INF8111

Fouille de données

Été 2019

Examen final - Partie I

Enseignant: Daniel Aloise

Cet examen comporte 5 questions

Pondération: 50% de la note de l'examen final

Directives:
Un aide-mémoire recto verso de format 8 1/2 X 11 est permis à l'examen.
Calculatrice permise.
Vous remettez seulement le cahier d'examen.
Durée de l'examen: 2 heures

1 Question 1 (2 points)

Plusieurs algorithmes pour la fouille de données utilisent la matrice de covariance Σ pour effectuer une partie de leurs calculs.

Rappel:
$(\Sigma)_{st} = \frac{1}{n}\sum_{i=1}^{n}(X_i^s - \mu^s)\cdot(X_i^t - \mu^t)$
où μ^s et μ^t correspondent aux moyennes des features s et t dans X, respectivement

Pour une matrice de données numériques X quelconque:

(a) Supposons que les données soient centrées par la moyenne dans toutes les dimensions de X. C.a.d:

$$X_i^s \leftarrow X_i^s - \mu^s \quad \forall i = 1,\ldots,n; \forall s = 1,\ldots,d$$

Montrez si tel changement affecte (ou non) la matrice de covariance Σ entre les features de X.

Supposons que nous avons deux features q et r La covariance entre ces deux attributs est calculée comme suit:

$$\Sigma_{qr} = \frac{1}{n}\sum_{i=1}^{n}(X_i^q - \mu^q)\cdot(X_i^r - \mu^r)$$

Écrivons les variables centrées sur chaque dimension comme

$$X' = X - \mu$$

Nous aurions alors:

$$\Sigma_{qr} = \frac{1}{n}\sum_{i=1}^{n}(X_i'^q - \mu'^q)\cdot(X_i'^r - \mu'^r)$$

Mais, puisque après le centrage $\mu' = 0$, alors:

$$\Sigma_{qr} = \frac{1}{n}\sum_{i=1}^{n}X_i'^q \cdot X_i'^r$$

ce qui revient à la matrice de covariance original.

(b) Supposons que les données soient standardisées. C.a.d:

$$X_i^s \leftarrow \frac{X_i^s - \mu^s}{\sigma^s} \quad \forall i = 1,\ldots,n; \forall s = 1,\ldots,d$$

où σ^s est l'écart type du feature s en X.

Montrez si tel changement affecte (ou non) la matrice de covariance Σ entre les features de X.

Après le standardisation, nous avons

$$X' = \frac{X - \mu}{\sigma}$$

Nous aurons encore une fois

$$\Sigma_{qr} = \frac{1}{n} \sum_{i=1}^{n} X_i'^q \cdot X_i'^r$$

puisque $\mu' = 0$ après la standardisation. Pourtant, maintenant

$$\begin{aligned} \Sigma_{qr} &= \frac{1}{n} \sum_{i=1}^{n} X_i'^q \cdot X_i'^r \\ &= \frac{1}{n} \sum_{i=1}^{n} \frac{X_i^q - \mu_q}{\sigma^q} \cdot \frac{X_i^r - \mu_r}{\sigma^r} \\ &= \frac{\Sigma_{qr}}{\sigma^q \sigma^r} \end{aligned}$$

En conséquence, la matrice de covariance est modifiée.

2 Question 2 (2 points)

Soit un échantillon de N objets, dont k appartiennent à la classe positive. Montrez que l'espérance du $recall$ d'un classificateur aléatoire qui classifie un objet de la classe positive avec probabilité q vaut q

La probabilité que notre classificateur aléatoire produise un vrai positif (noté TP) est:

$$\begin{aligned} P(TP) &= P(\text{réel} = 1,\ \text{prédit} = 1) \\ &= P(\text{réel} = 1) \times P(\text{prédit} = 1) \qquad \text{car } P(\text{réel}) \text{ et } P(\text{prédit}) \text{ sont indépendants} \\ &= \frac{k}{N} \times q \end{aligned}$$

De même, la probabilité qu'un classificateur aléatoire produise un faux négative (noté FN):

$$\begin{aligned} P(FN) &= P(\text{réel} = 1,\ \text{prédit} = 0) \\ &= P(\text{réel} = 1) \times P(\text{prédit} = 0) \qquad \text{car } P(\text{réel}) \text{ et } P(\text{prédit}) \text{ sont indépendants} \\ &= \frac{k}{N} \times (1 - q) \end{aligned}$$

Le recall $r = \frac{TP}{TP+FN}$. L'éspérance du recall est donc :

$$\mathbb{E}[r] = \frac{P(TP)}{P(TP) + P(FN)} = \frac{\frac{k}{N} \times q}{\frac{k}{N} \times q + \frac{k}{N} \times (1 - q)} = q$$

3 Question 3 (2 points)

Supposons que nous utilisons une machine à vecteurs de support (SVM) pour trouver un hyperplan de séparation entre un ensemble de N points rouges et bleus. Supposons maintenant que nous supprimons tous les points qui ne sont pas des vecteurs de support et que nous utilisons de nouveau notre SVM pour trouver le meilleur hyperplan pour ce qui reste. Ce hyperplan pourrait-il être différent de celui obtenu en premier? Justifiez.

Non. Le problème optimisé par une SVM s'écrit comme: $\begin{array}{ll} \min & f(w) = \|w\|^2/2 \\ s.t. & y_i(w^T X_i + b) \geq 1 \quad \forall i = 1, \ldots, N \end{array}$

Notons $\mathcal{I}^* = \{i | y_i(w^T X_i + b) = 1\}$ dans la solution optimale de la SVM.

Or, pour la SVM linéaire, nous avons que la valeur du primal est égale à la valeur du dual $O = L_D$ (ref. dans les diapositives). Alors,

$$\begin{aligned} f(w^*) &= L_D(\lambda^*) \\ &= \inf_w \left(f(w) - \sum_{i=1}^{N} \lambda_i^* [y_i(w^T X_i + b) - 1] \right) \\ &\leq f(w^*) - \sum_{i=1}^{N} \lambda_i^* [y_i(w^{*T} X_i + b) - 1] \\ &\leq f(w^*) \end{aligned}$$

Du coup, les deux dernières inégalités sont des égalités, et donc

$$\lambda_i^* [y_i(w^{*T} X_i + b) - 1] = 0 \quad \forall i = 1, \ldots, N$$

Ceci est vrai pour les vecteurs de supporte, puisque:

$$y_i(w^{*T} X_i + b) - 1 = 0 \quad \forall i \in \mathcal{I}^*$$

Pour les indices $i \notin \mathcal{I}^*$, il faut que les variables λ_i^* soient égales à zero.

En conséquence, ces contraintes sont inutiles pour l'optimisation du problème primal originale parce que leur violations n'impliquent aucune pénalité. Nous pouvons alors les enlever et entraîner la SVM à nouveau menant au même hyperplan séparateur.

4 Question 4 (2 points)

Supposez que vous avez entrâiné deux modèles A e B avec le même ensemble de données X.

Considérez les situations suivantes:

(a) Les modèles A et B ont optimisé la même fonction d'erreur pour leurs entraînements sur X. À la fin de l'entraînement, l'erreur du modèle A est inférieure à celle du modèle B. Ceci veut dire que A performera mieux que B pour un ensemble de test Y? Justifiez votre réponse.

Non, en raison du surapprentissage, il se peut que A performe moins bien que B sur l'ensemble de test.

(b) Les modèles A et B ont optimisé deux fonctions d'erreur différentes pour leurs entraînements sur X. À la fin de l'entraînement, les erreurs des deux modèles sont égales à zéro. Ceci veut dire qu'ils vont performer pareil pour un ensemble de test Y? Justifiez votre réponse.

Non, le fait qu'ils ont obtenu des erreurs égales à zéro sur l'ensemble d'entraînement avec des modèles d'optimisation différents n'établit aucune relation avec la performance de A et B sur l'ensemble de teste. Ceci est vrai même si les modèles avaient optimisé la même fonction d'erreur en raison entre autres des hyperparamètres particuliers à chacun d'eux.

5 Question 5 (2 points)

Considérez les quatre faces suivantes illustrées à la Figure 1. L'obscurité ou le nombre de points représentent la densité. Les lignes servent uniquement à démarquer les régions (c.a.d., nez, bouche, yeux) et ne représentent pas des points.

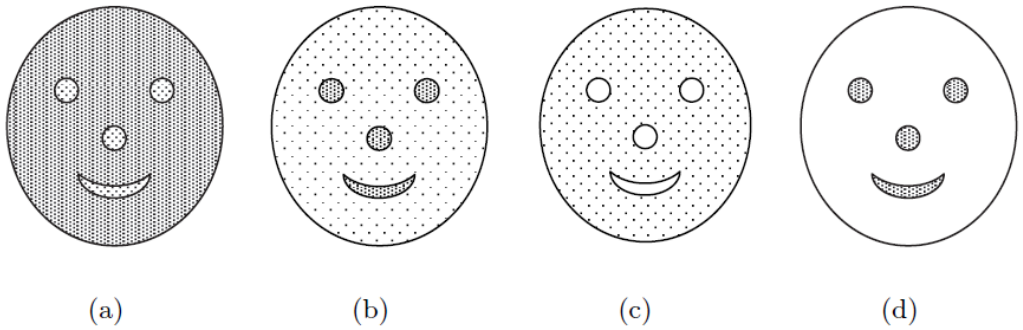

Figure 1:

(a) Pour chaque face, pourriez-vous utiliser l'algorithme *single-linkage* pour trouver les clusters représentées par le nez, les yeux et la bouche? Justifiez.

Juste pour (b) et (d). Pour (b), les points dans le nez, les yeux et la bouche sont beaucoup plus proches entre eux que les points entre ces régions. Pour (d), il n'y a que des espaces vides entre les régions.

(b) Pour chaque face, pourriez-vous utiliser l'algorithme *k-means* pour trouver les clusters représentées par le nez, les yeux et la bouche? Justifiez.

Juste pour (b) et (d), mais en fixant $k = 4$. Pour (b), k-means trouverait le nez, les yeux et la bouche, mais les autres points seraient aussi inclus dans la partition.

(c) Quelles sont les limites des algorithmes de clustering dans la détection du nez, des yeux et de la bouche sur la Figure 1 (c)?

Les algorithmes de clustering trouvent de groupes de points qui se ressemblent, pas des espaces vides.