Statistical Learning

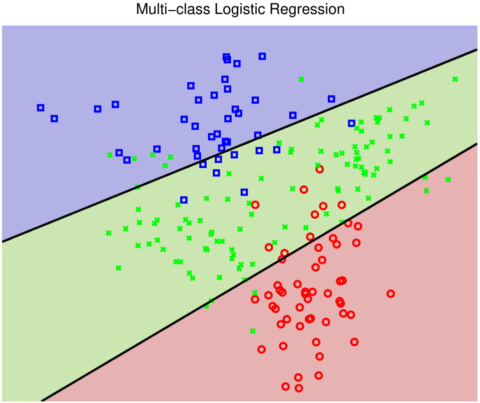
Master Spécialisé Intelligence Artificielle de Confiance (IAC)
@ Centrale Supélec en partenariat avec l'IRT SystemX
2024/2025.

Faïcel Chamroukhi

chamroukhi.com

Classification

1. Logistic Regression
   - Iteratively Reweighted Least Squares (IRLS)
2. Multi-class logistic regression
   - IRLS for Multi-class logistic regression



Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0, 1\}, \{-1, +1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.
  - ↪ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$
  - ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0, 1\}, \{-1, +1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

$\hookrightarrow$ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.

  $\hookrightarrow$ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$

  $\hookrightarrow$ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0, 1\}, \{-1, +1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.
  - ↪ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$
  - ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0, 1\}, \{-1, +1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

$\hookrightarrow$ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.
  - $\hookrightarrow$ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$
  - $\hookrightarrow$ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0, 1\}, \{-1, +1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

$\hookrightarrow$ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.
  - $\hookrightarrow$ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$
  - $\hookrightarrow$ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \rightarrow \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0, 1\}, \{-1, +1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.

  ↪ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$

  ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0,1\}, \{-1,+1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

↪ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X,h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.

  ↪ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$

  ↪ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0, 1\}, \{-1, +1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

$\hookrightarrow$ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.

  $\hookrightarrow$ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$

  $\hookrightarrow$ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Learning Framework for Classification

- The data are represented by a random pair $(\boldsymbol{X}, Y) \in \mathcal{X} \times \mathcal{Y}$ where $\boldsymbol{X}$ is a vector of descriptors for some variable of interest Y
- The objective is **Prediction**, i.e. to seek for a prediction function $h : \mathcal{X} \to \mathcal{Y}$ for which $\widehat{y} = h(\boldsymbol{x})$ is a good approximation of the true output y
- In a **classification** problem : typically $\boldsymbol{X} \in \mathcal{X} \subset \mathbb{R}^p$ and $Y \in \mathcal{Y} = \{0, 1\}, \{-1, +1\}$ (binary classification) or $\{1, \cdots, K\}$ (multiclass classification)

$\hookrightarrow$ We will mainly focus on parametric probabilistic models of the form

$$Y = h(X) + \epsilon, \epsilon \sim p_\theta$$

with the conditional distr. $P(Y|X, h)$ can be computed in terms of $P_\theta(Y - h(X))$.

- Data : a random sample $(\boldsymbol{X}_i, Y_i)_{i=1}^n$ with observed values $\mathcal{D}_n = (\boldsymbol{x}_i, y_i)_{i=1}^n$
- **Data-Scientist's role :** given the **data**, choose a **prediction function** h from a class $\mathcal{H}$ that attempts to "minimize" the prediction error for of all possible data (**risk**) $R(h)$, under a **loss** function ℓ measuring the error of predicting Y by $h(X)$.
  - $\hookrightarrow$ minimize the **empirical risk** (data-$\mathcal{D}_n$-driven) $R_n(h)$
  - $\hookrightarrow$ Minimizing $R_n(h)$ may require an optimization **algorithm** $\mathcal{A}$
- Data-Scientist's "**Toolbox**" **:** {Data, loss, hypothesis, algorithm}

Prediction/decision function

Def. Classifier or classification rule

$$\begin{aligned} h \colon \mathcal{X} &\to \mathcal{Y} \\ x &\mapsto h(x) \end{aligned}$$

is a decision/prediction function, parametric or not, linear or not, ...

Example : Linear predictors

$$\begin{aligned} h \colon \mathbb{R}^p &\to \mathbb{R} \\ x &\mapsto \langle x, \theta \rangle = \theta^T x \end{aligned}$$

The **predicted** values of Y_i's for new covariates $X_i = x_i$s correspond to

$$\widehat{y}_i = h(x_i)$$

Example : Linear predictors (cont.) : $\widehat{y}_i = \langle x_i, \theta \rangle = \theta^T x_i$

Q : How good we are in prediction on a particular pair (x, y) ?

Prediction/decision function

Def. Classifier or classification rule

$$h\colon \mathcal{X} \to \mathcal{Y}$$
$$x \mapsto h(x)$$

is a decision/prediction function, parametric or not, linear or not, ...

Example : Linear predictors

$$h\colon \mathbb{R}^p \to \mathbb{R}$$
$$x \mapsto \langle x, \theta \rangle = \theta^T x$$

The **predicted** values of Y_i's for new covariates $X_i = x_i$s correspond to

$$\widehat{y}_i = h(x_i)$$

Example : Linear predictors (cont.) : $\widehat{y}_i = \langle x_i, \theta \rangle = \theta^T x_i$

Q : How good we are in prediction on a particular pair (x, y) ?

Prediction/decision function

Def. Classifier or classification rule

$$\begin{aligned} h \colon \mathcal{X} &\to \mathcal{Y} \\ x &\mapsto h(x) \end{aligned}$$

is a decision/prediction function, parametric or not, linear or not, ...

Example : Linear predictors

$$\begin{aligned} h \colon \mathbb{R}^p &\to \mathbb{R} \\ x &\mapsto \langle x, \theta \rangle = \theta^T x \end{aligned}$$

The **predicted** values of Y_i's for new covariates $X_i = x_i$s correspond to

$$\widehat{y}_i = h(x_i)$$

Example : Linear predictors (cont.) : $\widehat{y}_i = \langle x_i, \theta \rangle = \theta^T x_i$

Q : How good we are in prediction on a particular pair (x, y) ?

Prediction/decision function

Def. Classifier or classification rule

$$h\colon \mathcal{X} \to \mathcal{Y}$$
$$x \mapsto h(x)$$

is a decision/prediction function, parametric or not, linear or not, ...

Example : Linear predictors

$$h\colon \mathbb{R}^p \to \mathbb{R}$$
$$x \mapsto \langle x, \theta \rangle = \theta^T x$$

The **predicted** values of Y_i's for new covariates $X_i = x_i$s correspond to

$$\widehat{y}_i = h(x_i)$$

Example : Linear predictors (cont.) : $\widehat{y}_i = \langle x_i, \theta \rangle = \theta^T x_i$

Q : How good we are in prediction on a particular pair (x, y)?

Loss

Def. Loss function

$$\ell\colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$
$$(y, h(x)) \mapsto \ell(y, h(x))$$

It measures how good we are on a particular (x, y) pair.

Examples of loss functions in classification

- "0-1" loss : $\ell(y, h(x)) = \mathbb{1}_{h(x)\neq y}$
- logarithmic loss : $\ell(y, h_\theta(x)) = -\log(p_\theta(x, y))$

  Denoting $\ell(y, h(x)) = \phi(yh(x))$
- Hinge loss $\phi_{\text{hinge}}(u) = (1 - u)_+$
- Logistic loss $\phi_{\text{logistic}}(u) = \log(1 + \exp(-u))$
- Exponential loss $\phi_{\exp}(u) = \exp(-u)$

Loss

Def. Loss function

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$
$$(y, h(x)) \mapsto \ell(y, h(x))$$

It measures how good we are on a particular (x, y) pair.

Examples of loss functions in classification

- "0-1" loss : $\ell(y, h(x)) = \mathbb{1}_{h(x) \neq y}$
- logarithmic loss : $\ell(y, h_\theta(x)) = -\log(p_\theta(x, y))$

  Denoting $\ell(y, h(x)) = \phi(yh(x))$
- Hinge loss $\phi_{\text{hinge}}(u) = (1 - u)_+$
- Logistic loss $\phi_{\text{logistic}}(u) = \log(1 + \exp(-u))$
- Exponential loss $\phi_{\exp}(u) = \exp(-u)$

Loss

Def. Loss function

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$
$$(y, h(x)) \mapsto \ell(y, h(x))$$

It measures how good we are on a particular (x, y) pair.

Examples of loss functions in classification

- "0-1" loss : $\ell(y, h(x)) = \mathbb{1}_{h(x) \neq y}$
- logarithmic loss : $\ell(y, h_\theta(x)) = -\log(p_\theta(x, y))$

  Denoting $\ell(y, h(x)) = \phi(yh(x))$
- Hinge loss $\phi_{\text{hinge}}(u) = (1-u)_+$
- Logistic loss $\phi_{\text{logistic}}(u) = \log(1 + \exp(-u))$
- Exponential loss $\phi_{\exp}(u) = \exp(-u)$

Loss

Def. Loss function

$$\ell \colon \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$$
$$(y, h(x)) \mapsto \ell(y, h(x))$$

It measures how good we are on a particular (x, y) pair.

Examples of loss functions in classification

- "0-1" loss : $\ell(y, h(x)) = \mathbb{1}_{h(x) \neq y}$
- logarithmic loss : $\ell(y, h_\theta(x)) = -\log(p_\theta(x, y))$

  Denoting $\ell(y, h(x)) = \phi(yh(x))$
- Hinge loss $\phi_{\text{hinge}}(u) = (1 - u)_+$
- Logistic loss $\phi_{\text{logistic}}(u) = \log(1 + \exp(-u))$
- Exponential loss $\phi_{\exp}(u) = \exp(-u)$

Examples of loss functions in classification

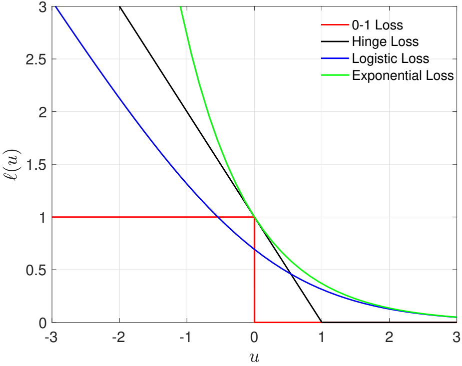


FIGURE – Some loss functions in classification : curves of $\ell(u)$ for $u = yh(x)$; $y \in \{-1, 1\}$. [plot_losses_classification.m]

For $y \in \{-1, 1\}$, with $u = yh(x)$:

- "0-1" loss : $\ell(u) = \mathbb{1}_{\text{sign}(u)\neq 1}$
- Hinge loss $\ell(_{\text{hinge}}(u) = (1-u)_+$
- Logistic loss $\ell(_{\text{logistic}}(u) = \log(1 + \exp(-u))$
- Exponential loss $\ell(_{\exp}(u) = \exp(-u)$

Risk

- **Risk** : the *Expected loss* :

$$R(h) = \mathbb{E}_P[\ell(Y, h(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(x)) dP(x, y)$$

↪ the error of approximating Y by model/hypothesis $h(X)$ as measured by a chosen loss function $\ell(Y, h(X))$ given the pair (X, Y) with (unknown) joint distribution P,

↪ prediction error : measures the generalization performance of the function h.

- **"0-1" Risk :** Under the "0-1"-loss $\ell(y, h(x)) = \mathbb{1}_{h(x)\neq y}$:

$$R(h) = \mathbb{E}_P[\mathbb{1}_{h(x)\neq y}] = \mathbb{P}(h(X) \neq Y). = \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{h(x)\neq y} dP(x, y)$$

↪ This is the most used risk in classification

Q : what is the best function h ? or equivalently, when the risk $R(h)$ is optimal ?

Risk

- **Risk** : the *Expected loss* :

$$R(h) = \mathbb{E}_P[\ell(Y, h(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(x))dP(x, y)$$

↪ the error of approximating Y by model/hypothesis $h(X)$ as measured by a chosen loss function $\ell(Y, h(X))$ given the pair (X, Y) with (unknown) joint distribution P,

↪ prediction error : measures the generalization performance of the function h.

- **"0-1" Risk :** Under the "0-1"-loss $\ell(y, h(x)) = \mathbb{1}_{h(x)\neq y}$:

$$R(h) = \mathbb{E}_P[\mathbb{1}_{h(x)\neq y}] = \mathbb{P}(h(X) \neq Y). = \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{h(x)\neq y} dP(x, y)$$

↪ This is the most used risk in classification

Q : what is the best function h ? or equivalently, when the risk $R(h)$ is optimal ?

Risk

- **Risk** : the *Expected loss* :

$$R(h) = \mathbb{E}_P[\ell(Y, h(X))] = \int_{\mathcal{X}\times\mathcal{Y}} \ell(y, h(x)) dP(x, y)$$

↪ the error of approximating Y by model/hypothesis $h(X)$ as measured by a chosen loss function $\ell(Y, h(X))$ given the pair (X, Y) with (unknown) joint distribution P,

↪ prediction error : measures the generalization performance of the function h.

- **"0-1" Risk :** Under the "0-1"-loss $\ell(y, h(x)) = \mathbb{1}_{h(x)\neq y}$:

$$R(h) = \mathbb{E}_P[\mathbb{1}_{h(x)\neq y}] = \mathbb{P}(h(X) \neq Y). = \int_{\mathcal{X}\times\mathcal{Y}} \mathbb{1}_{h(x)\neq y} dP(x, y)$$

↪ This is the most used risk in classification

Q : what is the best function h ? or equivalently, when the risk $R(h)$ is optimal ?

Optimal prediction function I

Theorem (The Bayes classifier)

Under the (0-1)-loss, $\ell(Y, h(X)) = \mathbb{1}_{h(X)\neq Y}$, *the classification function* $h^*(x)$ *minimizing the risk (the Bayes classifier)*

$$R(h) = \mathbb{P}(Y \neq h(X)) = \int_{\mathcal{X}} \mathbb{P}(Y \neq h(X)|X = x)dP_X(x)$$

is given by

$$\forall x \in \mathcal{X}, \quad h^*(x) = \arg\max_{k\in\mathcal{Y}} \mathbb{P}(Y = k|X = x).$$

Def. Decision boundaries

The decision bounadry between each pair of classes k and ℓ, $(k, \ell) \in \mathcal{Y} \times \mathcal{Y}$ is defined by

$$\eta_{k,\ell}(x) = \{x : \mathbb{P}(Y = k|X = x) = \mathbb{P}(Y = \ell X = x)\}$$

Optimal prediction function II

Proof. Optimal classifier.

Given $X = x$, the conditional risk under the 0-1 loss is

$$\begin{aligned} r(h|X=x) &= \mathbb{E}_{Y|X=x}[\ell(Y,h(X))|X=x] = \mathbb{E}_{Y|X=x}[\mathbb{1}_{Y \neq h(X)}|X=x] \\ &= \mathbb{P}[Y \neq h(X)|X=x] \\ &= 1 - \mathbb{P}[Y = h(X)|X=x]. \end{aligned}$$

By noting that

$$\begin{aligned} \min_{k \in \mathcal{Y}} r(h|X=x) &= -1 + \max_{k \in \mathcal{Y}} \mathbb{P}(Y=k|X=x) \\ \arg\min_{k \in \mathcal{Y}} r(h|X=x) &= \arg\max_{k \in \mathcal{Y}} \mathbb{P}(Y=k|X=x) \end{aligned}$$

we see that $h^*(x) = \arg\max_{k \in \mathcal{Y}} \mathbb{P}(Y=k|X=x)$ achieves the minimized risk $r(h|X=x)$.
Then the risk $R(h^*) = \mathbb{E}_X[-1 + \max_{k \in \mathcal{Y}} \mathbb{P}(Y=k|X=x)]$ is Bayes. □

Goal : estimate h^*, knowing only the data sample $D_n = (X_i, Y_i)_{i=1}^n$ and loss ℓ.

Empirical Risk Minimization & MLE I

- Then *Expected loss* $R(h)$ depends on the joint distribution P of the pair (X, Y). In real situations P is in unknown, as we only have a sample $D_n = (X_i, Y_i)_{1 \leq i \leq n}$,

$\hookrightarrow$ We attempt to minimize the **Empirical Risk**

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} l(Y_i, h(X_i))$$

to estimate h^* (within a family $\mathcal{H}$)

$\hookrightarrow$ ERM : $\widehat{h}_n = \arg\min_{h \in \mathcal{H}} R_n(h)$ is the **ERM** of h

- 0-1 Risk : Under the 0-1 loss (standard in classification) : $\ell_{0-1}(y, h(x)) = \mathbb{1}_{y \neq h(x)}$, the empirical 0-1 risk is

$$R_n(h) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{Y_i \neq h(X_i)}$$

ERM and MLE : Conditional maximum likelihood risks :

- MLE (density estimation framework) : We seek for an esitmator of the parameters θ of the joint distribution $p_\theta(x, y)$.

Empirical Risk Minimization & MLE II

- In discriminative learning (eg. logistic regression), we are interested in estimating the conditional distribution $P(Y|X)$, rather than the joint distribution $P(X,Y)$.
- Consider the log-loss : $\ell(y, h_\theta(x)) = -\log(p_\theta(y|x))$. We therefore have the conditional log-likehood risks

$$R(\theta) = -\mathbb{E}[\log p_\theta(Y|X)] \text{ and } R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i|x_i).$$

- For an i.i.d sample $\{(x_i, y_i)_{i=1}^n\}$, the conditional log-likelihood function of θ is :

$$\log L(\theta) = \sum_{i=1}^{n} \log p_\theta(y_i|x_i)$$

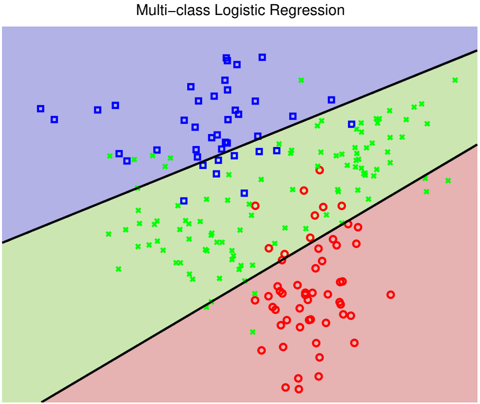
Then

$$R_n(\theta) = -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i|x_i) = -\frac{1}{n}\log L(\theta)$$

$\hookrightarrow$ With this log-loss, ERM coincides with conditional MLE.

- Liner classifier : Consider $\mathcal{H} = \{h_\theta(x) = \alpha + \beta^T x\}$, the set of linear functions in x

Logistic Regression



Logistic Regression

- We model the random pair $(\boldsymbol{X}, Y)$ where $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ is the predictor and the response $Y \in \mathcal{Y} = \{0, 1\}$ is the class label of $\boldsymbol{X}$
- Logistic Regression : Probabilistic Discriminative approach to model $\mathbb{P}(Y|\boldsymbol{X})$ as

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \text{Logistic}(x^T\theta) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}.$$

- $Y|X = x$ is Bernoulli with probability of success $\pi_{\boldsymbol{\theta}}(\boldsymbol{x})$, i.e.

$$\forall y \in \{0, 1\}, \mathbb{P}_{\boldsymbol{\theta}}(Y = y|X = x) = \pi_{\boldsymbol{\theta}}(x)^y(1 - \pi_{\boldsymbol{\theta}}(x))^{1-y}$$

  where $\pi(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta})$ is the sigmoid function.
- Classification rule : We have $h(x)$ is defined as

$$h_\theta(x) = \begin{cases} 1 \text{ if } \mathbb{P}(Y = 1|X = x) = \text{Logistic}(x^T\theta) > \frac{1}{2}, \\ 0, \text{otherwise.} \end{cases} \quad Eq. : h_\theta(x) = \begin{cases} 1 \text{ if } \theta^T x > 0, \\ 0, \text{otherwise.} \end{cases}$$

- The latter comes from the linear bounadry $\{x : \log \frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \beta_0 + \beta^T \boldsymbol{x} = 0\}$
- The parameter vector of the model $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^{d+1}$
- **Q** : Fit $\boldsymbol{\theta}$ from the training data.

Logistic Regression

- We model the random pair $(\boldsymbol{X}, Y)$ where $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ is the predictor and the response $Y \in \mathcal{Y} = \{0, 1\}$ is the class label of $\boldsymbol{X}$
- Logistic Regression : Probabilistic Discriminative approach to model $\mathbb{P}(Y|\boldsymbol{X})$ as

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \text{Logistic}(x^T\theta) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}\,.$$

- $Y|X = x$ is Bernoulli with probability of success $\pi_{\boldsymbol{\theta}}(\boldsymbol{x})$, i.e.

$$\forall y \in \{0, 1\}, \mathbb{P}_{\boldsymbol{\theta}}(Y = y|X = x) = \pi_{\boldsymbol{\theta}}(x)^y(1 - \pi_{\boldsymbol{\theta}}(x))^{1-y}$$

  where $\pi(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta})$ is the sigmoid function.
- Classification rule : We have $h(x)$ is defined as

$$h_\theta(x) = \begin{cases} 1 \text{ if } \mathbb{P}(Y = 1|X = x) = \text{Logistic}(x^T\theta) > \frac{1}{2}, \\ 0, \text{otherwise.} \end{cases} \quad Eq. : h_\theta(x) = \begin{cases} 1 \text{ if } \theta^T x > 0, \\ 0, \text{otherwise.} \end{cases}$$

- The latter comes from the linear bounadry $\{x : \log \frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \beta_0 + \beta^T \boldsymbol{x} = 0\}$
- The parameter vector of the model $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^{d+1}$
- **Q** : Fit $\boldsymbol{\theta}$ from the training data.

Logistic Regression

- We model the random pair $(\boldsymbol{X}, Y)$ where $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ is the predictor and the response $Y \in \mathcal{Y} = \{0, 1\}$ is the class label of $\boldsymbol{X}$
- Logistic Regression : Probabilistic Discriminative approach to model $\mathbb{P}(Y|\boldsymbol{X})$ as

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \text{Logistic}(x^T\theta) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}\,.$$

- $Y|X = x$ is Bernoulli with probability of success $\pi_{\boldsymbol{\theta}}(\boldsymbol{x})$, i.e.

$$\forall y \in \{0, 1\}, \mathbb{P}_{\boldsymbol{\theta}}(Y = y|X = x) = \pi_{\boldsymbol{\theta}}(x)^y (1 - \pi_{\boldsymbol{\theta}}(x))^{1-y}$$

  where $\pi(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta})$ is the sigmoid function.

- Classification rule : We have $h(x)$ is defined as

$$h_\theta(x) = \begin{cases} 1 \text{ if } \mathbb{P}(Y = 1|X = x) = \text{Logistic}(x^T\theta) > \frac{1}{2}, \\ 0, \text{otherwise.} \end{cases} \quad Eq. : h_\theta(x) = \begin{cases} 1 \text{ if } \theta^T x > 0, \\ 0, \text{otherwise.} \end{cases}$$

- The latter comes from the linear bounadry $\{x : \log \frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \beta_0 + \beta^T \boldsymbol{x} = 0\}$
- The parameter vector of the model $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^{d+1}$
- **Q** : Fit $\boldsymbol{\theta}$ from the training data.

Logistic Regression

- We model the random pair $(\boldsymbol{X}, Y)$ where $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ is the predictor and the response $Y \in \mathcal{Y} = \{0,1\}$ is the class label of $\boldsymbol{X}$
- Logistic Regression : Probabilistic Discriminative approach to model $\mathbb{P}(Y|\boldsymbol{X})$ as

$$\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x}) = \text{Logistic}(x^T\theta) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}{1+\exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}\,.$$

- $Y|X=x$ is Bernoulli with probability of success $\pi_{\boldsymbol{\theta}}(\boldsymbol{x})$, i.e.

$$\forall y \in \{0,1\},\, \mathbb{P}_{\boldsymbol{\theta}}(Y=y|X=x) = \pi_{\boldsymbol{\theta}}(x)^y(1-\pi_{\boldsymbol{\theta}}(x))^{1-y}$$

  where $\pi(\boldsymbol{x};\boldsymbol{\theta}) = \mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x};\boldsymbol{\theta})$ is the sigmoid function.
- Classification rule : We have $h(x)$ is defined as

$$h_\theta(x) = \begin{cases} 1 \text{ if } \mathbb{P}(Y=1|X=x) = \text{Logistic}(x^T\theta) > \frac{1}{2}, \\ 0, \text{otherwise.}\end{cases} \quad Eq.: h_\theta(x) = \begin{cases} 1 \text{ if } \theta^T x > 0, \\ 0, \text{otherwise.}\end{cases}$$

- The latter comes from the linear bounadry $\{x : \log\frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \beta_0 + \beta^T\boldsymbol{x} = 0\}$
- The parameter vector of the model $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^{d+1}$
- **Q** : Fit $\boldsymbol{\theta}$ from the training data.

Logistic Regression

- We model the random pair $(\boldsymbol{X}, Y)$ where $\boldsymbol{X}_i \in \mathcal{X} \subset \mathbb{R}^d$ is the predictor and the response $Y \in \mathcal{Y} = \{0, 1\}$ is the class label of $\boldsymbol{X}$
- Logistic Regression : Probabilistic Discriminative approach to model $\mathbb{P}(Y|\boldsymbol{X})$ as

$$\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \text{Logistic}(x^T\theta) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x})} \cdot$$

- $Y|X = x$ is Bernoulli with probability of success $\pi_{\boldsymbol{\theta}}(\boldsymbol{x})$, i.e.

$$\forall y \in \{0, 1\}, \mathbb{P}_{\boldsymbol{\theta}}(Y = y|X = x) = \pi_{\boldsymbol{\theta}}(x)^y(1 - \pi_{\boldsymbol{\theta}}(x))^{1-y}$$

  where $\pi(\boldsymbol{x}; \boldsymbol{\theta}) = \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}; \boldsymbol{\theta})$ is the sigmoid function.
- Classification rule : We have $h(x)$ is defined as

$$h_\theta(x) = \begin{cases} 1 \text{ if } \mathbb{P}(Y = 1|X = x) = \text{Logistic}(x^T\theta) > \frac{1}{2}, \\ 0, \text{otherwise.} \end{cases} \quad Eq. : h_\theta(x) = \begin{cases} 1 \text{ if } \theta^T x > 0, \\ 0, \text{otherwise.} \end{cases}$$

- The latter comes from the linear bounadry $\{x : \log \frac{\mathbb{P}(Y=1|\boldsymbol{X}=\boldsymbol{x})}{\mathbb{P}(Y=0|\boldsymbol{X}=\boldsymbol{x})} = \beta_0 + \beta^T\boldsymbol{x} = 0\}$
- The parameter vector of the model $\boldsymbol{\theta} = (\beta_0, \boldsymbol{\beta}^\top)^\top \in \mathbb{R}^{d+1}$
- **Q** : Fit $\boldsymbol{\theta}$ from the training data.

Linear decision boundary :

$$
\begin{aligned}
\eta_{1,0}(\boldsymbol{x}) &= \{\boldsymbol{x} : h_1(\boldsymbol{x}) = h_0(\boldsymbol{x})\} \\
&= \{\boldsymbol{x} : \mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Y = 0|\boldsymbol{X} = \boldsymbol{x})\} \\
&= \{\boldsymbol{x} : \log \frac{\mathbb{P}(Y = 1|\boldsymbol{X} = \boldsymbol{x})}{\mathbb{P}(\boldsymbol{X} = \boldsymbol{x})} = 0\} \\
&= \{\boldsymbol{x} : \log \frac{\frac{\exp(\beta_0+\boldsymbol{\beta}^\top\boldsymbol{x})}{1+\exp(\beta_0+\boldsymbol{\beta}^\top\boldsymbol{x})}}{\frac{1}{1+\exp(\beta_0+\boldsymbol{\beta}^\top\boldsymbol{x})}} = 0\} \\
&= \{\boldsymbol{x} : \beta_0 + \boldsymbol{\beta}^\top\boldsymbol{x} = 0\}
\end{aligned}
$$

Logistic Regression

$\hookrightarrow$ Maximum conditional likelihood.

- The conditional log-likelihood function :

$$
\begin{aligned}
L(\boldsymbol{\theta}) &= \log \mathbb{P}(Y_1 = y_1, \ldots, Y_n = y_n | \boldsymbol{X}_1 = \boldsymbol{x}_1, \ldots, \boldsymbol{X}_n = \boldsymbol{x}_n; \boldsymbol{\theta}) \\
&= \log \prod_{i=1}^{n} \mathbb{P}(Y_i = y_i | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta}) \\
&= \log \prod_{i=1}^{n} \mathbb{P}(Y_i = 1 | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta})^{y_i} \mathbb{P}(Y_i = 0 | \boldsymbol{X}_i = \boldsymbol{x}_i; \boldsymbol{\theta})^{1-y_i} \\
&= \sum_{i=1}^{n} y_i \log \pi(\boldsymbol{x}_i; \boldsymbol{\theta}) + (1 - y_i) \log \left(1 - \pi(\boldsymbol{x}_i; \boldsymbol{\theta})\right) \\
&= \sum_{i=1}^{n} y_i(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i) - \log(1 + \exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i)) \\
&= \sum_{i=1}^{n} y_i(1, \boldsymbol{x}_i)^\top \boldsymbol{\theta} - \log\{1 + \exp((1, \boldsymbol{x}_i)^\top \boldsymbol{\theta})\}. \\
&= \sum_{i=1}^{n} y_i \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta} - \log\{1 + \exp(\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})\}.
\end{aligned}
$$

Logistic Regression II

- A concave function in $\boldsymbol{\theta} \hookrightarrow$ Global maximization
- However, it does not admit a closed-form solution

  $\hookrightarrow$ Numerical optimization : Iterative Reweighted Least Squares (IRLS) Algorithm.

ERM for logistic regression

- Conditional log-likelihood

$$\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} y_i (1, \boldsymbol{x}_i)^\top \boldsymbol{\theta} - \log\{1 + \exp((1, \boldsymbol{x}_i)^\top \boldsymbol{\theta})\}.$$

- Conditional ERM : Consider the log-loss :

$$\ell(y, h_\theta(x)) = -\log(p_\theta(y|x))$$

and the hypothesis

$$h_Y(\boldsymbol{X};\boldsymbol{\theta}) = \mathbb{P}_{\boldsymbol{\theta}}(Y|\boldsymbol{X}) = \pi_{\boldsymbol{\theta}}(\boldsymbol{X})^Y (1 - \pi_{\boldsymbol{\theta}}(\boldsymbol{X}))^{1-Y}$$

- The corresponding conditional empirical risk is by definition

$$\begin{aligned} R_n(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h_\theta(x_i)) \\ &= -\frac{1}{n}\sum_{i=1}^{n} \log p_\theta(y_i|x_i) \\ &= -\frac{1}{n}\log L(\boldsymbol{\theta}) \end{aligned}$$

$\hookrightarrow$ With the log-loss, the conditional ERM coincides with conditional MLE.

Newton-Raphson iteration : $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\nabla^2 L(\boldsymbol{\theta}^{(t)})\right]^{-1} \nabla L(\boldsymbol{\theta}^{(t)})$

- Let $\widetilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^\top)^\top$, then : $L(\boldsymbol{\theta}) = \sum_{i=1}^{n} y_i \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta} - \log\{1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})\}$.
- Gradient vector :

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{i=1}^{n} \left[\frac{\partial}{\partial \boldsymbol{\theta}} y_i \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta} - \frac{\partial}{\partial \boldsymbol{\theta}} \log(1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta}))\right] = \sum_{i=1}^{n} y_i \widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_i \pi(\boldsymbol{x}_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i; \boldsymbol{\theta})) \cdot \qquad (1)\end{aligned}$$

- Hessian matrix :

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} &= -\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i \frac{\partial}{\partial \boldsymbol{\theta}^\top} \left\{\frac{\exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}{1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}\right\} - \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i \frac{\widetilde{\boldsymbol{x}}_i^\top \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}{\left(1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})\right)^2} \\ &= -\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i; \boldsymbol{\theta})(1 - \pi(\boldsymbol{x}_i; \boldsymbol{\theta})) \qquad (2)\end{aligned}$$

- The Newton-Raphson iterative update of $\boldsymbol{\theta}$ has therefore the following expression :

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left[\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)})(1 - \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}))\right]^{-1} \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}))$$

Iteratively Reweighted Least Squares (IRLS)

Newton-Raphson iteration : $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\nabla^2 L(\boldsymbol{\theta}^{(t)})\right]^{-1} \nabla L(\boldsymbol{\theta}^{(t)})$

- Let $\widetilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^\top)^\top$, then : $L(\boldsymbol{\theta}) = \sum_{i=1}^n y_i \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta} - \log\{1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})\}$.
- Gradient vector :

$$\begin{aligned}\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{i=1}^n \big[\frac{\partial}{\partial \boldsymbol{\theta}} y_i \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta} - \frac{\partial}{\partial \boldsymbol{\theta}} \log(1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta}))\big] = \sum_{i=1}^n y_i \widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_i \pi(\boldsymbol{x}_i; \boldsymbol{\theta}) \\ &= \sum_{i=1}^n \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i; \boldsymbol{\theta})) \cdot \qquad (1)\end{aligned}$$

- Hessian matrix :

$$\begin{aligned}\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} &= -\sum_{i=1}^n \widetilde{\boldsymbol{x}}_i \frac{\partial}{\partial \boldsymbol{\theta}^\top} \{\frac{\exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}{1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}\} - \sum_{i=1}^n \widetilde{\boldsymbol{x}}_i \frac{\widetilde{\boldsymbol{x}}_i^\top \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}{\left(1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})\right)^2} \\ &= -\sum_{i=1}^n \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i; \boldsymbol{\theta})(1 - \pi(\boldsymbol{x}_i; \boldsymbol{\theta})) \qquad (2)\end{aligned}$$

- The Newton-Raphson iterative update of $\boldsymbol{\theta}$ has therefore the following expression :

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \big[\sum_{i=1}^n \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)})(1 - \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}))\big]^{-1} \sum_{i=1}^n \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}))$$

Iteratively Reweighted Least Squares (IRLS)

Newton-Raphson iteration : $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\nabla^2 L(\boldsymbol{\theta}^{(t)})\right]^{-1} \nabla L(\boldsymbol{\theta}^{(t)})$

- Let $\widetilde{\boldsymbol{x}}_i = (1, \boldsymbol{x}_i^\top)^\top$, then : $L(\boldsymbol{\theta}) = \sum_{i=1}^n y_i \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta} - \log\{1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})\}$.
- Gradient vector :

$$\begin{aligned}
\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} &= \sum_{i=1}^n \left[\frac{\partial}{\partial \boldsymbol{\theta}} y_i \widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta} - \frac{\partial}{\partial \boldsymbol{\theta}} \log(1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta}))\right] = \sum_{i=1}^n y_i \widetilde{\boldsymbol{x}}_i - \widetilde{\boldsymbol{x}}_i \pi(\boldsymbol{x}_i; \boldsymbol{\theta}) \\
&= \sum_{i=1}^n \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i; \boldsymbol{\theta})) \cdot \qquad (1)
\end{aligned}$$

- Hessian matrix :

$$\begin{aligned}
\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} &= -\sum_{i=1}^n \widetilde{\boldsymbol{x}}_i \frac{\partial}{\partial \boldsymbol{\theta}^\top} \left\{\frac{\exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}{1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}\right\} - \sum_{i=1}^n \widetilde{\boldsymbol{x}}_i \frac{\widetilde{\boldsymbol{x}}_i^\top \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})}{\left(1 + \exp(\widetilde{\boldsymbol{x}}_i^\top \boldsymbol{\theta})\right)^2} \\
&= -\sum_{i=1}^n \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i; \boldsymbol{\theta})(1 - \pi(\boldsymbol{x}_i; \boldsymbol{\theta})) \qquad (2)
\end{aligned}$$

- The Newton-Raphson iterative update of $\boldsymbol{\theta}$ has therefore the following expression :

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \left[\sum_{i=1}^n \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)})(1 - \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}))\right]^{-1} \sum_{i=1}^n \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i; \boldsymbol{\theta}^{(t)}))$$

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \big[\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)})(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)}))\big]^{-1} \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)}))$$

Matrix form the NR iteration update :

Let

- $\widetilde{\mathbf{X}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n)^\top$ matrix whose rows are the augmented input vectors $(1, \boldsymbol{x}_i^\top)$
- $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ the vector on binary labels y_i
- $\mathbf{p} = (\pi(\boldsymbol{x}_1;\boldsymbol{\theta}), \ldots, \pi(\boldsymbol{x}_n;\boldsymbol{\theta}))^\top$ the vector of logistic probabilities
- $\mathbf{W} = \mathrm{diag}(\mathbf{p} \odot (\mathbf{1}_n - \mathbf{p}))$ diagonal matrix with $(\mathbf{W})_{ii} = \pi(\boldsymbol{x}_i;\boldsymbol{\theta})\,(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))$
- $\widetilde{\boldsymbol{y}} = \widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)} + (\mathbf{W}^{(t)})^{-1}(\boldsymbol{y} - \mathbf{p}^{(t)})$ the current approximate response

  Then

↪ Vectorial form of the Gradient : $\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \widetilde{\mathbf{X}}^\top(\boldsymbol{y} - \mathbf{p}^{(t)})$

↪ Vectorial form of the Hessian matrix : $\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = -\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}$

Iteratively Reweighted Least Squares (IRLS)

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \Big[\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)})(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)}))\Big]^{-1} \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)}))$$

Matrix form the NR iteration update :

Let

- $\widetilde{\mathbf{X}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n)^\top$ matrix whose rows are the augmented input vectors $(1, \boldsymbol{x}_i^\top)$
- $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ the vector on binary labels y_i
- $\mathbf{p} = (\pi(\boldsymbol{x}_1;\boldsymbol{\theta}), \ldots, \pi(\boldsymbol{x}_n;\boldsymbol{\theta}))^\top$ the vector of logistic probabilities
- $\mathbf{W} = \mathrm{diag}(\mathbf{p} \odot (\mathbf{1}_n - \mathbf{p}))$ diagonal matrix with $(\mathbf{W})_{ii} = \pi(\boldsymbol{x}_i;\boldsymbol{\theta})\,(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))$
- $\widetilde{\boldsymbol{y}} = \widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)} + (\mathbf{W}^{(t)})^{-1}(\boldsymbol{y} - \mathbf{p}^{(t)})$ the current approximate response

  Then

↪ Vectorial form of the Gradient : $\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \widetilde{\mathbf{X}}^\top(\boldsymbol{y} - \mathbf{p}^{(t)})$

↪ Vectorial form of the Hessian matrix : $\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = -\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}$

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \Big[\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)})(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)}))\Big]^{-1} \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)}))$$

Matrix form the NR iteration update :

Let

- $\widetilde{\mathbf{X}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n)^\top$ matrix whose rows are the augmented input vectors $(1, \boldsymbol{x}_i^\top)$
- $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ the vector on binary labels y_i
- $\mathbf{p} = (\pi(\boldsymbol{x}_1;\boldsymbol{\theta}), \ldots, \pi(\boldsymbol{x}_n;\boldsymbol{\theta}))^\top$ the vector of logistic probabilities
- $\mathbf{W} = \mathrm{diag}(\mathbf{p} \odot (\mathbf{1}_n - \mathbf{p}))$ diagonal matrix with $(\mathbf{W})_{ii} = \pi(\boldsymbol{x}_i;\boldsymbol{\theta})\,(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))$
- $\widetilde{\boldsymbol{y}} = \widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)} + (\mathbf{W}^{(t)})^{-1}(\boldsymbol{y} - \mathbf{p}^{(t)})$ the current approximate response

  Then

↪ Vectorial form of the Gradient : $\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \widetilde{\mathbf{X}}^\top(\boldsymbol{y} - \mathbf{p}^{(t)})$

↪ Vectorial form of the Hessian matrix : $\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^\top} = -\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}$

Iteratively Reweighted Least Squares (IRLS)

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} + \big[\sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i \widetilde{\boldsymbol{x}}_i^\top \pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)})(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)}))\big]^{-1} \sum_{i=1}^{n} \widetilde{\boldsymbol{x}}_i (y_i - \pi(\boldsymbol{x}_i;\boldsymbol{\theta}^{(t)}))$$

Matrix form the NR iteration update :

Let

- $\widetilde{\mathbf{X}} = (\widetilde{\boldsymbol{x}}_1, \ldots, \widetilde{\boldsymbol{x}}_n)^\top$ matrix whose rows are the augmented input vectors $(1, \boldsymbol{x}_i^\top)$
- $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$ the vector on binary labels y_i
- $\mathbf{p} = (\pi(\boldsymbol{x}_1;\boldsymbol{\theta}), \ldots, \pi(\boldsymbol{x}_n;\boldsymbol{\theta}))^\top$ the vector of logistic probabilities
- $\mathbf{W} = \mathrm{diag}(\mathbf{p} \odot (\mathbf{1}_n - \mathbf{p}))$ diagonal matrix with $(\mathbf{W})_{ii} = \pi(\boldsymbol{x}_i;\boldsymbol{\theta})\,(1-\pi(\boldsymbol{x}_i;\boldsymbol{\theta}))$
- $\widetilde{\boldsymbol{y}} = \widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)} + (\mathbf{W}^{(t)})^{-1}(\boldsymbol{y} - \mathbf{p}^{(t)})$ the current approximate response

  Then

↪ Vectorial form of the Gradient : $\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \widetilde{\mathbf{X}}^\top(\boldsymbol{y} - \mathbf{p}^{(t)})$

↪ Vectorial form of the Hessian matrix : $\frac{\partial^2 L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^\top} = -\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}$

Then we get the Matrix form :

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\nabla^2 L(\boldsymbol{\theta}^{(t)})\right]^{-1} \nabla L(\boldsymbol{\theta}^{(t)}) & (3)\\
&= \boldsymbol{\theta}^{(t)} + \left(\widetilde{\mathbf{X}}^\top \mathbf{W} \widetilde{\mathbf{X}}\right)^{-1} \widetilde{\mathbf{X}}^\top (\boldsymbol{y} - \mathbf{p}^{(t)}) & \\
&= \boldsymbol{\theta}^{(t)} + (\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T (\boldsymbol{y} - \mathbf{p}^{(t)}) & (4)\\
&= (\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \widetilde{\mathbf{X}})^{-1} \left[\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \widetilde{\mathbf{X}} \boldsymbol{\theta}^{(t)} + \widetilde{\mathbf{X}}^T (\boldsymbol{y} - \mathbf{p}^{(t)})\right] & (5)\\
&= (\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \widetilde{\mathbf{X}}^T \left[\mathbf{W}^{(t)} \widetilde{\mathbf{X}} \boldsymbol{\theta}^{(t)} + (\boldsymbol{y} - \mathbf{p}^{(t)})\right] & (6)\\
&= (\widetilde{\mathbf{X}}^\top \mathbf{W}^{(t)} \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \mathbf{W}^{(t)} \widetilde{\boldsymbol{y}} & (7)
\end{aligned}$$

Algorithm 1 Pseudo Code for Training Logistic Regression IRLS.

Inputs : n sample $(\boldsymbol{x}_i, y_i)_{i=1}^n$ arranged as $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$ and $\boldsymbol{y} = (y_1, \ldots, y_n)^\top$

Construct $\widetilde{\mathbf{X}} = [\mathbf{1}_n, \mathbf{X}]$

Initialization : $\boldsymbol{\theta}^{(0)}$; set $t \leftarrow 0$ (IRLS iteration)

while *increment in log-likelihood* $> \epsilon$ *(eg. 1e-6)* **do**

$\mathbf{p}^{(t)} = (\pi(\boldsymbol{x}_1; \boldsymbol{\theta}^{(t)}), \ldots, \pi(\boldsymbol{x}_n; \boldsymbol{\theta}^{(t)}))^\top = \exp(\widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)}) \oslash (\mathbf{1}_n + \exp(\widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)}))$

$\mathbf{W}^{(t)} = \mathrm{diag}(\mathbf{p}^{(t)} \odot (\mathbf{1}_n - \mathbf{p}^{(t)}))$

$\widetilde{\boldsymbol{z}} = \widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)} + (\mathbf{W}^{(t)})^{-1}(\boldsymbol{y} - \mathbf{p}^{(t)})$

$\boldsymbol{\theta}^{(t+1)} = (\widetilde{\mathbf{X}}^\top \mathbf{W}^{(t)} \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^\top \mathbf{W}^{(t)} \widetilde{\boldsymbol{z}}$

`% Convergence test`

`log-lik` $= \sum\{\boldsymbol{y} \odot (\widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)}) - \log(\mathbf{1}_n + \exp(\widetilde{\mathbf{X}}\boldsymbol{\theta}^{(t)}))\}$. `% log-likelihood.`

end

Result: $\widehat{\boldsymbol{\theta}} = \boldsymbol{\theta}^{(t)}$ the MLE of $\boldsymbol{\theta}$

Algorithm 2 Pseudo Code for Predicting with Logistic Regression.

Inputs : Test sample $(\boldsymbol{x}_i)_{i=1}^n$ arranged as $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^\top$, and parameter vector $\boldsymbol{\theta}$

Construct $\widetilde{\mathbf{X}} = [\mathbf{1}_n, \mathbf{X}]$

$\mathrm{probs} = \exp(\widetilde{\mathbf{X}}\boldsymbol{\theta}) \oslash (\mathbf{1}_n + \exp(\widetilde{\mathbf{X}}\boldsymbol{\theta}))$ `% Conditional probabilities`

$\widehat{\boldsymbol{y}} = \mathbb{1}_{\mathbf{probs} \geq 1/2}$ `% Predicted labels using Bayes rule` $(\arg\max)$

Result: $\widehat{\boldsymbol{y}}$ the predicted class labels

"Optimal" decision boundaries

Def. Decision boundaries

The decision bounadry between each pair of classes k and ℓ, $(k, \ell) \in \mathcal{Y} \times \mathcal{Y}$ is defined by

$$\eta_{k,\ell}(\boldsymbol{x}) = \{\boldsymbol{x} : \mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}) = \mathbb{P}(Y = \ell|\boldsymbol{X} = \boldsymbol{x})\}$$

- Plugin classifier : Prediction by the Bayes' decision rule

$$\widehat{h}(\boldsymbol{x}) = \arg\max_{k \in \mathcal{Y}} \mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}; \widehat{\boldsymbol{\theta}}) \quad (8)$$

- Plugin Decision boundaries : The decision bounadry between each pair of classes k and ℓ is defined by

$$\eta_{k,\ell}(\boldsymbol{x}; \widehat{\boldsymbol{\theta}}) = \{\boldsymbol{x} : \mathbb{P}(Y = k|\boldsymbol{X} = \boldsymbol{x}; \widehat{\boldsymbol{\theta}}) = \mathbb{P}(Y = \ell|\boldsymbol{X} = \boldsymbol{x}; \widehat{\boldsymbol{\theta}})\}$$

ERM vs MLE : Logistic Regression : $y \in \{0,1\}$ with $p_{\boldsymbol{\theta}}(y|\boldsymbol{x}) = \pi_{\boldsymbol{\theta}}(\boldsymbol{x})^y (1 - \pi_{\boldsymbol{\theta}}(\boldsymbol{x}))^{1-y}$, and $\pi_{\boldsymbol{\theta}}(\boldsymbol{x}) = \sigma(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x}) = \frac{\exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}{1 + \exp(\beta_0 + \boldsymbol{\beta}^T \boldsymbol{x})}$ is the logistic function.

$$R_n(\theta) = -\frac{1}{n} \sum_{i=1}^{n} \log p_{\boldsymbol{\theta}}(y_i|x_i) = -\frac{1}{n} \underbrace{\sum_{i=1}^{n} y_i(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i) - \log(1 + \exp(\beta_0 + \boldsymbol{\beta}^\top \boldsymbol{x}_i))}_{\text{Conditional log-likelihood}}$$

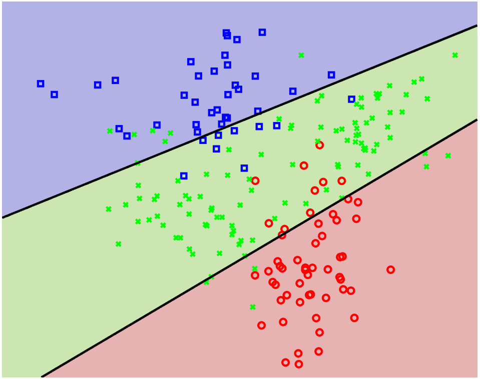
ERM for logistic regression / Logistic Loss

- Conditional log-likelihood ($y_i \in \{0,1\}$)
  $\log L(\boldsymbol{\theta}) = \sum_{i=1}^{n} y_i (1, \boldsymbol{x}_i)^\top \boldsymbol{\theta} - \log\{1 + \exp((1, \boldsymbol{x}_i)^\top \boldsymbol{\theta})\}$.
- Conditional ERM : Consider the logistic loss : $\ell(y, h_\theta(x)) = \log(1 + \exp(-y_i h_{\boldsymbol{\theta}}(x_i)))$, $y_i \in \{-1, +1\}$ and the hypothesis $h_{\boldsymbol{\theta}}(\boldsymbol{X}) = \beta_0 + \beta^T \boldsymbol{X}$
- The corresponding conditional empirical risk is by definition

$$
\begin{aligned}
R_n(h) &= \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, h_\theta(x_i)) = \frac{1}{n}\sum_{i=1}^{n} \log(1 + e^{-y_i h_{\boldsymbol{\theta}}(x_i)}) \\
&= \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{1 + e^{y_i h_{\boldsymbol{\theta}}(x_i)}}{e^{y_i h_{\boldsymbol{\theta}}(x_i)}}\right) = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{e^{y_i h_{\boldsymbol{\theta}}(x_i)}}{1 + e^{y_i h_{\boldsymbol{\theta}}(x_i)}}\right) \\
&= -\frac{1}{n}\sum_{i=1}^{n} \left\{ y_i h_{\boldsymbol{\theta}}(x_i) - \log\left(1 + e^{y_i h_{\boldsymbol{\theta}}(x_i)}\right)\right\};\ y_i \in \{-1, +1\} \\
&= -\frac{1}{n}\begin{cases} \sum_{i=1}^{n} \left\{ y_i h_{\boldsymbol{\theta}}(x_i) - \log\left(1 + e^{h_{\boldsymbol{\theta}}(x_i)}\right)\right\} & ; y_i = 1 \\ \sum_{i=1}^{n} \left\{ -h_{\boldsymbol{\theta}}(x_i) - \log\left(1 + e^{-h_{\boldsymbol{\theta}}(x_i)}\right)\right\} & ; y_i = -1 \end{cases} \\
&= -\frac{1}{n}\begin{cases} \sum_{i=1}^{n} \left\{ y_i h_{\boldsymbol{\theta}}(x_i) - \log\left(1 + e^{h_{\boldsymbol{\theta}}(x_i)}\right)\right\} & ; y_i = 1 \\ \sum_{i=1}^{n} \left\{ -\log\left(1 + e^{h_{\boldsymbol{\theta}}(x_i)}\right)\right\} & ; y_i = -1 \end{cases} \\
&= -\frac{1}{n}\log L(\boldsymbol{\theta})
\end{aligned}
$$

$\hookrightarrow$ With the logistic loss, the conditional ERM coincides with conditional MLE.

Multi-class logistic regression



Multi-class logistic regression I

- $\boldsymbol{X} \in \mathcal{X} = \mathbb{R}^d$ and $Y_i \in \mathcal{Y} = \{1, \cdots, K\}$
- Conditional (Discriminative) model : for $k = 1, \cdots, K-1$

$$\mathbb{P}(Y = k|\mathbf{x}; \boldsymbol{\theta}) = \frac{\exp(\alpha_k + \boldsymbol{\beta}_k^T \mathbf{x})}{1 + \sum_{\ell=1}^{K-1} \exp(\alpha_\ell + \boldsymbol{\beta}_\ell^T \mathbf{x})} = \pi_k(\boldsymbol{x}_i; \boldsymbol{\theta})$$

- for $k = K$, $\mathbb{P}(Y = K|\mathbf{x}; \boldsymbol{\theta}) = 1 - \sum_{k=1}^{K-1} \mathbb{P}(Y = k|\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{1 - \sum_{\ell=1}^{K-1} \exp(\alpha_\ell + \boldsymbol{\beta}_\ell^T \mathbf{x})}$.
  This is equivalent to setting $(\alpha_K, \boldsymbol{\beta}_K^T)^T = \mathbf{0}$.
- Link function : for $k = 1, \cdots, K$

$$\log \frac{\mathbb{P}(Y = k|\mathbf{x}; \boldsymbol{\theta})}{\mathbb{P}(Y = K|\mathbf{x}; \boldsymbol{\theta})} = \alpha_k + \boldsymbol{\beta}_k^T \mathbf{x}$$

Multi-class logistic regression II

- The model parameter : $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_K)$ with $\boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\beta}_k^T)^T$ $(k = 1, \cdots, K-1)$
- Maximum conditional likelihood estimation :The conditional log-likelihood of $\boldsymbol{\theta}$

$$\begin{aligned} L(\boldsymbol{\theta}) &= \log \prod_{i=1}^{n} \mathbb{P}(Y_i|\mathbf{x}_i;\boldsymbol{\theta}) = \log \prod_{i=1}^{n}\prod_{k=1}^{K} \mathbb{P}(Y_i = k|\mathbf{x}_i;\boldsymbol{\theta})^{y_{ik}} \\ &= \sum_{i=1}^{n}\sum_{k=1}^{K} y_{ik} \log \pi_k(\boldsymbol{x}_i;\boldsymbol{\theta}) \end{aligned}$$

where we have used the notation $y_{ik} = \mathbb{1}_{y_i \neq k}$, i.e. $y_{ik} = 1$ iff $y_i = k$

- This log-likelihood is convex but can not be maximized in a closed form.
- The Newton-Raphson (NR) algorithm :
$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}}$

Newton-Raphson for Multi-class LR

- The Newton-Raphson algorithm is an iterative numerical optimization algorithm
- starts from an initial arbitrary solution $\boldsymbol{\theta}^{(0)}$, and updates the estimation of $\boldsymbol{\theta}$
- A single NR update is given by :

$$\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]^{-1} \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad (9)$$

  where the Hessian and the gradient of $\mathcal{L}(\boldsymbol{\theta})$ (which are respectively the second and first derivative of $\mathcal{L}(\boldsymbol{\theta})$) are evaluated at $\boldsymbol{\theta} = \boldsymbol{\theta}^{(t)}$.
- NR can be stopped when the relative variation of $\mathcal{L}(\boldsymbol{\theta})$ is below a prefixed threshold.

IRLS for Multi-class logistic regression I

- **Gradient vector** : $\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \Big((\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_1})^T, \ldots, (\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_{K-1}})^T \Big)^T$ where $\forall k \in [K-1]$:

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_k} = \sum_{i=1}^{n} \big(y_{ik} - \pi_k(\mathbf{x}_i;\boldsymbol{\theta}) \big) \boldsymbol{x}_i = \mathbf{X}^T (\mathbf{y}_k - \mathbf{p}_k)$$

i) $\mathbf{X} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n)^T$: $n \times (d+1)$ matrix whose rows are the inputs $\mathbf{x}_i$,

ii) $\mathbf{y}_k = (y_{1k}, \ldots, y_{nk})^T$: $n \times 1$ vector of indicator variables y_{ik}

iii) $\mathbf{p}_k = (\pi_k(\mathbf{x}_1;\boldsymbol{\theta}), \ldots, \pi_k(\mathbf{x}_n;\boldsymbol{\theta}))^T$: $n \times 1$ vector of logistic probabilities

- Vectorized form of the gradient of $\mathcal{L}(\boldsymbol{\theta})$ for all the logistic components :

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \begin{pmatrix} \mathbf{X}^T & 0 & \ldots & 0 \\ 0 & \mathbf{X}^T & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \mathbf{X}^T \end{pmatrix} \begin{pmatrix} \mathbf{y}_1 - \mathbf{p}_1 \\ \mathbf{y}_2 - \mathbf{p}_2 \\ \vdots \\ \mathbf{y}_{K-1} - \mathbf{p}_{K-1} \end{pmatrix} = \widetilde{\mathbf{X}}^T (\mathbf{Y} - \mathbf{P}) \qquad (10)$$

i) $\mathbf{Y} = (\mathbf{y}_1^T, \ldots, \mathbf{y}_{K-1}^T)^T$: $n \times (K-1)$ column vector

ii) $\mathbf{P} = (\mathbf{p}_1^T, \ldots, \mathbf{p}_{K-1}^T)^T$: $n \times (K-1)$ column vector

iii) $\widetilde{\mathbf{X}} = (\mathbf{X}^T, \ldots, \mathbf{X}^T)^T$: $(n \times (K-1))$ by $(d+1)$ matrix of $K-1$ copies of $\mathbf{X}$.

IRLS for Multi-class logistic regression II

- **Hessian matrix** : composed of $(K-1)\times(K-1)$ block matrices $\left\{\frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_\ell\partial\boldsymbol{\theta}_k^T}\right\}_{k,\ell=1}^{K-1}$

$$\frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}^T} = \begin{pmatrix} \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_1^T} & \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_2^T} & \cdots & \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_{K-1}^T} \\ \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_2\partial\boldsymbol{\theta}_1^T} & & \cdots & \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_{K-1}^T} \\ \vdots & \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_\ell\partial\boldsymbol{\theta}_k^T} & & \vdots \\ \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_{K-1}\partial\boldsymbol{\theta}_1^T} & & \cdots & \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_{K-1}\partial\boldsymbol{\theta}_{K-1}^T} \end{pmatrix}$$

where each block matrix is of dimension $(d+1)\times(d+1)$ and is given by :

$$\begin{aligned} \frac{\partial^2\mathcal{L}(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}_\ell\partial\boldsymbol{\theta}_k^T} &= -\sum_{i=1}^{n}\pi_k(\mathbf{x}_i;\boldsymbol{\theta})\left(\delta_{k\ell}-\pi_\ell(\mathbf{x}_i;\boldsymbol{\theta})\right)\boldsymbol{x}_i\boldsymbol{x}_i^T \\ &= -\mathbf{X}^T\mathbf{W}_{k\ell}\mathbf{X} \end{aligned}$$

i) $\mathbf{W}_{k\ell}$: $n\times n$ diagonal matrix whose diagonal elements are $\pi_k(\mathbf{x}_i;\boldsymbol{\theta})\left(\delta_{k\ell}-\pi_\ell(\mathbf{x}_i;\boldsymbol{\theta})\right)$ for $i=1,\ldots,n$.

IRLS for Multi-class logistic regression III

- For all the logistic components $(k, \ell = 1, \ldots, K-1)$, the Hessian takes the form :

$$\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = -\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}} \quad (11)$$

  $\rightarrow \mathbf{W}$: $(n \times (K-1))$ by $(n \times (K-1))$ matrix composed of $(K-1)) \times (K-1))$ block matrices, each block is $\boldsymbol{\theta}_{k\ell}$ $(k, \ell = 1, \ldots, K-1)$.

- It can be shown that the Hessian matrix for the multi-class logistic regression model is positive semi definite and therefore the log-likelihood is concave.

$$\begin{aligned}
\mathbf{H} &= -\begin{pmatrix} \mathbf{X}^T\mathbf{W}_{1,1}\mathbf{X} & \ldots & \mathbf{X}^T\mathbf{W}_{1,K-1}\mathbf{X} \\ \vdots & \ddots & \vdots \\ \mathbf{X}^T\mathbf{W}_{K-1,1}\mathbf{X} & \ldots & \mathbf{X}^T\mathbf{W}_{K-1,K-1}\mathbf{X} \end{pmatrix} \\
&= -\begin{pmatrix} \mathbf{X}^T & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \mathbf{X}^T \end{pmatrix} \begin{pmatrix} \mathbf{W}_{1,1} & \ldots & \mathbf{W}_{1,K-1} \\ \vdots & \ddots & \vdots \\ \mathbf{W}_{K-1,1} & \ldots & \mathbf{W}_{K-1,K-1} \end{pmatrix} \begin{pmatrix} \mathbf{X} & \ldots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \ldots & \mathbf{X} \end{pmatrix} \\
&= -\widetilde{\mathbf{X}}^T \mathbf{W} \widetilde{\mathbf{X}}
\end{aligned}$$

IRLS for Multi-class logistic regression IV

The NR algorithm in this case can therefore be reformulated as the IRLS

$$\begin{aligned}
\boldsymbol{\theta}^{(t+1)} &= \boldsymbol{\theta}^{(t)} - \left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\right]^{-1}_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\Big|_{\boldsymbol{\theta}=\boldsymbol{\theta}^{(t)}} \\
&= \boldsymbol{\theta}^{(t)} + (\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T (\mathbf{Y} - \mathbf{P}^{(t)}) \\
&= (\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \widetilde{\mathbf{X}})^{-1} \left[\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \widetilde{\mathbf{X}} \boldsymbol{\theta}^{(t)} + \widetilde{\mathbf{X}}^T (\mathbf{Y} - \mathbf{P}^{(t)})\right] \\
&= (\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \mathbf{X})^{-1} \widetilde{\mathbf{X}}^T \left[\mathbf{W}^{(t)} \widetilde{\mathbf{X}} \boldsymbol{\theta}^{(t)} + (\mathbf{Y} - \mathbf{P}^{(t)})\right] \\
&= (\widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \widetilde{\mathbf{X}})^{-1} \widetilde{\mathbf{X}}^T \mathbf{W}^{(t)} \tilde{\mathbf{Y}}
\end{aligned}$$

where $\tilde{\mathbf{Y}} = \widetilde{\mathbf{X}} \boldsymbol{\theta}^{(t)} + (\mathbf{W}^{(t)})^{-1} (\mathbf{Y} - \mathbf{P}^{(t)})$ which yields in the Iteratively Reweighted Least Squares (IRLS) algorithm.

Tasks :

- Implement (from the scratch) each of the following functions and apply them to the given data :
  - `train_reglog` and `predict_reglog`
  - `irls` should be in a separate function

  Datasets :
  - Training data Xtrain.txt and ytrain.txt
  - Testing data : Xtest.txt
- Plot the results by highlighting the classification and the generative model for each class
- compare your results to those you could obtain by using standard packages

  `from sklearn.linear_model import LogisticRegression`

  or `GLM` from `statsmodels`