Exercice 1 : Analyse Discriminante Probabiliste

On considère un échantillon de n couples de données observées $\mathcal{D} = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ issue de K groupes (classes) homogènes où chaque individu $\boldsymbol{x}_i$, $i = 1, \ldots, n$ est décrit par d variables x_{ij}, $j = 1, \ldots, d$. La variable y_i représente la classe du $i$ème individu et est à valeurs dans $\{1, \ldots, K\}$: $y_i = k \in \{1, \ldots, K\}$. On suppose que les données de chaque classe k sont générées selon une loi normale $(\boldsymbol{\mu}_k, \boldsymbol{\Sigma})$ où la matrice $\boldsymbol{\Sigma}$ est supposée la même pour toutes les classes.

1. Quel est le modèle conditionnel $\mathbb{P}(Y|\boldsymbol{X} = \boldsymbol{x})$ induit par ce modèle génératif ?
2. Montrer qu'il correspond à un modèle probabiliste de classification linéaire (frontières de décision linéaires entres les classes)
3. Quelles frontières de décision obtient-on si l'on autorise chaque classe à avoir sa propre matrice de covariance ?
4. Appliquez le modèle aux données en utilisant la méthode du maximum de vraisemblance pour estimer ses paramètres

Exercice 2 : Analyse Factorielle Discriminante

On considère les données des départements français. Les données sont disponibles dans le fichiers `depart.data`.

On se propose de mettre en evidence les plus grandes disparités inter-régionales et donc de rechercher les variables ou combinai- sons de variables expliquant au mieux le decoupage regional. Pour simplifier, nous procedons a des regroupements afin de construire des regions moins nombreuses comprenant des nombres de départements plus semblables. D ?autre part, la region "Ile de France", trop particulière et donc trop facilement discriminée, est laissée a part.

Les donnees des departements franc ?ais Les donnees proviennent du Groupe d ?Etude et de Reflexion Inter-Regional (GERI). Elles decrivent, quatre grands themes : la demographie, l ?emploi, la fiscalite directe locale, la criminalite, de chacun des departements franc ?ais m etropolitains et de la Corse. Les indicateurs ont ete observes pendant l ?annee 1990, ils sont, pour la plupart, des taux calcules relativement a la population totale du departement concerne. Voici leur liste :

— identificateur : numero du departement,
— identificateur : code du departement,
— identificateur : code de la region
— `URBR` : indicateur de concentration de la population mesurant le caractere urbain ou rural diun departement,
— `TXCR` : taux de croissance de la population sur la p eriode intercensitaire 1882-1990,
— `JEUN` : part des 0-19 ans dans la population totale,
— `AGE` : part des plus de 65 ans dans la population totale,
— `FE90` : taux de fecondite (pour 1000) egal au nombre de naissances rapport es au nombre de femmes agees de 15 a 49 ans en moyenne triennale,
— `ETRA` : part des etrangers dans la population totale,
— `CHOM` : taux de chomage,
— `CRIM` : taux de criminalit e : nombre de d elits par habitant,
— `FISC` : produit, en francs constants 1990 et par habitant des quatre taxes directes locales (profes- sionnelle, habitation, foncier bati, foncier non bati). Suivent ensuites les parts de chaque profession et categorie socioprofessionnelle (PCS) dans la population active occupee du departement :
— `AGRI` : agriculteurs,
— `ARTI` : artisans,
— `CADR` : cadres sup erieurs,

— `EMPL` : employ es,
— `OUVR` : ouvriers,
— `PROF` : professions interm ediaires

— Charger les données
— Afficher les
— Regrouper les en régions : Nd (Nord) ; Es (Est) ; Ws (West) ; CN (Centre Nord) ; Centre West (CW) ; Centre Est (CE) ; Sud West (SW) ; Sud Est (SE). (vous pouvez utiliser `select`)
— Appliquer une analyse factorielle discriminante et analyser les résultats

Exercice 2 : Analyse Discriminante décisionnelle

On considère d'infarctus disponible dans le fichier `infarctu.txt`.

Ce jeu de données provient d'une étude épidémiologique où le but de l'étude était d'évaluer l'existence d'un risque plus élevé de survenue d'un infarctus du myocarde chez les femmes qui utilisent ou ont utilisé des contraceptifs oraux.

101 observations et 8 variables. FRCAR Frequence Cardiaque (i.e. heart rate)

INCAR Index Cardique (cardiac index)

INSYS Index Systolique (systolic index)

PRDIA Pression Diastolique (diastolic pressure)

PAPUL Pression Arterielle Pulmonaire (pulmonary artery pressure)

PVENT Pression Ventriculaire (ventricular pressure)

REPUL Resistance Pulmonaire (pulmonary resistance)

PRONO Pronostic (prognosis) : a factor with levels dead and survive

— Charger les données
— Afficher les
— Appliquer une analyse discriminante linéaire et analyser les résultats
— Appliquer une analyse discriminante quadratique et analyser les résultats
— Dans chacun des deux cas précédents, considérer le cas où les proportions des classes sont égales ou non
— utiliser une validation croisée (cross-validation en anglais) pour évaluer les capacités en généralisation de l'analyse discriminante quadratique pour ce jeux de données.