German University in Cairo
Faculty of Media Engineering and Technology
Spring 2019

CSEN1083 – Data Mining

Problem Set #3

Problem 1

Consider the problem of finding the K nearest neighbors of a data object. A programmer designs Algorithm 2.1 for this task.

Algorithm 2.1 Algorithm for finding K nearest neighbors.

```
1: for i = 1 to number of data objects do
2:     Find the distances of the i^th object to all other objects.
3:     Sort these distances in decreasing order.
       (Keep track of which object is associated with each distance.)
4:     return the objects associated with the first K distances of the sorted list
5: end for
```

(a) Describe the potential problems with this algorithm if there are duplicate objects in the data set. Assume the distance function will only return a distance of 0 for objects that are the same.

(b) How would you fix this problem?

Solution:

(a) There are several problems. First, the order of duplicate objects on a nearest neighbor list will depend on details of the algorithm and the order of objects in the data set. Second, if there are enough duplicates, the nearest neighbor list may consist only of duplicates. Third, an object may not be its own nearest neighbor.

(b) There are various approaches depending on the situation. One approach is to keep only one object for each group of duplicate objects. In this case, each neighbor can represent either a single object or a group of duplicate objects.

CSEN1083 – Data Mining

Problem Set #3

Problem 2

You are given a set of m objects that is divided into K groups, where the i^{th} group is of size m_i. If the goal is to obtain a sample of size $n < m$, what is the difference between the following two sampling schemes? (Assume sampling with replacement.)

(a) We randomly select $n*m_i/m$ elements from each group.

(b) We randomly select n elements from the data set, without regard for the group to which an object belongs.

Solution:

The first scheme is guaranteed to get the same number of objects from each group, while for the second scheme, the number of objects from each group will vary. More specifically, the second scheme only guarantees that, on average, the number of objects from each group will be $n*m_i/m$.

Problem 3

Consider a document-term matrix, where tf_{ij} is the frequency of the i^{th} word (term) in the j^{th} document and m is the number of documents. Consider the variable transformation that is defined by

$$tf'_{ij} = tf_{ij} * \log \frac{m}{df_i},$$

where df_i is the number of documents in which the i^{th} term appears and is known as the document frequency of the term. This transformation is known as the inverse document frequency transformation.

(a) What is the effect of this transformation if a term occurs in one document? In every document?

(b) What might be the purpose of this transformation?

Solution:

(a) Terms that occur in every document have 0 weight, while those that occur in one document have maximum weight, i.e., $\log m$.

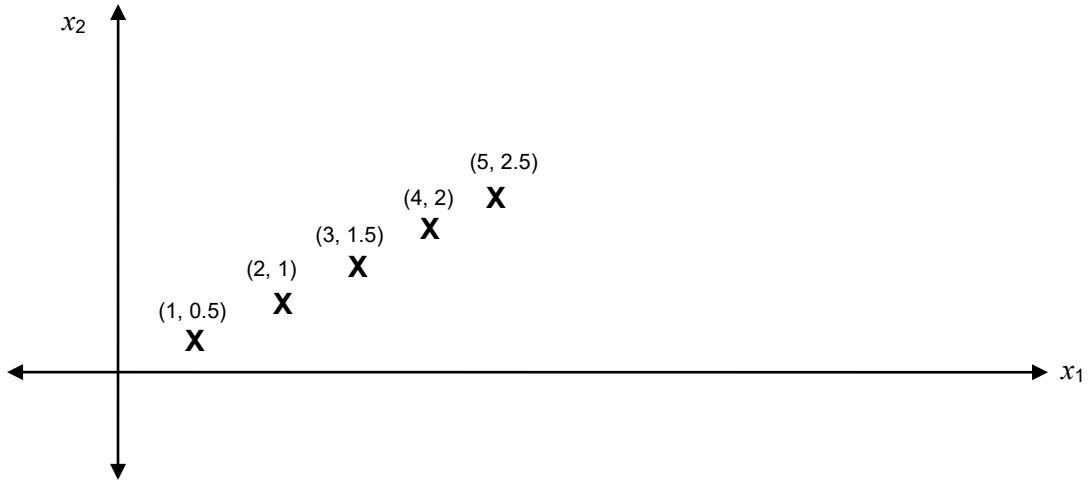
(b) This normalization reflects the observation that terms that occur in every document do not have any power to distinguish one document from another, while those that are relatively rare do.

CSEN1083 – Data Mining

Problem Set #3

Problem 4

Apply Principal Component Analysis (PCA) to the data given below.



Solution

The first step is to compute the mean of the points using the equation $\bar{\mathbf{x}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$

$$\bar{\mathbf{x}} = \frac{1}{5}\left(\begin{bmatrix}1\\0.5\end{bmatrix}+\begin{bmatrix}2\\1\end{bmatrix}+\begin{bmatrix}3\\1.5\end{bmatrix}+\begin{bmatrix}4\\2\end{bmatrix}+\begin{bmatrix}5\\2.5\end{bmatrix}\right)=\begin{bmatrix}3\\1.5\end{bmatrix}$$

We next compute the covariance matrix **S** given by $\mathbf{S} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n-\bar{\mathbf{x}})(\mathbf{x}_n-\bar{\mathbf{x}})^{\mathrm{T}}$

$$\mathbf{S} = \frac{1}{5}\left[\left(\begin{bmatrix}1\\0.5\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)\left(\begin{bmatrix}1\\0.5\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)^T+\left(\begin{bmatrix}2\\1\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)\left(\begin{bmatrix}2\\1\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)^T+\left(\begin{bmatrix}3\\1.5\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)\left(\begin{bmatrix}3\\1.5\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)^T\right.$$
$$\left.+\left(\begin{bmatrix}4\\2\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)\left(\begin{bmatrix}4\\2\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)^T+\left(\begin{bmatrix}5\\2.5\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)\left(\begin{bmatrix}5\\2.5\end{bmatrix}-\begin{bmatrix}3\\1.5\end{bmatrix}\right)^T\right]$$
$$=\begin{bmatrix}2 & 1\\1 & 0.5\end{bmatrix}$$

Next, we should find the eigenvectors and eigenvalues of **S**. To do so, we should solve the eigen decomposition problem given by: $\mathbf{u}_1^{\mathrm{T}}\mathbf{S}\mathbf{u}_1 = \lambda_1$

We compute the characteristic equation given by $|\mathbf{S}-\lambda I| = 0$

CSEN1083 – Data Mining

Problem Set #3

$$\left|\begin{bmatrix}2 & 1\\ 1 & 0.5\end{bmatrix}-\lambda\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right|=0$$

$$\left|\begin{bmatrix}2-\lambda & 1\\ 1 & 0.5-\lambda\end{bmatrix}\right|=0$$

$$(2-\lambda)(0.5-\lambda)-(1)^2=0$$

Solving this equation, we get two eigenvalues:

$\lambda_1 = 2.5, \lambda_2 = 0$

To get, principal component 1 (PC1), we find the eigenvector corresponding to the largest eigenvalue λ_1.

$$\begin{bmatrix}2-\lambda_1 & 1\\ 1 & 0.5-\lambda_1\end{bmatrix}\begin{bmatrix}u_{11}\\ u_{12}\end{bmatrix}=0$$

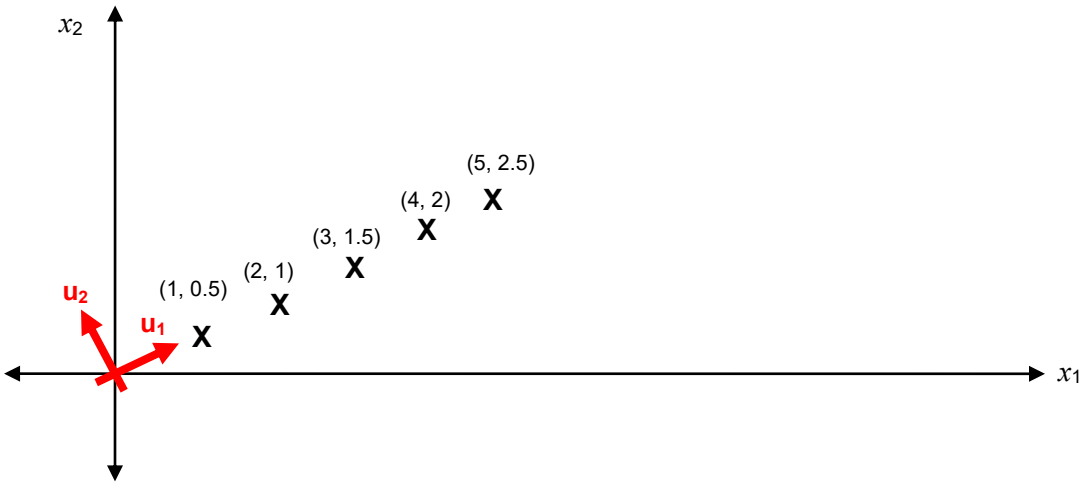
$$\begin{bmatrix}-0.5 & 1\\ 1 & -2\end{bmatrix}\begin{bmatrix}u_{11}\\ u_{12}\end{bmatrix}=0$$

$$-0.5u_{11}+u_{12}=0 \rightarrow u_{11}=2u_{12}$$

Therefore, PC1 is given by $\mathbf{u_1} = c\begin{bmatrix}2\\ 1\end{bmatrix}$.

Since, the principal components have to be normal, therefore $\mathbf{u_1}=\begin{bmatrix}2/\sqrt{2^2+1^2}\\ 1/\sqrt{2^2+1^2}\end{bmatrix}=\begin{bmatrix}0.8944\\ 0.4472\end{bmatrix}$

The other principal component $\mathbf{u_2}$ will be orthogonal to $\mathbf{u_1}$ $\mathbf{u_2}=\begin{bmatrix}-0.4472\\ 0.8944\end{bmatrix}$



CSEN1083 – Data Mining

Problem Set #3

Problem 5

For the following vectors, x and y, calculate the indicated similarity or distance measures.

(a) x = (1, 1, 1, 1), y = (2,2,2,2) cosine, correlation, Euclidean

(b) x = (0, 1,0, 1), y = (1,0, 1,0) cosine, correlation, Euclidean, Jaccard

(c) x = (0, -1,0, 1), y = (1,0, -1,0) cosine, correlation, Euclidean

(d) x = (1, 1,0, 1,0, 1), y = (1, 1, 1,0,0, 1) cosine, correlation, Jaccard

(e) x = (2,-1,0,2,0,-3), y = (-1, 1, -1,0,0, -1) cosine, correlation

Solution

(a) cos(x,y) = 2, corr(x,y) = 0/0 (undefined), Euclidean(x,y)=2

(b) cos(x,y) = 0, corr(x,y)=−1, Euclidean(x,y) = 2, Jaccard(x,y)=0

(c) cos(x,y) = 0, corr(x,y)=0, Euclidean(x,y)=2

(d) cos(x,y) = 0.75, corr(x,y) = 0.25, Jaccard(x,y) = 0.6

(e) cos(x,y) = 0, corr(x,y)=0