

Outlier Detection with One-class Kernel Fisher Discriminant by Volker Roth

Outlier Detection with One-class Kernel Fisher Discriminants

Volker Roth
ETH Zurich, Institute of
Computational Science

Julie Carreau 11 février 2005

Détection de valeurs aberrantes et classification:

- Séparation des observations en deux classes (typiques et atypiques)
- Frontière entre les deux classes est déterminée par un quantile (contour de densité constante)

Détection de valeurs aberrantes et classification:

- Séparation des observations en deux classes (typiques et atypiques)
- Frontière entre les deux classes est déterminée par un quantile (contour de densité constante)

Difficulté :

- Peu ou pas d'exemples étiquetés de la classe "atypique"
- Partiellement non-supervisé : connaissance a priori de la fraction espérée de valeurs aberrantes ==> difficile à justifier

Détection de valeurs aberrantes et classification:

- Séparation des observations en deux classes (typiques et atypiques)
- Frontière entre les deux classes est déterminée par un quantile (contour de densité constante)

Difficulté :

- Peu ou pas d'exemples étiquetés de la classe "atypique"
- Partiellement non-supervisé : connaissance a priori de la fraction espérée de valeurs aberrantes ==> difficile à justifier

Proposition :

- Utilisation d'un **classifieur à noyau** à une classe tel que dans l'espace des *features* induit, les observations aient une densité **Gaussienne**.
- La détection de valeurs aberrantes se réduit à identifier les déviations d'une loi Gaussienne.

Approche basée sur un test des distances de Mahalanobis

$$d(x, \mu) = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

Approche basée sur un test des distances de Mahalanobis

$$d(x, \mu) = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

1) Supposons que $X \sim N(\mu, \Sigma), X \in \mathfrak{R}^d$

- séparation du jeu de données à l'origine : LDA sur $X' = (X, -X)^T$

- la solution *ridge* de LDA est $\beta^*(\gamma)$

- distance de Mahalanobis : $d(x, \mu) = (\beta^* x - m)^2 + D_{\perp}$
 m est la projection de μ par LDA

Approche basée sur un test des distances de Mahalanobis

$$d(x, \mu) = (x - \mu)^T \Sigma^{-1} (x - \mu)$$

1) Supposons que $X \sim N(\mu, \Sigma), X \in \mathfrak{R}^d$

- séparation du jeu de données à l'origine : LDA sur $X' = (X, -X)^T$

- la solution *ridge* de LDA est $\beta^*(\gamma)$

- distance de Mahalanobis : $d(x, \mu) = (\beta^* x - m)^2 + D_{\perp}$

m est la projection de μ par LDA

2) Supposons que $\phi: \mathfrak{R}^d \rightarrow \mathfrak{R}^p$ telle que $\phi(X)$ soit Gaussien. Soit le noyau K tel que $K(x_i, x_j) = \phi(x_i) \cdot \phi(x_j)$. On se place dans l'espace des *features*.

- la solution *ridge* de LDA $\alpha_*(\gamma)$ et posons $k(x) = (k(x, x_1), \dots, k(x, x_n))^T$

- On peut exprimer la distance de Mahalanobis:

$$d(x, \mu) = (\alpha_*^T k(x) - m)^2 + D_{\perp}$$

Avantages:

- distances de Mahalanobis sans calcul explicite de μ et Σ
- contrôle de la complexité par la régression *ridge*
- calcul analytique des **degrés de liberté effectif**

Détection de valeurs aberrantes

Soit $X \sim N_d(\mu, \Sigma)$, alors $\Delta \equiv (X - \mu)^T \Sigma^{-1} (X - \mu)$ suit une loi chi-carrée χ^2 de d degrés de liberté.

Nombre effectif de degrés de liberté:

LDA : $df = \text{trace}(X(X^T X + \gamma I)^{-1} X^T)$

Modèle à noyau: $df = \text{trace}(K(K + \gamma I)^{-1})$ où K est la matrice de Gram

Quantile-quantile plot : test de l'hypothèse $\Delta \sim \chi^2(d)$

- graphe des quantiles empiriques contre les quantiles théoriques χ^2
- régression linéaire robuste \implies idéalement une ligne droite
- intervalle de confiance (IC) de niveau ϵ
- observations en dehors de l'IC = valeurs aberrantes

Sélection de modèle : $\theta = (\sigma, \gamma)$

- estimation de la **vraisemblance** estimé par **validation croisée**
- performance asymptotique équivalente à celle du meilleur modèle dans la classe

- modèle dans l'espace des entrées : $p_n(x | X_n, \theta) = Z^{-1} \exp\{-\frac{1}{2}D(x; X_n, \theta)\}$

Richesse du modèle : noyaux RBF $k(x_i, x_j) = \exp(-\|x_i - x_j\|^2 / \sigma)$

Lorsque $\sigma \rightarrow 0$, $p_n(x | X_n, \theta)$ converge vers une fenêtre de Parzen.

Parzen est un estimateur non-biaisé de toute densité continue lorsque $\sigma \rightarrow 0$

Expérience : reconnaissance de visage (Olivetti face database)

- dix images différentes de 40 sujets distincts
- corruption artificielle: ajout de deux images aberrantes avec lunettes de soleil



- **but : détecter les deux images disparates**
- chaque image est représentée par un vecteur de caractéristiques de longueur 10
- sélection des hyperparamètres $\theta = (\sigma, \gamma)$ par validation croisée (2 plis)
- $\gamma = 10^{-4}$ est fixe; σ varie

Résultat:

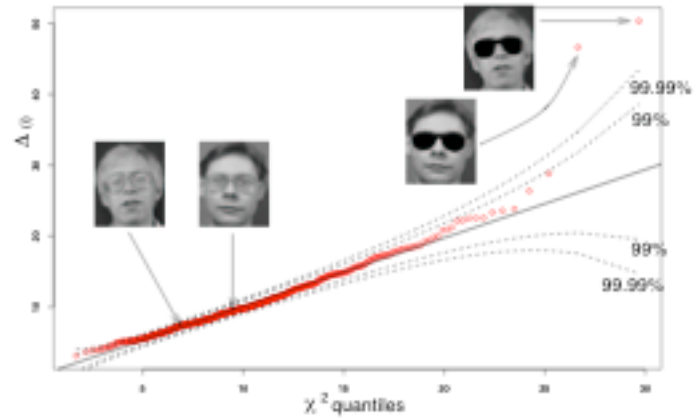


Figure 3: Quantile plot with linear fit (solid) and envelopes (99% and 99.99 %, dashed).