Exploratory Data Analysis: Learning Goals

1. Principal Component Analysis
2. Singular Value Decomposition
3. Robust Computation
4. Identification of Outliers

Exploratory Data Analysis

- So far, we have discussed clustering
- Before we start discussed supervised learning, we discuss other unsupervised approaches
- These are often called exploratory methods – to allow us to take a quick look at the data
- Here we focus on: principal components analysis and outlier analysis
- In fact, both can be used *before* clustering

Reducing Redundancy in Data

- Everyday example: AM-FM radio transmission
- Signals of left speaker strongly correlated with those of right speaker
- If we have only one channel to transmit, instead of two, which channel is good?
- Which channel is optimal?
- For some applications (e.g., news), one channel is sufficient (use AM)
- For some applications which require fidelity, (e.g., music), full signals are desirable (use FM)

Principal Component Analysis

- Perform a space rotation
- As the x-axis rotates, projects the data points onto the new axis
- Notice how the projected data range (i.e., variance) expands initially and then shrinks
- The first principal component is the new x-axis when the variance is maximized
- Construct the second principal component that is orthogonal to the first
- Notice the smaller variance on the second component

Principal Component Analysis (2)

- In k-d space, the principal components are descendingly ordered in variances, i.e.,

  $$\mathrm{var}(PC_1) \geq \mathrm{var}\ (PC_2) \geq \ldots \geq \mathrm{var}\ (PC_k)$$

- Because PCA performs a space rotation, we can reconstruct the full data if we use all k PCs
- However, if the variances of later PCs are small, we can consider those as not "carrying too much signals" and remove them
- Can we quantify $\mathrm{var}(PC_i)$?

Singular Value Decomposition

- This is really matrix algebra, rather than data mining, machine learning
- Use it for PCA, as well as for various supervised learning methods later on, including regression
- Given a matrix X, it can be factorized into:

  $$X = U\,D\,V^T$$

  - D is a diagonal matrix where $d_1 \geq d_2 \geq \ldots \geq d_k \geq 0$, and $d_1, \ldots, d_k$ are the *singular values* of X
  - U and V are orthogonal and contain the left-singular and right-singular vectors of X

Principal Component Analysis (3)

- Let X be the input data matrix, N objects/rows with k attributes/columns
- Covariance matrix $S = X^T X/N$
- SVD on X gives: $X^T X = (U D V^T)^T (U D V^T)$

$$= V D U^T U D V^T$$

$$= V D^2 V^T$$

- $\mathrm{var}(PC_i) = d_i^2 / N$
- In practice: rule of thumb, keep the top-p PC's such that they account for 80% of the sum of all d_i^2

E.g., Food Group PCA

- Notice some initial clusterings
- Also some outliers

Robust Data Analysis

- A branch of Statistics in *managing* outliers, i.e., minimizing the impact of outliers
- E.g., trimmed mean (1D)
  1. Remove the top-m biggest values, and the bottom-m smallest values
  2. Get the mean of the remaining values
- E.g. depth-contours (higher D)
  1. Remove data points that are on the convex hull
  2. Continue with step (1) for a few iterations
  3. Get the mean of the remaining data points

Robust Data Analysis (2)

- PCA is sensitive to the existence of outliers
  - There are methods for computing robust PCA
- More fundamentally, we have seen more than once the use of covariance matrix
  - Important to compute a robust covariance matrix
- Later on, when we will talk about a family of regression methods for supervised learning, they too are sensitive to outliers

Outlier Detection Methods

- Two reasons to detect outliers
- First, for robustness, we need to identify outliers and then minimize their impact
  - E.g., depth contours
- Second, we are indeed keen to know which data objects are outliers
  - Sometimes for data cleansing
  - Other times, genuine new, unexpected knowledge, e.g., credit card frauds

Existing Outlier Detection Methods

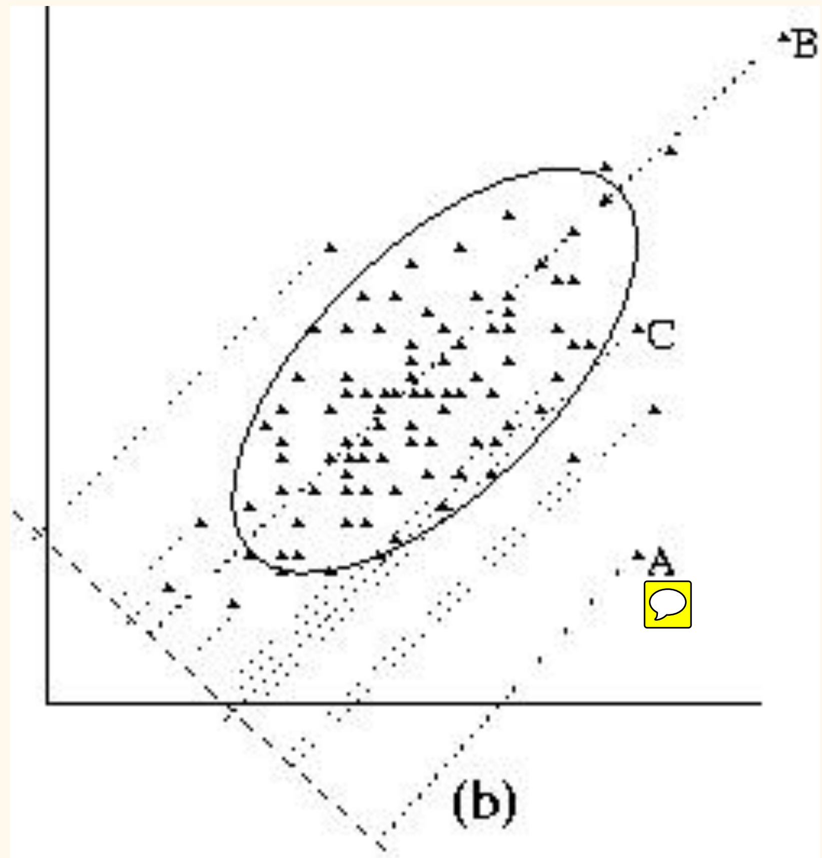
- Visual-Based (low-D only)
  - Boxplot (1-D), Scatterplot (2-D), Spin Plot (3-D)
  - Time-consuming, subjective
- Distribution-Based
  - Statistical discordancy tests
  - Requires Prior Knowledge of Distribution, # of Outliers, Types of Outliers, Mostly Univariate

Standardization-based Outlier Detection

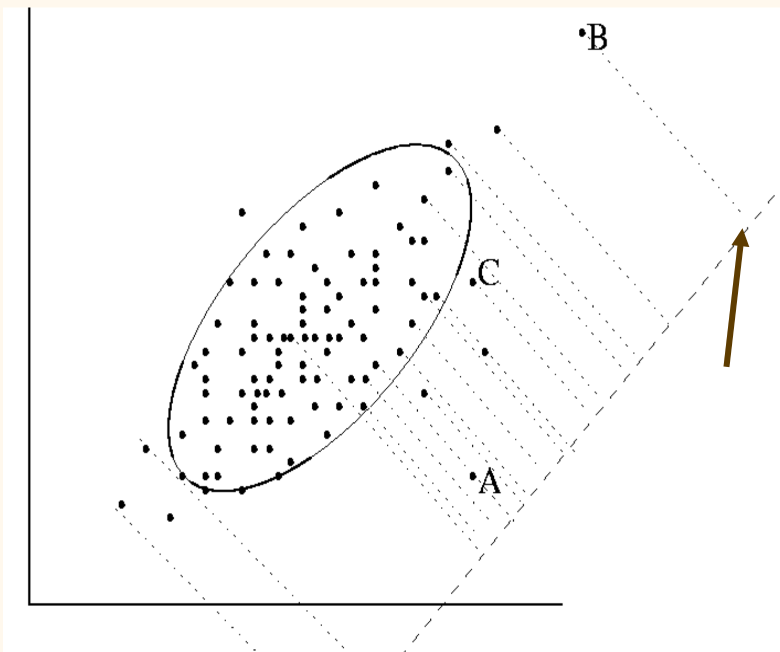
- Rule-of-thumb: $|x_{i,w} - \mu_w| / \sigma_w > 3$, then $x_{i,w}$ is outlying
- If only 1 attribute is outlying for an object, do we consider this object outlying?
- In statistics, objects that are outlying in at least one attribute/dimension are called *extreme* outliers

Structural Outliers

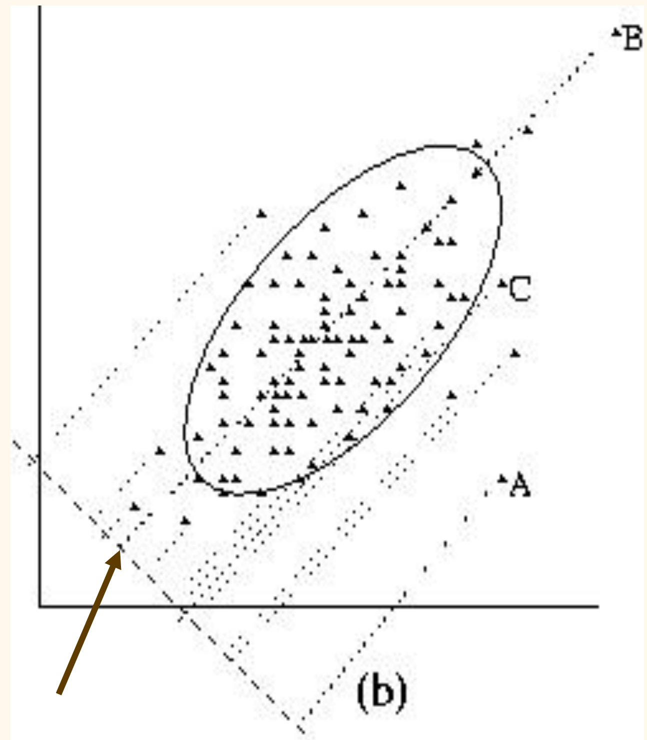
- Object B is an extreme outlier
- Object A is outlying but is not an extreme outlier
- This is called a *structural* outlier



Finding Structural Outliers by projection vectors (dashed lines)



B is outlying, but *not* A, C



B is not outlying here

Using PCA

- May need many projection vectors to identify structural outliers
- One heuristic is to use PCA (even though PCA itself is sensitive to outliers)
- Given the data matrix X, PCA-transform to X'
- Find extreme outliers in X'; these are structural outliers in X

Concluding Remarks: Big Data

- PCA requires matrix factorization
- Though it is worse than quadratic time, many methods have been developed for parallelizing SVD and sampling
- Scalable methods also exist for detecting outliers in large, high-dimensional datasets
- Many of those methods are highly parallelizable