

Outlier Detection: Probabilistic and Clustering-Based Methods

Mining Massive Datasets

Prof. Carlos Castillo

Topic 20

Sources

- Data Mining, The Textbook (2015) by Charu Aggarwal (chapter 8) – slides by Lijun Zhang

Probabilistic methods

Related to probabilistic model-based clustering

- Assume data is generated from a **mixture-based generative model**
- **Learn** the parameters of the model from data
 - EM algorithm
- Evaluate the **probability** of each data point being generated by the model
 - Points with low values are outliers

Mixture-based generative model

- Data is generated by a **mixture** of k distributions with probability distributions G_1, \dots, G_k
- Each point \bar{X} is generated as follows:
 - 1) Select a mixture component with probability α_i
 - Suppose it's component r
 - 2) Sample a data point from distribution G_r

Learning parameters from data

- Probability of generating a point

$$\begin{aligned} f^{\text{point}}(\overline{X_j}|\mathcal{M}) &= \sum_{i=1}^k P(\mathcal{G}_i, \overline{X_j}) \\ &= \sum_{i=1}^k P(\mathcal{G}_i)P(\overline{X_j}|\mathcal{G}_i) \\ &= \sum_{i=1}^k \alpha_i f^i(\overline{X_j}) \end{aligned}$$

Learning parameters from data

- Probability of generating a point

$$f^{\text{point}}(\overline{X}_j | \mathcal{M}) = \sum_{i=1}^k \alpha_i f^i(\overline{X}_j)$$

- Probability of generating a dataset

$$f^{\text{data}}(\mathcal{D} | \mathcal{M}) = \prod_{j=1}^n f^{\text{point}}(\overline{X}_j | \mathcal{M})$$

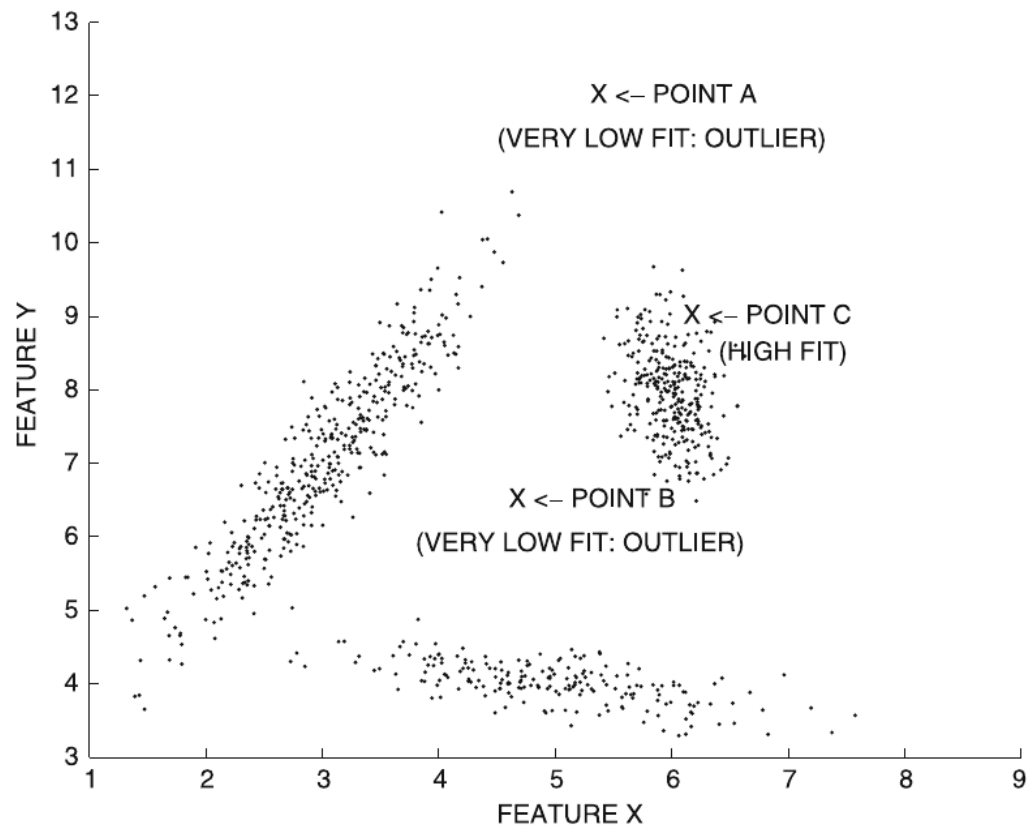
- Learning: min log loss

$$\min \mathcal{L}(\mathcal{D} | \mathcal{M}) = \log \left(\prod_{j=1}^n f^{\text{point}}(\overline{X}_j | \mathcal{M}) \right) = \sum_{j=1}^n \log \left(\sum_{i=1}^k \alpha_i f^i(\overline{X}_j) \right)$$

Identifying an outlier

Outlier score:

$$f^{\text{point}}(\overline{X_j} | \mathcal{M}) = \sum_{i=1}^k \alpha_i f^i(\overline{X_j})$$



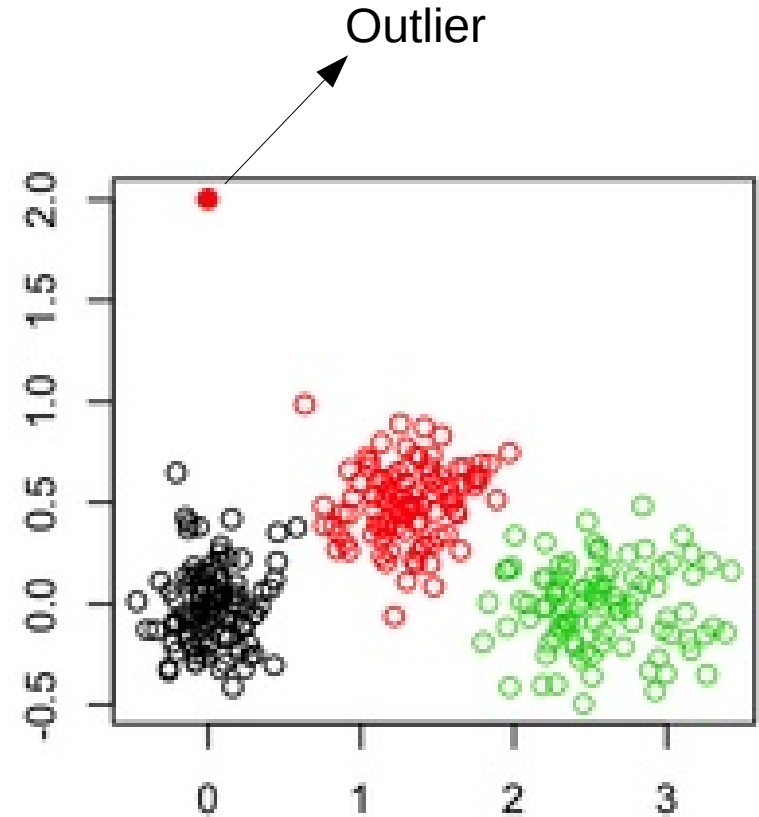
Clustering-based methods

Clustering for outlier analysis

- Clustering associate points to similar points
- Points either clearly belong to a cluster or are outliers
- Some clustering algorithms also detect outliers
 - Examples: DBSCAN, DENCLUE

Simple method

- Cluster data, associating each point to a centroid, e.g., using k-means
- Outlier score = distance of point to its centroid

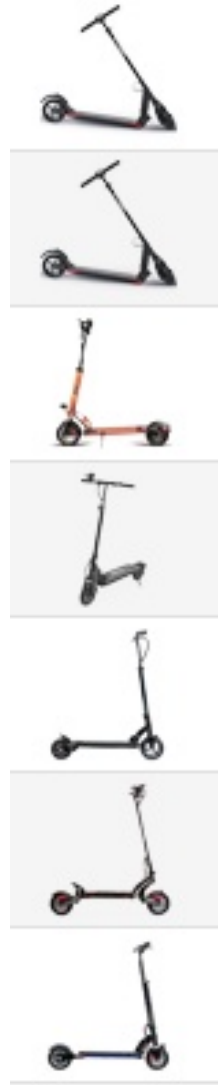


Exercise

Spreadsheet does k-means to cluster the electric scooter database

- 1) Re-run with a new initial clustering
- 2) Do you see any interesting pattern in the final clustering assignment?
- 3) Find outliers according to the method from the previous slide

Answer in
Google Spreadsheet



Improved method

- Cluster data
- Outlier score = local Mahalanobis distance with respect to center of cluster r

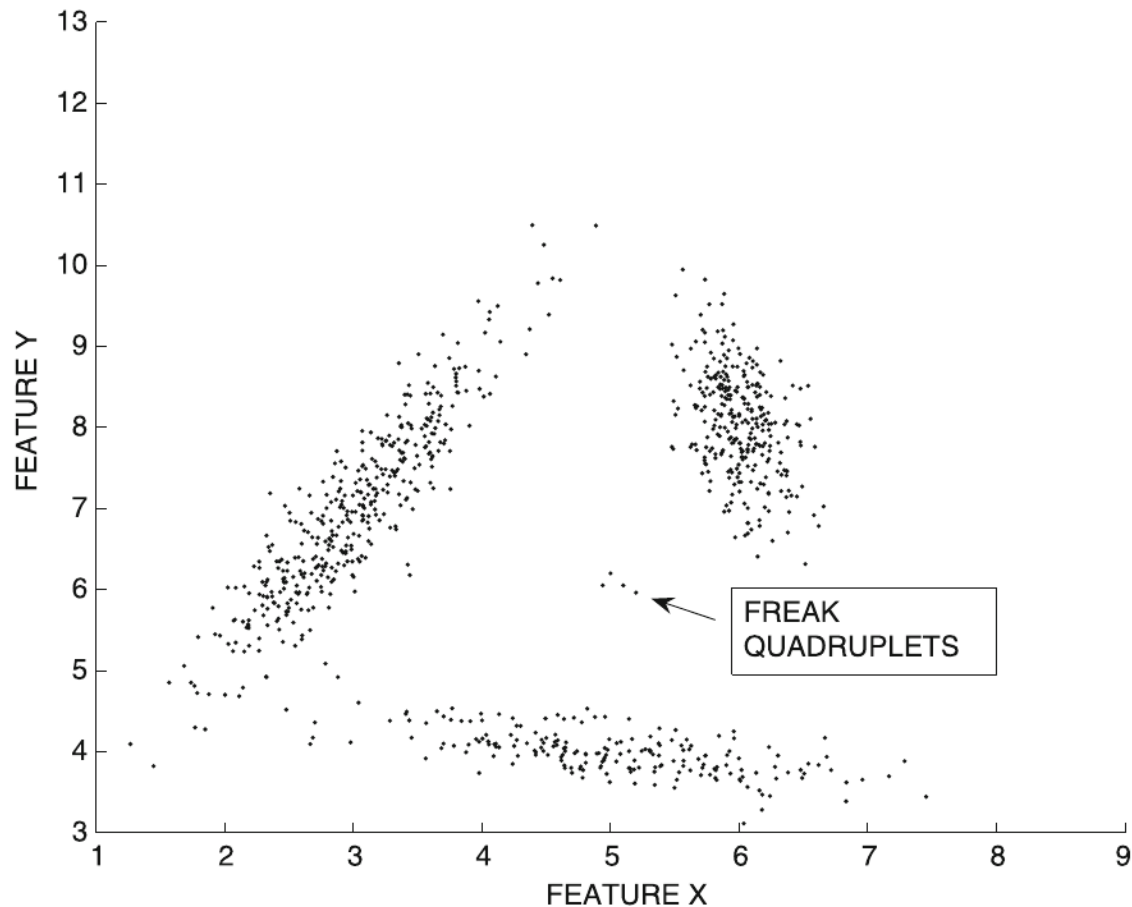
$$\text{Maha}(\overline{X}, \overline{\mu_r}, \Sigma_r) = \sqrt{(\overline{X} - \overline{\mu_r}) \Sigma_r^{-1} (\overline{X} - \overline{\mu_r})^T}$$

$\overline{\mu_r}$ is the mean of the cluster r

Σ_r is the covariance matrix of cluster r

Improved method (cont.)

- Remove tiny clusters



Summary

Things to remember

- Probabilistic methods
- Clustering-based methods

Exercises for TT19-TT21

- Data Mining, The Textbook (2015) by Charu Aggarwal
 - Exercises 8.11 → all except 10, 15, 16, 17