Mastère M2 MVA 2012/2013 - Modèles graphiques

Exercices à rendre pour le 17 Octobre 2012.

Ces exercices peuvent s'effectuer par groupe de deux élèves.

Si vous souhaitez nous envoyer votre devoir au format pdf, merci de le nommer `MVA_DM1_<votre nom>.pdf` ou `MVA_DM1_<nom1>_<nom2>.pdf`

1 Apprentissage dans les modèles discrets

On considère le modèle suivant où z et x sont des variables discrètes pouvant prendre respectivement M et K valeurs : $p(z = m) = \pi_m$, $p(x = k|z = m) = \theta_{mk}$.

Calculer l'estimateur du maximum de vraisemblance de π et θ à partir de données i.i.d. (On attend le calcul lui-même et pas seulement la solution).

2 Classification linéaire

Les fichiers `classificationA.train`, `classificationB.train` et `classificationC.train` contiennent des ensembles de données (x_n, y_n) où $x_n \in \mathbb{R}^2$ et $y_n \in \{0, 1\}$ (chaque ligne est composee des 2 composantes de x_n puis de y_n). Le but de cet exercice est d'implémenter les méthodes de classification linéaire et de les tester sur ces 3 jeux de données. Le langage de programmation est libre (MATLAB, Octave, Scilab ou R sont néanmoins recommandés). Le code source doit être remis avec les résultats. On demande néanmoins une présentation rédigée des résultats indépendante du code.

1. Modèle génératif (LDA). Etant donnée la classe, les données sont normales avec des moyennes différentes et la même matrice de covariance :

$$y \sim \text{Bernoulli}(\pi), \quad x|\{y = i\} \sim \text{Normale}(\mu_i, \Sigma).$$

   (a) Faire le calcul de l'estimateur du maximum de vraisemblance pour ce modèle. *Indication* : le modèle a été vu en cours mais pas le calcul du maximum de vraisemblance. On pourra s'inspirer de la Section 7.2 du polycopié (qui traîte le cas où Σ est diagonale).

   (b) Quelle est la forme de la loi conditionelle $p(y = 1|x)$? Comparer à la régression logistique.

   (c) Implémenter l'estimateur du maximum de vraisemblance pour ce modèle et l'appliquer aux données. Représenter graphiquement les données comme un nuage de points dans $\mathbb{R}^2$ ainsi que la droite définie par

$$p(y = 1|x) = 0.5\,.$$

2. Régression logistique : implémenter la régression logistique (pour une fonction affine $f(x) = w^\top x + b$) en utilisant l'algorithme IRLS (Newton-Raphson) décrit en cours et dans le polycopié (ne pas oublier le terme constant).
   (a) Donner les valeurs numériques des paramètres appris.
   (b) Représenter graphiquement les données comme un nuage de points dans $\mathbb{R}^2$ ainsi que la droite définie par

$$p(y = 1|x) = 0.5\,.$$

3. Régression linéaire : en considérant la classe y comme variable réelle prenant les valeurs 0 et 1, implémenter la régression linéaire (pour une fonction affine $f(x) = w^\top x + b$) par résolution de l'équation normale.
   (a) Donner les valeurs numériques des paramètres appris.
   (b) Représenter graphiquement les données comme un nuage de points dans $\mathbb{R}^2$ ainsi que la droite définie par

$$p(y = 1|x) = 0.5\,.$$

4. Les données contenues dans les fichiers `classificationA.test`, `classificationB.test` et `classificationC.test` ont respectivement été générées par la même distribution que les données dans les fichiers `classificationA.train`, `classificationB.train` et `classificationC.train`. Tester les différents modèles sur ces données de test. En particulier :
   (a) Calculer pour chaque modèle le taux d'erreur sur les données d'entraînement et calculer le taux d'erreur sur les données de test.
   (b) Comparer les résultats des différentes méthodes sur chaque jeu de données. Le taux d'erreur est-il plus fort, plus faible ou similaire sur les données d'apprentissage et les données de test ? Pourquoi ? Quelles méthodes donnent des résultats semblables/très différents ? Quelles méthodes donnent de meilleurs résultats sur les différents jeu de données ? Interpréter.
5. Modèle génératif (QDA). Etant donnée la classe, les données sont normales avec des moyennes et des matrices de covariance différentes :

$$y \sim \text{Bernoulli}(\pi), \quad x|y = i \sim \text{Normale}(\mu_i, \Sigma_i).$$

   Implémenter le maximum de vraisemblance pour ce modèle et l'appliquer aux données.
   (a) Donner les valeurs numériques des paramètres appris.
   (b) Représenter graphiquement les données ainsi que la conique définie par

$$p(y = 1|x) = 0.5\,.$$

   (c) Calculer les taux d'erreur du modèles QDA sur les données d'entraînement et de test.
   (d) Commenter les résultats comme précédemment.