Détection d'*outliers*

Quentin Fournier <quentin.fournier@polymtl.ca>

Les diapositives ont été créées par Daniel Aloise <daniel.aloise@polymtl.ca>

13 décembre 2019

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

Définition

données suffisament anormale pour devenir intéressante

- Un *outlier* (ou donnée aberrante) est un enregistrement très différent de la majorité des données :
  "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." D. Hawkins (1980)
- Les outliers peuvent être :
  - des déviations d'intérêts (phénomène normal inconnu ou rare)
  - du bruit provenant des capteurs
  - des anomalies (erreurs, artefacts, attaques, etc.)
- L'identification d'*outliers* est un problème complémentaire au *clustering* : les *outliers* appartiennent à des *clusters* petits ou creux.

Applications

- Nettoyage des données
- Détection de fraudes (bancaires, assurances, etc.)
- Détection d'intrusions
- Diagnostic médical
- Détection de fautes dans les systèmes informatiques

Exercice

1. À partir des méthodes déjà vues en classe, imaginez votre algorithme de détection d'*outliers*.
2. Implémentez votre algorithme et le testez sur les données `.csv` disponibles sur Moodle.
3. Quels sont les indexes des enregistrements identifiés comme des *outliers* ?

Modèles probabilistes

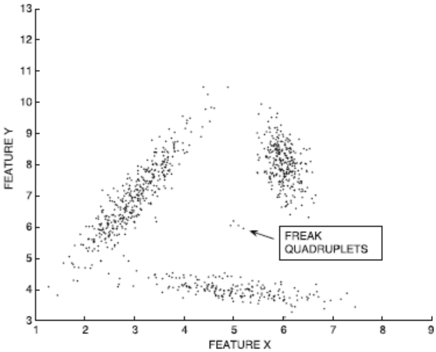
- Adaptation directe de l'algorithme EM (*gaussian mixture model*)
- On calcule la somme $j = 1, \ldots, k$

$$\sum_{j=1}^{k} \pi_j \mathcal{N}(X_i | \mu_j, \Sigma_j)$$

- Ceci est un outlier score
  - Si trop petit $\Rightarrow$ outlier trouvé

Modèles par clustering

- Le *clustering* et les détections d'*outliers* partagent une relation complémentaire bien connue.
- La détection d'*outliers* en tant que produit secondaire des méthodes de *clustering* n'est cependant pas une approche appropriée.
- Les algorithmes de *clustering* ne sont pas optimisés pour la détection d'*outliers*

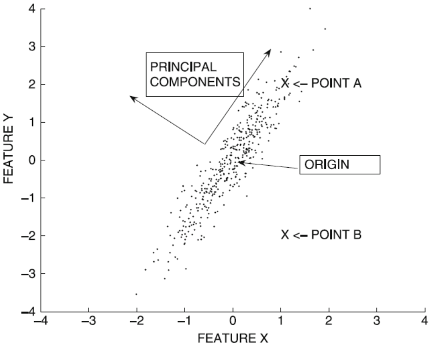


source : Aggarwal, 2015

Modèles par clustering

grouper le données et utiliser un point du centre du groupe
à l'aide de ce point comparer les autres points

- La manière simple de définir l'*outlier score* d'un enregistrement consiste à grouper d'abord l'ensemble de données, puis à utiliser la distance de chaque enregistrement vers son centre le plus proche.
- Cependant, on peut faire mieux !



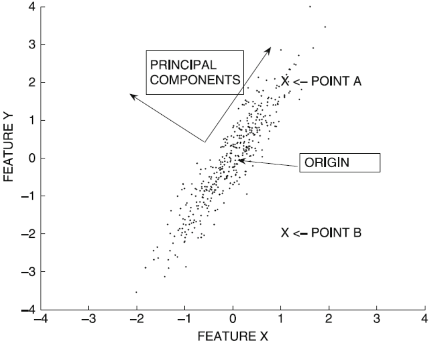
source : Aggarwal, 2015

Modèles par clustering

- Ces distances doivent-elles dépendre de la répartition des autres points dans l'espace ?

Modèles par clustering

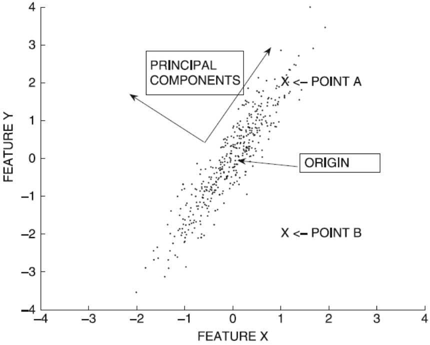
- Ces distances doivent-elles dépendre de la répartition des autres points dans l'espace ?
- La réponse est OUI.



source : Aggarwal, 2015

Modèles par clustering

- La droite de O à A est alignée avec une direction de variance élevée, et statistiquement, il est plus probable que les points soient plus éloignés dans cette direction.
- D'autre part, le segment de O à B est faiblement peuplé.
- Statistiquement, il est beaucoup moins probable que B soit aussi loin de O dans cette direction
- Par conséquent, la distance de O à A devrait être inférieure à celle de O à B .



source : Aggarwal, 2015

Distance de Mahalanobis

- Soit Σ la matrice de covariance $d \times d$ de X.
- Dans ce cas, l'entrée $\Sigma[i][j]$ est égale à la covariance entre les dimensions i et j.
- La distance de Mahalanobis $Maha(X, \mu, \Sigma)$ est donnée par :

$$Maha(X, \mu, \Sigma) = \sqrt{(X - \underbrace{\mu}_{\text{moyenne}})\Sigma^{-1}(X - \mu)^T}$$

- Normalise les données en se basant sur les covariances entre les dimensions.

Modèles par clustering

- Considèrent k clusters
- Soit :
  - μ_r le centroïde du *cluster* r
  - Σ_r la matrice de covariance des données groupées dans le *cluster* r
- L'*outlier score* d'un point X_i est calculé par rapport à son *cluster* r :

$$Maha(X_i, \mu_r, \Sigma_r) = \sqrt{(X_i - \mu_r)\Sigma_r^{-1}(X_i - \mu_r)^T}$$

Modèles par clustering

- Basés sur une analyse globale :
  - masse critique nécessaire pour avoir un *cluster* (plus d'un certain nombre des enregistrements - hyperparamètre[1])
- Ne distinguent pas très bien les données générées par du bruit de celles qui sont vraiment des anomalies.
- La distance d'un point à son centre le plus proche n'est pas très informative dans certains cas.
- On a besoin d'une analyse locale.

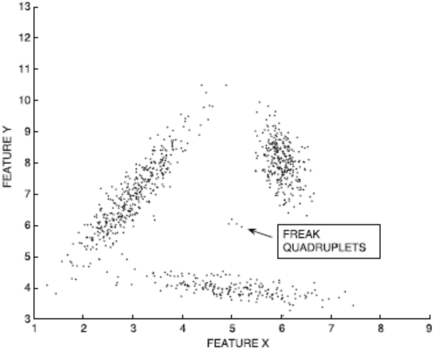
1. Paramètre choisi a priori, qui n'est pas appris par la méthode

Modèles basés sur des distances

- Basé sur les k-plus proches voisins.
- L'*outlier score* d'un enregistrement est donné par sa distance à son k^{eme}-plus proche voisin.
  - Des variantes considèrent la moyenne des k plus proches voisins
  - La méthode est sensible au choix de la distance
- k est un hyperparamètre.

Exemple

- Quelle serait une bonne valeur de k pour cet exemple ?



- $k = 2$
- $k = 3$
- $k = 4$
- $k = 5$

Modèles basés sur des distances

- Normalement les données avec beaucoup de bruit n'ont pas de grands *outliers scores* selon ce modèle...

Modèles basés sur des distances

- Normalement les données avec beaucoup de bruit n'ont pas de grands *outliers scores* selon ce modèle...
- ... mais les vrais *outliers* oui

Modèles basés sur des distances

- Normalement les données avec beaucoup de bruit n'ont pas de grands *outliers scores* selon ce modèle...
- ... mais les vrais *outliers* oui
- Cette distinction est perdue dans les méthodes de *clustering* où la distance par rapport au centre le plus proche ne reflète pas précisément l'isolation d'un certain enregistrement.

Modèles basés sur des distances

- Cette analyse plus raffinée a pourtant un prix.
- Déterminer la distance d'un enregistrement à son k^{eme} plus proche voisin nécessite un temps $O(n)$.
- $O(n^2)$ pour tous les enregistrements.

Modèles basés sur des distances

- Cette analyse plus raffinée a pourtant un prix.
- Déterminer la distance d'un enregistrement à son k^{eme} plus proche voisin nécessite un temps $O(n)$.
- $O(n^2)$ pour tous les enregistrements.
- Comment pourrait-on accélérer cet algorithme ?

Modèles basés sur des distances

- Dans la plupart des applications, on ne calcule pas les *outliers scores* de tous les enregistrements.
  points (indiviidus)
  on n'est pas obligés de claculer les scores de tt les individus, slm ceux qui sont les plus aberrants
- Il est suffisant de renvoyer des étiquettes binaires pour les top-r *outliers*, avec leurs scores.

EXAMEN dans le passé

Sampling

- Choisissez un échantillon $\mathcal{S}$ d'un ensemble de données $\mathcal{D}$ de taille $s \ll n$.
- Calculer toutes les distances par paires entre les points dans $\mathcal{S}$ et les points dans $\mathcal{D}$: $O(sn)$. points dans s * n
- Les k-plus proches voisins sont connus pour les points dans $\mathcal{S}$.
- Le r^{eme} *outlier score* dans l'échantillon est déterminé.
- Il est un lower bound (L) pour le r^{eme} *outlier score* dans l'ensemble total $\mathcal{D}$.

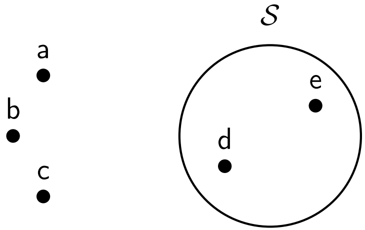
Sampling

- Pour chaque point dans $\mathcal{D} - \mathcal{S}$, nous connaissons seulement un *upper bound* par rapport à son k^{eme}-plus proche voisin.
- Si ce *upper bound* est plus petit que L, alors le point de $\mathcal{D} - \mathcal{S}$ peut-être exclu comme possible top-r *outliers*.
- L'algorithme reprend son exécution avec les points restants.

Exemple *Sampling*

Soient un ensemble de points $\mathcal{D} = \{a, b, c, d, e\}$ et $\mathcal{S} = \{d, e\}$ un échantillon de $\mathcal{D}$.

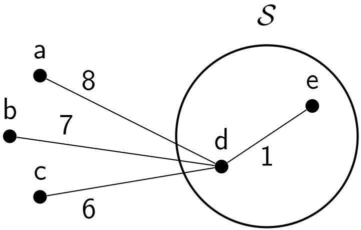
dès qu'on trouve les deux plus grands on va s'arrêter



Soient $k = 2$ (2^e^ plus proche voisin) et $r = 2$ (top-2).

Exemple *Sampling*

Calculons les distances entre tous les points ($\mathcal{D}$) et ceux dans l'échantillon ($\mathcal{S}$).

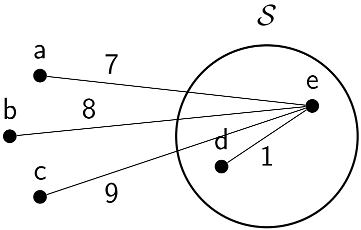


Top-r	Points	Scores	
1	d	6	plus proche voisin de d
2			

L'*outlier* score du point d est $s(d) = 6$.

Exemple *Sampling*

Calculons les distances entre tous les points ($\mathcal{D}$) et ceux dans l'échantillon ($\mathcal{S}$).



e prends la place du point le plus aberrant

Top-r	Points	Scores
1	e	7
2	d	6

L'*outlier* score du point e est $s(e) = 7$.

POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

Exemple *Sampling*

trouver une valeur plus grande mais pas suppérieure
--> c'est une borne inférieure

Les *outlier* scores des top-r sont des *lower bound*.

le point le + isolé

Le score du top-1 *outlier* vaut au moins 7. Il est possible de trouver un point avec un score plus élevé, mais un score plus petit ne peut pas remplacer celui déjà connu.

Top-r	Points	Scores
1	e	7
2	d	6

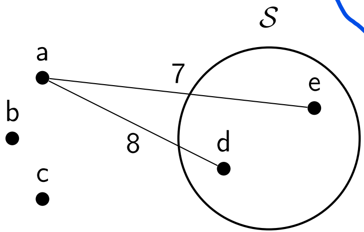
POLYTECHNIQUE MONTRÉAL
LE GÉNIE EN PREMIÈRE CLASSE

EXAMEN Sampling

Exemple *Sampling*

Considérons le point a. Seules les distances $d(a, e)$ et $d(a, d)$ sont connues.

borne suppérieure (je px trouver un score plus faible, mais pas plus grand)



Le score estimé de a est un *upper bound*. Trouver un score plus élevé signifie trouver un plus proche voisin plus éloigné que ceux déjà connus.

L'*outlier score* estimé du point a est $\hat{s}(a) = 8$.

score maximum que j'aurais pour a = 8

Exemple *Sampling*

- Si le score estimé (*upper bound*) est inférieur au plus petit score des r-*outliers* (le plus petit des *lower bound*) alors le point n'est pas un candidat et peu être élagué.

- Si le score estimé est plus grand, alors il faut calculer son score exact. Cela met aussi à jour le score des autres points (pas encore considéré).

$\hat{s}(a) = 8 > s(d) = 6$: le score estimé de a est supérieure au plus petit *lower bound*, il faut calculer la distance entre a et tous les autres points.