MAP553 Projet 1

Apprentissage semi-supervisé

année 2014-2015

Les groupes pour ce projet seront de 2 personnes.

Le projet se compose de deux parties : une première partie sur l'apprentissage supervisé par SVM (application directe du cours) et une seconde partie plus créative sur l'apprentissage semi-supervisé.

Pour réaliser votre projet, vous pouvez utiliser `R` (http://cran.r-project.org/), Scilab, Matlab, Python, C, etc. Vous me rendrez deux fichiers : un fichier pdf incluant les codes, les sorties, et les explications nécessaires pour comprendre votre démarche et un fichier "code" (code.R ou code.m) contenant votre code "opérationnel" (c'est à dire que lorsque je le lancerai il marchera sans intervention de ma part et fournira des résultats complets). Vos codes seront annotés de façon à être lisibles facilement (description des fonctions, entrées / sorties, etc).

Vous allez travailler sur des données publiques, proposées dans le cadre du challenge "active learning"
http://www.causality.inf.ethz.ch/activelearning.php

Données HIVA

"HIVA is a chemoinformatics dataset. The task of HIVA is to predict which compounds are active against the AIDS HIV infection. The original data has 3 classes (active, moderately active, and inactive). We brought it back to a two-class classification problem (active vs. inactive). We represented the data as 1617 sparse binary input variables. The variables represent properties of the molecule inferred from its molecular structure. The problem is therefore to relate structure to activity (a QSAR=quantitative structure-activity relationship problem) to screen new compounds before actually testing them (a HTS=high-throughput screening problem). The original data were made available by The National Cancer Institute (USA). The 3d molecular structure was obtained by the CORINA software and the features were derived with the ChemTK software."

Format : le jeu de données est constitué de 2 fichiers : un fichier "data" qui contient les variables explicatives (features) et un fichiers "label" qui contient les labels (-1 ou 1). La taille de l'échantillon est de $n = 42678$ et le nombre de features est $p = 1617$.

data format txt : (il faut désarchiver le fichier pour récupérer hiva.data)
http://www.cmap.polytechnique.fr/~giraud/MAP553/Projets/hiva.zip
data format matlab :

http://www.cmap.polytechnique.fr/~giraud/MAP553/Projets/hiva.mat
labels format txt :
http://www.cmap.polytechnique.fr/~giraud/MAP553/Projets/hiva.label

Conseils pour le développement : dans la phase de "développement", travaillez avec un sous-échantillon des données pour réduire les temps de calculs. Une fois que votre code et algorithme sont au point, vous pouvez alors le confronter aux données complètes.

Si vous avez des problèmes de gestion de ma mémoire, réduisez le taille du fichier data en ne conservant que les 500 premières colonnes.

1 Apprentissage supervisé par SVM

1. Première étape : séparer les données en deux parties "training" et "test". La matrice `TrainingData` contiendra les 5000 premières lignes de la matrice "data" et la matrice `TrainingLabels` les 5000 premières entrées du vecteur "label". Les données restantes seront stockées dans `TestData` et `TestLabels`.
2. Calculer un classifieur SVM (éventuellement à noyaux) à partir de `TrainingData` et `TrainingLabels`. La sélection des paramètres, du noyau, etc, peut être réalisée à l'aide d'une 10-fold Cross-Validation.
3. Notons $h_{training}$ le classifieur calculé ci-dessus. Appliquer ce classifieur à `TestData` et évaluer le pourcentage de données "test" mal classifiées (en comparant les labels prédits par $h_{training}$ aux vrais labels `TestLabels`).
4. Rafinez ensuite votre évaluation en regardant d'une part le pourcentage de données de label "1" étiquétées comme "1" par le classifieur et d'autre part le ratio entre le nombre de données étiquetées "1" à tort par le classifieur et le nombre de toutes les données étiquetées "1" par le classifieur (FDR).

Vous pouvez travailler avec le logiciel de votre choix. Vous pouvez consulter cet article : http://www.jstatsoft.org/v15/i09/paper (en particulier si vous choisissez de travailler avec `R`) et vous trouverez quelques indications pratiques ci-dessous.

Quelques astuces sous `R`

Chargement des données : pour charger les données dans `R`, il faut utiliser la commande `read.table`. Exemple :

```
features <- read.table("hiva.data")
labels <- read.table("hiva.label")
labels <- as.vector(labels[,])
features <- as.matrix(features)
```

Enfin, tapez `?read.table` pour avoir plus d'explications.

2 Apprentissage semi-supervisé

Obtenir le label (-1 ou 1) de données est souvent couteux, difficile et requiert un gros effort d'annotation humaine. Il est par contre souvent peu couteux d'obtenir des données sans-label.

Le but de l'apprentissage semi-supervisé est d'utiliser simultanément des données avec label (peu abondantes) et des données sans label (abondantes) pour construire des classifieurs. Nous allons utiliser les mêmes données que précédemment. La différence, c'est que seule une partie des labels va être utilisée pour construire le classifieur.

1. Selon vous, quel profit peut-on tirer des données sans labels ?
2. Proposez une (ou plusieurs) démarche(s) pour tirer profit de cet avantage.
3. Construire le vecteur `PartialTrainingLabels` en ne conservant que les 1000 premières entrées du vecteur `TrainingLabels` construit précédemment.
4. Proposer un algorithme original construisant un classifieur uniquement à partir de la matrice `TrainingData` et du vecteur `PartialTrainingLabels` (soyez créatifs !)
5. Notons $\hat{h}$ le classifieur calculé ci-dessus. Appliquer ce classifieur à `TestData` et évaluer le pourcentage de données "test" mal classifiées (en comparant les labels donnés par $\hat{h}$ aux vrais labels `TestLabels`).
6. Comparez la performance de votre classifieur à celle d'un SVM appliqué aux seules données avec labels. Faites-vous mieux ?

Dans le document que vous me rendrez, vous expliquerez en détails votre algorithme, vous motiverez votre choix et vous donnerez en plus du code le pourcentage de données "test" mal classifiées. Si vous avez essayé plusieurs méthodes, vous expliquerez vos différents essais.

Notation

La note finale se répartit selon les proportions suivantes
– 1/2 pour la qualité de l'analyse statistique effectuée. Toute approche originale sera considérée positivement.
– 1/6 pour l'originalité de l'approche proposée
– 1/3 pour la qualité du rapport : explications, profondeur d'analyse, mise en perspective, qualité des sorties graphiques, pertinence du choix des résultats montrés, analyse critique des résultats.

Référence

Vous pouvez vous appuyer par exemple sur le livre "The Element of Statistical Learning" de Hastie, Tibshirani et Friedman pour développer vos analyses :
http://www-stat.stanford.edu/~tibs/ElemStatLearn/