Datamining 2: Classification supervisée Partie 5 : estimation de densité, vers les méthodes bayesiennes

M2 STD

October 15, 2015

Principe

Si jamais on connaissait les densités conditionnelles $f(x|Y=j)$ on aurait une méthode de décision très simple via la formule de Bayes:

$$P(Y=k|X=x) = \frac{f(X=x|Y=k)P(Y=k)}{\sum_k f(X=x|Y=j)P(Y=j)}$$

Et bien sur on choisirait la classe la plus probable

Principe

$$P(Y = k|X = x) = \frac{f(X = x|Y = k)P(Y = k)}{\sum_k f(X = x|Y = j)P(Y = j)}$$

Le probleme de classification devient alors un problème d'estimation de densité classe par classe (l'estimation de $P(Y = j)$ étant un probleme "trivial").
Plusieur méthodes d'estimation de densités

Le cas paramétrique

Bien sur si on connait la “forme” de la densité (par exemple on sait que la densité est gaussienne) on estime les parametres puis on estime la densité. C’est le cas paramétrique (qui est, relativement, simple) Ce qui nous interresse c’est le cas non paramétrique (ie : on ne sait pas a priori quelle est la forme de la densité)

EM (Expectation Maximisation)

L'hypothèse ici est que la loi de X est un mélange de gaussiennes. soit que la densité s'écrive comme:

$$f(x) = \sum_{i=1}^{K} p_i \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-0.5(x-\mu_i)\Sigma_i^{-1}(x-\mu_i)'\right)$$

ou les p_i sont positifs et $\sum p_i = 1$ les μ_i sont des vecteurs de $\mathbb{R}^d$ et les Σ_i des matrices symétriques définies positives de dimension d. On a donc $K + Kd + K((d-1)^2/2)$ parametres a estimer...

EM (Expectation Maxisation)

L'algorithme *EM* propose une méthode de maximisation de la vraisemblance par montée de gradient avec des étapes simples !!! On initialise avec une classification en K classes (par exemple classification hierarchique, mixte, kmeans...) on initialise $\pi_i(j) = P(Y = i|X = X_j)$ par 1 si l'individu j appartient a la classe i et 0 sinon.

EM (Expectation Maxisation)

puis on itére:

- **Etape E: Expectation** actualisation des μ_i, Σ_i et p_i soit :
  1. $\mu_i := \sum_j \pi_i(j) X_j$
  2. $\Sigma_i := \sum_j \pi_i(j)(X_j - \mu_i)'(X_j - \mu_i)$
  3. $p_i = \sum_j \pi_i(j)$
- **Etape M: Maximisation** Calcul des "poids" $\pi_i(j)$ par affectation de chaque individu a sa classe la plus probable d'après la règle de Bayes:

$$\pi_i(j) = \frac{p_i \frac{1}{(2\pi)^{d/2}|\Sigma_i|^{1/2}} \exp\left(-0.5(X_i - \mu_k)\Sigma_i^{-1}(X_j - \mu_i)'\right)}{\sum_{k=1}^{K} p_k \frac{1}{(2\pi)^{d/2}|\Sigma_k|^{1/2}} \exp\left(-0.5(X_j - \mu_k)\Sigma_k^{-1} X_j - \mu_k)'\right)}$$

Estimation de densité par les k_n plus proches voisins

$$\hat{f}_{k_n}(x) = \frac{k_n}{n\omega_d r_{k_n}^d(x)}$$

Ou $r_{k_n}(x)$ est la distance entre x et son $k_n$eme voisin (ie l'observation X_i qui est la $k_n$eme plus proche de x)
Theoreme : si $k_n \to \infty$ et $k_n/n \to 0$ alors on a un estimateur convergent.
Idée : si on regarde la probabilité de tomber dans une boule $\mathcal{B}(x, \rho_n) = f(x)\rho_n^d\omega_d$ (si $\rho_n \to 0$) et donc le nombre moyen de points dans $E(K) = \mathcal{B}(x, \rho_n) = nf(x)\rho_n^d\omega_d$ ici on "fixe le nombre de points" mais on autorise le rayon a etre une variable aléatoire... on "suppose" qu'on peut inverser tout ca et on a une relation de type $k_n = nf(x)\rho_n^d\omega_d$ qui donne la formule précisée...

Estimation de densité par les noyaux (sphérique)

$$\hat{f}_h(x) = \frac{1}{Nh^d} \sum_i K((X_i - x)/h)$$

Ou K est une fonction positive d'integrale 1.
Théorie: si $h \to 0$ et $nh^d \to \infty$ alors on a un estimateur convergent.

Les parametres

Dans les trois cas le choix du parametre:

1. K le nombre de composantes du mélange pour EM
2. k_n le nombre de voisins pour les plus proches voisins
3. h la taille de fenetre pour la densité

est fondamental a la réussite de l'estimation

La taille de fenetre optimale

en dimension 1

$$h_n^{opt} = \left(\frac{\int K^2}{n \left(\int x^2 K \right)^2 \int (f'')^2} \right)^{1/5}$$

démonstration tableau

en dimension d :

$$h_n^{opt} = C(f) n^{-1/(d+4)}$$