TD 7 : ANALYSE DISCRIMINANTE LINÉAIRE

COURS D'APPRENTISSAGE, ECOLE NORMALE SUPÉRIEURE, AUTOMNE 2014

Rémi Lajugie
remi.lajugie@ens.fr

Résumé. Ce TP/TD passe en revue divers modèles abordés en cours dans le but de les appliquer sur des données synthétiques puis réelles. On cherchera en particulier à implémenter un modèle génératif simple mais néanmoins robuste en l'ajustant via le principe du maximum de vraisemblance. On essaiera de voir ensuite sa capacité à généraliser et le comparera avec une méthode simple de classification par moyennage local.

Vous devrez rendre un compte rendu pour la partie théorique de ce TP/TD d'ici le Vendredi 12/12/2014 à 23 :59 par mail au format PDF à `remi.lajugie@ens.fr`. Joignez y une version du code que vous aurez produit qui soit : 1) lisible (indentée et aérée) et 2) simple à lancer.

Remarque préliminaire : Les comptes-rendus comportant de belles figures permettant de bien visualiser les résultats seront valorisés.

Problème

Commencez par télécharger sur la page web du cours les jeux de données :
- `di.ens.fr/~slacoste/teaching/apprentissage-fall2014/TP/classificationA.train`,
- `di.ens.fr/~slacoste/teaching/apprentissage-fall2014/TP/classificationA.test`,
- `di.ens.fr/~slacoste/teaching/apprentissage-fall2014/TP/classificationB.train`,
- `di.ens.fr/~slacoste/teaching/apprentissage-fall2014/TP/classificationB.test`,
- `di.ens.fr/~slacoste/teaching/apprentissage-fall2014/TP/classificationC.train`,
- `di.ens.fr/~slacoste/teaching/apprentissage-fall2014/TP/classificationC.test`,

sur la page web du cours. Les fichiers sont constitués de trois colonnes. Les deux premières donnent les coordonnées de points d'entrées $x_i \in \mathbb{R}^2$. La troisième est la donnée de sortie correspondante $y_i \in \{0, 1\}$. A partir des données d'entrainement des fichiers `.train`, on veut construire des classifieurs qui seront efficaces sur les données des fichiers `.test`.

Génération des données.
Le premier jeu de données est issu d'un modèle où les points de chaque classe sont générés suivant des loi Gaussiennes de variance commune entre classe mais de moyennes différentes, i.e, $\mathbb{P}(X|Y = i)$ est une distribution Gaussienne de moyenne μ_i et de matrice de covariance commune Σ. Pour le deuxième jeu de données, le modèle de génération est le même mais cette fois les matrices de covariances aussi sont différentes. Concernant, le troisième jeu de données, une des classes est le mélange de deux Gaussiennes tandis que l'autre correspond à une seule Gaussienne.

Modèles de régression.

1) Pour chacun des ensembles d'apprentissages A, B et C, représentez les données sous forme d'un nuage bicolore de points en 2D.

2) Calculez le prédicteur de classification obtenu par plug-in du prédicteur de la régression linéaire.

3) Implémentez le calcul du prédicteur de classification obtenu par régression logistique en utilisant soit une méthode de descente de gradient, soit l'algorithme des moindres carrés repondérés.

Discriminant de Fisher.

On propose un autre modèle de classification dit *génératif*. On modélise les données de chaque classe comme un nuage gaussien de distribution $\mathcal{N}(\mu_1, \Sigma)$ pour la classe 1 (i.e., $Y = 1$) et $\mathcal{N}(\mu_0, \Sigma)$ pour la classe 0 (i.e., $Y = 0$), où la matrice de covariance Σ est supposée la même pour les deux classes.

4) a) Quel est le modèle conditionnel de $\mathbb{P}(Y|X = x)$ induit par ce modèle génératif ?

b) Montrez qu'il correspond à un modèle probabiliste de classification linéaire de la même forme que la régression logistique.

c) Appliquez le modèle aux données en utilisant le principe du maximum de vraisemblance pour apprendre les paramètres du modèle génératif.

d) Quelles frontières de classification obtient-on si l'on autorise les matrices de covariances à être différentes pour les deux classes ? On appelle cette méthode QDA (Quadratic Discriminant Analysis). Appliquez le modèle aux données en utilisant le principe du maximum de vraisemblance pour apprendre les paramètres du modèle génératif.

5) Pour chacun de ces quatre classifieurs (régression linéaire, régression logistique, LDA et QDA) et pour chacun des ensembles d'entraînement, représentez sur une figure le nuage de points des données et la frontière de séparation entre les deux classes correspondant au classifieur appris.

6) Calculez le taux d'erreur en classification, c'est-à-dire le risque empirique pour la perte $0-1$, sur les données d'entraînement d'une part et sur les données de test d'autre part pour ces quatre classifieurs.

7) Comparez les taux d'erreurs sur les données d'entraînement et sur les données de test. Comparer les performances relatives des différents algorithmes de classification en fonction de la structure des données A, B et C.

8) Sur ces mêmes jeux de données, implémentez une méthode des k plus proches voisins (en ajustant par validation croisée le paramètre), comparez les performances avec celles des modèles de la question précédente.

9) Reprenez les modèles génératifs de la partie sur le discriminant de Fisher et appliquez les aux données MNIST déjà utilisées lors des précédents TPs (disponibles sur la page du cours `www.di.ens.fr/~slacoste/teaching/apprentissage-fall2014/TP/mnist_digits.mat`). Comparez avec les méthodes de plus proches voisins vus au TP 2. En comparant les erreurs de test, discutez des performances relatives des modèles génératifs et discriminatifs sur cette tâche de classification.