'From Statistics to Data Mining' Standard Questions for the Exam

---

Linear Algebra

Let $A = \begin{pmatrix} -2 & -1 & 2 \\ 1 & 0.5 & 1 \\ 0 & 0 & 2 \end{pmatrix}$ be a 3×3 matrix.

1. Compute the determinant of A. What do you conclude?
2. Compute the eigenvalues of A. What do you conclude? Using the answer of this question, check the result of the first question.
3. Compute the eigenvectors of A.
4. Compute $A^T A$.

Let $B = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 1 & 1 \\ 2 & 1 & 0 \end{pmatrix}$ be a 3×3 matrix. Compute the inverse matrix B^{-1}.

Let us consider the following matrix $A = \begin{pmatrix} 1 & 2 & 1 \\ 1 & 1 & 3 \\ 2 & 3 & 2 \end{pmatrix}$.

1. Compute the inverse matrix A^{-1} using the method based on the **augmented matrix** $[A|I]$.
2. Compute the determinant of A.

Principal Component Analysis

Let $X = \{A, B, C\}$ be a set of 3 examples lying in a 3-dimensional feature space where $A = (1, 2, 4)$, $B = (0, 1, 2)$ and $C = (2, 3, 6)$.

1. Compute the 3×3 covariance matrix Σ from the **zero mean values** of A, B and C.
2. Find the eigenvalues of Σ. What do you conclude?
3. Compute the unit-eigenvector $\vec{u}$ corresponding to the largest eigenvalue.
4. Plot the 3 points in $\mathbb{R}$ according to $\vec{u}$.
5. Compute the matrix of Euclidean distances between those 3 points A, B, C both in $\mathbb{R}^3$ and $\mathbb{R}$. What do you conclude?

Clustering

Agglomerative algorithms

Consider the following proximity matrix:

$$P(X) = \begin{pmatrix} 0 & 16 & 8 & 32 & 64 \\ 16 & 0 & 10 & 22 & 100 \\ 8 & 10 & 0 & 2 & 4 \\ 32 & 22 & 2 & 0 & 22 \\ 64 & 100 & 4 & 22 & 0 \end{pmatrix}.$$

1. How many examples are we considering?
2. Apply the Single Link agglomerative clustering.
3. Apply the Complete Link agglomerative clustering.

Indications:

- Describe each iteration of the algorithms, and explain your choice.
- Show the dendograms obtained along with the similarity level.

K-means algorithm

Consider the following dataset X : $A = (1,1)$, $B = (-1,-1)$, $C = (-1,-2)$, $D = (2,-3)$, $E = (4,-3)$.
The objective is to apply the k-means algorithm with the Manhattan distance to cluster X into 3 clusters.

1. Compute $P(X)$ the proximity matrix of X.
2. Suppose that the initial seeds (centers of each cluster) are $C_1 = (0,0)$, $C_2 = (0,-3)$ and $C_3 = (3,-3)$. Run the k-means algorithm for the first iteration only. You have to:
   (a) describe the construction of the clusters (and compute them),
   (b) compute the centers of the new clusters.