Hasso-Plattner-Institute
Knowledge Discovery and Data Mining
Prof. Emmanuel Müller November 30th, 2015

Exercise 2 for the lecture
Big Data Analytics
WS 2015/2016

Next tutorial session:
December 9th (15:45 - 17:15), the time slot had to be moved for this tutorial (30 min.)!
Solutions will be presented by students and discussed during the tutorial. Please prepare your solutions in order to participate actively during the tutorial session.

1 Principal component analysis

Consider the following data set with five examples ($N = 5$) and two attributes ($D = 2$):

$$A = \begin{bmatrix} 2 & 2 \\ 3 & 3 \\ 3 & 2 \\ 4 & 3 \\ 5 & 5 \end{bmatrix}$$

(a) Perform PCA on this data set. You have to center the data, compute the covariance matrix of the centered data, and find the eigenvalues and eigenvectors of the covariance matrix (the principal components). Use the eigenvector matrix to perform the basis transformation, and finally reduce the dimensionality of the data by removing one of the two dimensions from the transformed data set.

(b) Perform naive dimensionality reduction by simply removing one of the two dimensions from the original data matrix. Compute the sample variance of the naive reductions and the PCA reductions from (a) with the formula

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - \mu)^2.$$

What do you observe? What does this tell you about projection quality?

(c) Dimensionality reduction with PCA works best if the data set is Gaussian distributed. Draw a non-Gaussian two-dimensional data set where PCA loses a lot of the original variance when only one dimension is kept after projecting the data. Come up with a more suitable method to reduce its dimensionality (informally). *Hint: consider a non-linear projection.*

2 Entropy-based discretization

Entropy-based discretization is a supervised binning approach that aims at finding boundaries for discretization that keep the class labels of the resulting bins as pure as possible. Consider the following set of sensor measurements a_i with class labels $c_i \in \{\text{OK}, \text{FAIL}\}$:

i	1	2	3	4	5	6	7	8
a_i	0.1	0.2	0.8	0.9	1.0	4.0	10.0	50.0
c_i	FAIL	FAIL	OK	OK	FAIL	OK	OK	OK

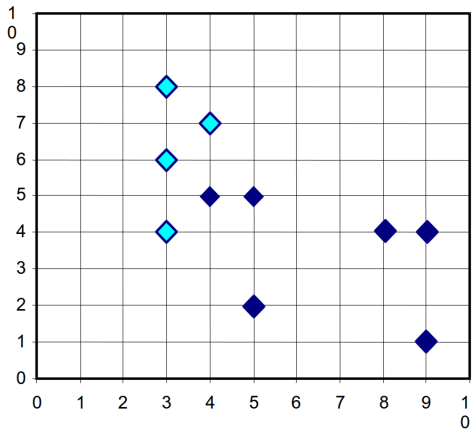
We now want to perform entropy-based discretization for the values $a_1, ..., a_8$.

(a) Compute the entropy for the candidate boundaries $T = 0.5$, $T = 0.95$, $T = 2.5$. Which boundary gives the best discretization? Use that boundary to discretize the data.

(b) Describe a method to decide which candidate boundaries to test.

3 k-Means and EM algorithm

Consider the data and its initial partitions depicted below:

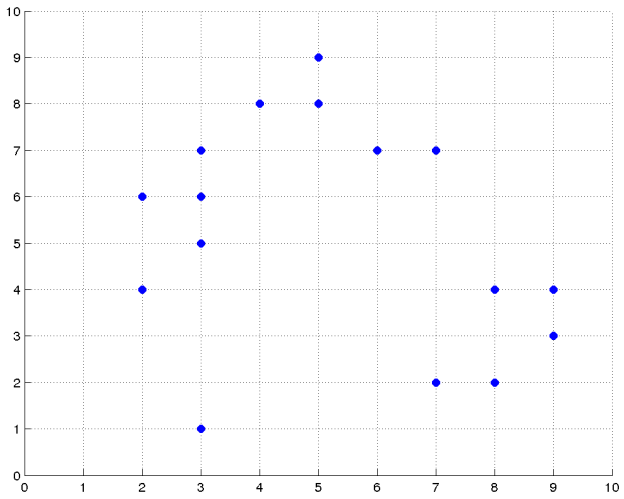


(a) Apply the k-Means algorithm to find two clusters. Use the Manhattan distance as metric and give the clustering results after each iteration of the algorithm, including the coordinates of the cluster means.

(b) k-Means is often used to find initial parameters for a Gaussian Mixture Model (GMM) clustering. These parameters are then refined with the EM algorithm. Determine the initial GMM parameters π_1, μ_1, Σ_1 and π_2, μ_2, Σ_2 for the two clusters C_1 and C_2 obtained in (a). *Note: You don't have to perform EM training.*

4 DBSCAN

Given the following two-dimensional data set: $P_1 = (2,4)$, $P_2 = (2,6)$, $P_3 = (3,1)$, $P_4 = (3,5)$, $P_5 = (3,6)$, $P_6 = (3,7)$, $P_7 = (4,8)$, $P_8 = (5,8)$, $P_9 = (5,9)$, $P_{10} = (6,7)$, $P_{11} = (7,2)$, $P_{12} = (7,7)$, $P_{13} = (8,2)$, $P_{14} = (8,4)$, $P_{15} = (9,3)$, $P_{16} = (9,4)$.



(a) Use the heuristic from the lecture to determine input parameters MinPts and ϵ for DBSCAN using the **Manhattan distance**.

(b) Apply DBSCAN to the given data set using the **Euclidean distance**, MinPts $= 4$ and $\epsilon = 2.2$. Give the resulting clusters as sets of points and a list of all core objects. *Note: You don't have to write down intermediate results.*