Data Mining and Decision Support Final

1. Assume we have a dataset (shown in (1)) consisting of people arriving at an emergency room at a hospital (0 and 1 indicate, respectively, that the patient did or did not have the corresponding problem). Find the following probabilities given this data: $P(\text{admission} = \text{morning})$, $P(\text{comatose} = 1|\text{admission} = \text{afternoon})$. Use a Naive Bayes classifier to find the most likely admission time for a patient without insurance suffering from a broken bone.

$$\begin{bmatrix} \text{admission} & \text{broken bone} & \text{comatose} & \text{dizzy} & \text{has insurance} \\ \text{morning} & 1 & 1 & 1 & 1 \\ \text{morning} & 1 & 1 & 1 & 1 \\ \text{afternoon} & 0 & 0 & 1 & 0 \\ \text{evening} & 1 & 1 & 1 & 0 \\ \text{afternoon} & 0 & 0 & 1 & 0 \\ \text{afternoon} & 1 & 1 & 1 & 0 \\ \text{afternoon} & 1 & 0 & 0 & 1 \\ \text{evening} & 1 & 1 & 0 & 1 \end{bmatrix} \tag{1}$$

To find $P(\text{admission} = \text{morning})$ divide the total number of people admitted in the morning by the total number of observations. $P(\text{admission} = \text{morning}) = 2/8$

$$P(\text{comatose} = 1|\text{admission} = \text{afternoon}) = \frac{\text{number comatose and morning}}{\text{number morning}} = 1/4 \tag{2}$$

Naive Bayes assumes the features are independent. Find $P(\text{admission} = x|\text{data})$. Where data are all of the conditional probabilities. Use Bayes theorem to invert conditional probability. Our data is (broken bone = 1), (comatose = 0),(dizzy = 0),(has insurance = 0).

$$P(\text{adm.} = x)P(\text{b. b.} = 1|\text{adm.} = x)P(\text{comat.} = 0|\text{adm.} = x)P(\text{diz.} = 0|\text{adm.} = x)P(\text{has insurance} = 0|\text{adm.} = x) \tag{3}$$

This expression is maximized for $x = \text{afternoon}$.

2. Find the covariance matrices for each of the datasets A,B,C, and D shown below. (Hint: all covariance matrices in this problem are 2×2)

$$\mathbf{A} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{B} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 1 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{C} = \begin{bmatrix} 0 & 0 \\ 0 & 0 \\ 2 & 0 \\ 0 & 0 \\ 0 & 0 \\ 0 & 0 \end{bmatrix}, \mathbf{D} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \end{bmatrix}, \tag{4}$$

Call the elements of the covariance matrix $R = (r_{ij})$. The formula for the covariance is

$$R_{ij} = \sigma_{ij}^2 = E[(X_i - \mu_{x_i})(X_j - \mu_{x_j})] \tag{5}$$

In our example R can also be expressed

$$\frac{1}{5}\left(X - \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}[\mu_1, \mu 2]\right)^T \left(X - \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}[\mu_1, \mu 2]\right) \tag{6}$$

Case A:

$$\frac{1}{5}\left(\begin{bmatrix}0&0\\0&0\\0&0\\0&0\\0&0\\0&0\end{bmatrix} - \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}[0,0]\right)^T \left(\begin{bmatrix}0&0\\0&0\\0&0\\0&0\\0&0\\0&0\end{bmatrix} - \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}[0,0]\right) = R_{A,ij} = \begin{bmatrix}0&0\\0&0\end{bmatrix} \tag{7}$$

Case B:

$$\frac{1}{5}\left(\begin{bmatrix}0&0\\0&0\\0&0\\0&1\\0&0\\0&0\end{bmatrix} - \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}[0,1/6]\right)^T \left(\begin{bmatrix}0&0\\0&0\\0&0\\0&1\\0&0\\0&0\end{bmatrix} - \begin{bmatrix}1\\1\\1\\1\\1\\1\end{bmatrix}[0,1/6]\right) = R_{B,ij} = \begin{bmatrix}0&0\\0&1/6\end{bmatrix} \tag{8}$$

Case C:

$$R_{C,ij} = \begin{bmatrix}2/3&0\\0&0\end{bmatrix} \tag{9}$$

Case D:

$$R_{D,ij} = \begin{bmatrix}3/10&-3/10\\-3/10&3/10\end{bmatrix} \tag{10}$$

3. Given a set of five points in 2D space (e.g. $p_1 = (x_1, y_1)$, $p_2 = (x_2, y_2)$, etc.) expressed as the rows of a matrix A. Treat y as the dependent variable and x as the independent variable. Show how set up a linear regression for this data, that is, find a function of the form $y = \alpha x + \beta$ that is optimal in the least squares sense. Show how to fit a cubic polynomial to the data. (Hint: set this up as a matrix multiplication and solve using the pseudoinverse technique).

$$\mathbf{A} = \begin{bmatrix}1&0\\1&1.5\\2&0\\3&0\\3.5&-0.5\end{bmatrix} \tag{11}$$

Define a new design matrix B that has a column represent a constant and a column representing a linear term, then solve the following by least squares

$$\begin{bmatrix} 1 & 1 \\ 1 & 1 \\ 1 & 2 \\ 1 & 3 \\ 1 & 3.5 \end{bmatrix} \begin{bmatrix} \beta \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ 1.5 \\ 0 \\ 0 \\ -0.5 \end{bmatrix} \tag{12}$$

Likewise for a general cubic polynomial solve

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 3.5 & 12.25 & 42.875 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ c_2 \\ c_3 \end{bmatrix} = \begin{bmatrix} 0 \\ 1.5 \\ 0 \\ 0 \\ -0.5 \end{bmatrix} \tag{13}$$

4. Consider solving for $\mathbf{x}$ in $\mathbf{Ax} = \mathbf{b}$ in the least-squares sense for the data below.

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 3 & 0 \\ 1 & 2 \end{bmatrix}, \mathbf{b} = \begin{bmatrix} -1 \\ 1 \\ -2 \end{bmatrix} \tag{14}$$

   (a) For this problem, how many *parameters* are obtained in the least-squares solution?

   There are 2 parameters.

   (b) Show what the objective function would be if we were solving this problem using LASSO regression with a tuning parameter λ.

$$J = ||\mathbf{Ax} - \mathbf{b}|| + \lambda \sum_i |x_i| \tag{15}$$

   (c) What problem would have in using the pseudo-inverse technique to solve this problem if A were as follows?

$$\mathbf{A} = \begin{bmatrix} 1 & 1 \\ 3 & 3 \\ 1 & 1 \end{bmatrix} \tag{16}$$

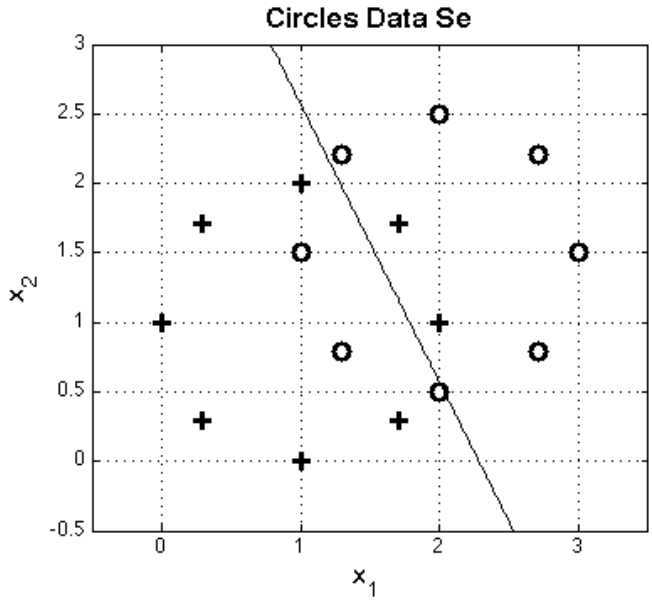
   In this case, we can not easily invert $A^T A$ to compute the pseudoinverse.

5. The *singular value decomposition* (SVD) is typically expressed as a factorization of a matrix $A = USV^T$. What are the properties of the three matrices U, S, and V in the SVD? Provide an algorithm for finding an optimal low rank (say rank 3) approximation of A using the SVD.

   Matrices U and V are orthogonal and matrix S is diagonal. The elements along the main diagonal are the singular values. To obtain an optimal low-rank approximation we can first obtain the SVD, set $n - 3$ smallest singular values to zero, and remultiply.

6. Explain what *latent semantic analysis* is and how it employs the singular value decomposition.

   Latent semantic analysis is an technique to analyze a corpus of textual documents by forming the *document-term* matrix, calculating its SVD, and using the SVD to find a low rank approximation.



7. The figure above (Circles Data Set) shows a two-class supervised machine learning problem. The classes are indicated, respectively by plus symbols and circles. Each class consists of 8 data points equally spaced on a circle of radius 1 centered at $(1, 1)$ and $(2, 1.5)$. Answer the following:

   (a) What is the total number of *features*?
   2 features

   (b) What is the total number of *data instances*?
   16 data instances

   (c) What are the within-class means for each of the classes, μ_{plus} and μ_{circle}? Draw and label the within-class means on the figure.
   In class means are the circle centers, $(1, 1)$ and $(2, 1.5)$

   (d) What are the in-class covariance matrices for the plus and circle classes? Explain why they are identical. (Hint: the covariance matrices are diagonal)
   The in-class covariances are both

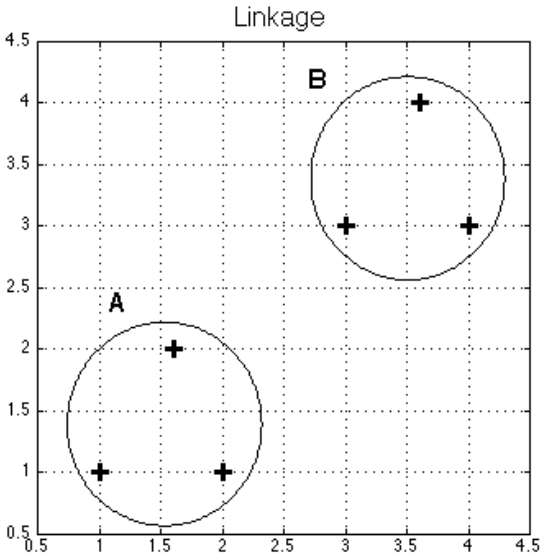
$$R = \begin{bmatrix} 4/7 & 0 \\ 0 & 4/7 \end{bmatrix} \tag{17}$$

   We expect the covariances to be the same since the instances in each class are arranged identically except for the mean, and the mean is removed before computing the covariance.

   (e) The solid line in the figure indicates a decision boundary for a classifier obtained by *linear discriminant analysis*. The classifier is of the form $w_1x_1 + w_2x_2 > \theta$ for a threshold parameter θ. Find w_1 and w_2.
   Using the formula for LDA $S_W^{-1}(\mu_1 - \mu_2)$ gives

$$\left(2\begin{bmatrix} 4/7 & 0 \\ 0 & 4/7 \end{bmatrix}\right)^{-1} ([1, 1]^T - [2, 1.5]^T) = [-7/8, -7/16]^T \tag{18}$$

   (f) Let plus indicate a *positive* result and circle indicate a *negative* result. Find a value of θ for which the LDA classifier has a false negative rate of 0 on the given data.
   We need to set θ so that only circles appear to the right of the line. This is true for $\theta = -2.5$.



8. The figure above (Linkage) shows a clustering problem. Find the distances between groups A and B using euclidean distances with average, complete, and single linkage. For reference the data in the figure is given below.

$$\begin{bmatrix} 1 & 1 \\ 2 & 1 \\ 1.6 & 2 \\ 3 & 3 \\ 4 & 3 \\ 3.6 & 4 \end{bmatrix} \tag{19}$$

Single linkage: 1.72, Complete linkage: 3.97, Average linkage: 2.89

9. What does the acronym SVG stand for and how is it used in data visualization in conjunction with `d3`?

SVG stands for scalable vector graphics and can be used by d3 to render data visually.

10. Consider the following `d3` code

```
d3.select("svg").append("g")
  .selectAll("rect")
  .data(dataset)
  .enter()
  .append("rect")
  .attr("width",25)
  .attr("height",25)
  .attr("x", function(d,i) { return d.x * 3;})
  .attr("y", function(d,i) { return d.y * 3;})
  .attr("fill","rgb(25,25,20)")
  .attr("stroke","black")
```

(a) What is the purpose of the `enter()` method in this code?
The enter method introduces new `rect` elements that are bound to the datapoints in dataset.

(b) Explain what `d` and `i` are in the anonymous function definitions.
`d` is the bound data point for the given element, and `i` is an index for the dataset.

(c) Based on this code, what type of object do expect `dataset` to be?
dataset should be an array of javascript objects with x and y fields.

11. List the components of a box-and-whisker diagram and sketch an example.

    The box-and-whisker diagram or box plot consists of a box that extends from the first two third quartiles. A central line indicating the median. The whiskers extend outwards from the main box and determine which samples are considered outliers. Outliers are drawn explicitly.

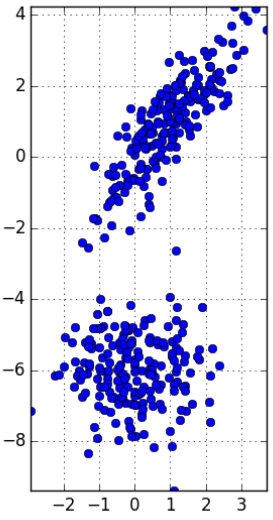
12. What is *principal components analysis* and how is it related to eigenvector decomposition? Categorize PCA as a supervised or unsupervised learning technique.

    PCA is a method that finds an orthogonal transformation that collects the variance of a data set into the fewest possible components. PCA is based upon the eigenvector decomposition of the sample covariance matrix. PCA is an unsupervised technique.

13. Sketch the expected output of the following python script.

```
#!/usr/bin/env python

import numpy as np
import matplotlib.pyplot as plt

mean = np.array([1,1])
cov = np.array([[1,1.2],[1.2,1.9]])

Num=200

samp = np.random.multivariate_normal(mean,cov,Num)
samp2 = np.random.multivariate_normal(np.array([0,-6]),np.array([[1,0],[0,1]]),Num)

out = np.vstack((samp,samp2))

plt.figure(1)
plt.grid()
plt.plot(out[:,0],out[:,1],'o')
plt.axis('image')
plt.show()
```

    Actual output,



14. Explain what bootstrapping is and how it can be used to deal with the problem of having a small amount of training data.

    Bootstrapping means to draw randomly with replacement. Drawing with replacement means that we have a degree of control over the size of the samples we draw even if our input data is limited.

15. In addition to the labeled dataset, what information should be provided when performing k-nearest neighbor classification?

    We must provide the number of nearest neighbors k to use in classification and also the metric used for computing distances, e.g. euclidean, hamming.

16. What is a *fold* in cross-validation? How many folds are created for *leave-one-out* cross validation?

    A fold is a partition of the data into groupings of equal size. One of the partitions is used for testing while the rest are used for training. Repetition of this process for each different choice of training test split are the folds. In leave-one-out, there are as many folds as data instances.