Premiers algorithmes

Web-Mining - Apprentissage statistique

Mardi 18 novembre 2011

Perceptron

1. Ecrire une procédure qui ajuste perceptron monocouche comme décrit dans la section 11.3 du HTF.
2. Générer des données séparables dans $\mathbb{R}^2$, et illustrer graphiquement la convergence de l'algorithme. Essayer avec un exemple de données non séparables.
3. Montrer que, si les données $(x_i, y_i)_{1 \leq i \leq n}$ sont linéairement séparables, et si $z_i = (x_i, 1)/\|(x_i, 1)\|$, il existe $\beta^* \in \mathbb{R}^2$ tel que pour tout $i \in \{1, \ldots, n\}$, $y_i \langle \beta^*, z_i \rangle \geq 1$. En déduire que chaque mise à jour $\beta_{t+1} \leftarrow \beta_t + y_i z_i$ d'un point (x_i, y_i) mal classé est telle que $\|\beta_{t+1} - \beta^*\|^2 \leq \|\beta_t - \beta^*\|^2 - 1$, et conclure sur la convergence de l'algorithme.

Analyse Discriminante Linéaire

On considère deux populations gaussiennes dans $\mathbb{R}^d$ ayant la même structure de covariance. On observe des points dans le mélange de ces deux populations. On note X les vecteurs aléatoires dans $\mathbb{R}^d$ et Y leur étiquette qui vaut +1 ou -1 selon la population d'appartenance. Les lois conditionnelles de X sachant $Y = +1$ (respectivement $Y = -1$) sont des gaussiennes multivariées $\mathbb{N}_d(\mu_+, \Sigma)$ (respectivement $\mathbb{N}_d(\mu_-, \Sigma)$). On notera leur densités respectives f_+ et f_-. On note également $p = \mathbb{P}\{Y = +1\}$. On rappelle que la densité de la loi $\mathbb{N}_d(\mu, \Sigma)$ est donnée par :

$$f(x) = \frac{1}{(2\pi)^{d/2}\sqrt{det(\Sigma)}} \exp\left\{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right\} .$$

et que la matrice de covariance d'un vecteur aléatoire X est définie par $\Sigma = \mathbb{E}\big((X - \mathbb{E}(X))(X - \mathbb{E}(X))^T\big)$.

1. En utilisant la formule de Bayes donner la formule des probabilités a posteriori : $\mathbb{P}\{Y = +1 \mid X = x\}$, $\mathbb{P}\{Y = -1 \mid X = x\}$, comme fonctions de f_+, f_- et p.
2. Exprimer le log-ratio des deux classes :

   $$\log\left(\frac{\mathbb{P}\{Y = +1 \mid X = x\}}{\mathbb{P}\{Y = -1 \mid X = x\}}\right)$$

   en fonction de p, μ_+, μ_- et Σ.
3. On dispose à présent d'un échantillon de ce mélange et on suppose que p, μ_+, μ_- et Σ sont des paramètres inconnus. On suppose que l'échantillon considéré contient n observations notées $(x_1, y_1), \ldots, (x_n, y_n)$ et que $\sum_{i=1}^n \mathbb{I}\{y_i = +1\} = m$. En utilisant la méthode des moments utilisée en estimation paramétrique, proposer des estimateurs $\widehat{p}$, $\widehat{\mu}_+$, $\widehat{\mu}_-$ et $\widehat{\Sigma}$ des paramètres.
4. Justifier le choix du classifieur qui classe +1, toute observation x telle que :

   $$x^T\widehat{\Sigma}^{-1}(\widehat{\mu}_+ - \widehat{\mu}_-) > \frac{1}{2}\widehat{\mu}_+^T\widehat{\Sigma}^{-1}\widehat{\mu}_+ - \frac{1}{2}\widehat{\mu}_-^T\widehat{\Sigma}^{-1}\widehat{\mu}_- + \log(1 - m/n) - \log(m/n) ,$$

   et -1, sinon.
5. On propose de comparer le classifieur précédent à celui obtenu par minimisation du critère des moindres carrés :

   $$\min_{\beta_0 \in \mathbb{R}, \beta \in \mathbb{R}^d} \sum_{i=1}^n (y_i - \beta_0 - \beta^T x_i)^2$$

Ecrire la condition d'annulation du gradient et mettre en évidence le fait que la solution $\widehat{\beta}$ doive satisfaire une équation de la forme :

$$(\alpha\widehat{\Sigma} + \gamma\widehat{\Sigma}_B)\widehat{\beta} = n(\widehat{\mu}_+ - \widehat{\mu}_-)$$

où $\widehat{\Sigma}_B = (\widehat{\mu}_+ - \widehat{\mu}_-)(\widehat{\mu}_+ - \widehat{\mu}_-)^T$ et α, γ sont des facteurs dépendant de n, m, à préciser.

6. Montrer alors que $\widehat{\Sigma}_B\widehat{\beta}$ est porté par la direction $(\widehat{\mu}_+ - \widehat{\mu}_-)$. En déduire que $\widehat{\beta}$ est proportionnel à $\widehat{\Sigma}^{-1}(\widehat{\mu}_+ - \widehat{\mu}_-)$.
7. Trouver $\widehat{\beta}_0$ par la résolution de la minimisation des moindres carrés. Montrer que le classifieur obtenu diffère de celui obtenu dans la question 4, sauf dans le cas où $m = n/2$.

- Application -

Il s'agit ici d'appliquer concrètement la méthode ci-dessus sur des observations simulées puis à une base de données réelles. Dans ce dernier cas, on scindera aléatoirement les données en deux échantillons : un échantillon d'apprentissage (70% des données environ) à partir duquel les paramètres régissant le score seront estimés et un échantillon test (les 30% restant) sur lequel on évaluera la performance de la règle de score précédemment apprise.

1. Appliquer l'analyse discriminante linéaire sur des données d'apprentissage gaussiennes bi-dimensionnelles. Estimer son erreur de prédiction au moyen de l'échantillon test (comparer l'estimation à l'"erreur d'apprentissage"). Tracer la frontière.
2. Mêmes questions avec des données non gaussiennes.
3. Mêmes questions (à l'exception du tracé de la frontière) avec des données issues de la base South African Heart Disease disponible sur le site `http://www-stat.stanford.edu/~tibs/ElemStatLearn/`

K plus proches voisins

Appliquer la méthode $K - NN$ (version multi-classe) aux données issues de la base zipcode disponibles sur le site `http://www-stat.stanford.edu/~tibs/ElemStatLearn/` avec différents choix de $K \geq 1$. Estimer la matrice de confusion $(\mathbb{P}\{C_K(X) = i, Y = j\})_{i,\,j}$ associée au classifieur C_K ainsi obtenu. Proposer une méthode pour choisir K et la mettre en oeuvre.

CART

Appliquer la méthode CART aux données spam disponibles sur le site www-stat.stanford.edu/ElemStatLearn . On rappelle que dans la version originale, à toute région R de l'espace d'entrée, on affecte la prédiction $\arg\max_{k=-1,\,+1} \widehat{p}_k(R)$ avec

$$\widehat{p}_k(R) = (1/N_R) \sum_{i:\ X_i \in N_R} \mathbb{I}\{Y_i = k\}.$$

Mettre en oeuvre CART avec les mesures d'impureté suivantes (à minimiser récursivement) :

- Indice de Gini : $2\widehat{p}_k(R)(1 - \widehat{p}_k(R))$
- Entropie croisée : $-\widehat{p}_k(R)\log(\widehat{p}_k(R)) - (1 - \widehat{p}_k(R))\log(1 - \widehat{p}_k(R))$.