COMP527
Data Mining and Visualisation
Problem Set 1

Danushka Bollegala

Question 1

A. State the two main types of data mining models. **(2 marks)**

Predictive models and descriptive models. Each point will be assigned 1 mark.

B. Consider that you measured the height and weight of 100 students for a health survey. For 20 students in your sample you could only measure either their height or weight, but not both values. Assume that we would like to train a binary classifier to predict whether a student is overweight compared to the students in this dataset. Answer the following questions about this experiment.

(a) State two algorithms that you can use to learn a binary classifier for this purpose. **(2 marks)**

logistic regression, SVM, perceptron, etc.

(b) What is meant by the *missing-value* problem in data mining? **(3 marks)**

Some of the feature values (attributes) in the data might be missing because either the measurements were not taken and/or the data is corrupted.

(c) State two disadvantages we will encounter if we ignore the 20 instances that we have incomplete measurements for and use the remaining 80 instances to train the classifier. **(4 marks)**

The dataset size will be too small and we might overfit to it. The dataset size might be too small to learn anything useful (under fitting). The missing data points might contain useful information about the target task.

(d) The average height of the students in this dataset is 169cm. Provide a reason for and a reason against using the average to fill the missing values. **(4 marks)**

For: It is a typical value for the height of the students. Against: The 20 *students for which we do not have height measurements could be outliers.*

(e) Assume that we would like to check whether there is any correlation between the height and the weight of the students in this dataset. How do we check this? **(4 marks)**

We could measure the Pearson correlation coefficient between the height and the weight, and if it is high we could conclude that there is a high correlation between the two variables.

(f) Given that there is a high correlation between the height and the weight of a student, how can we use this information to overcome the missing-value problem? **(4 marks)**

We could learn a linear relationship between the two variables using a technique such as the linear regression and then use the learnt predictor to predict the missing values. We can then train a binary classifier using this predicted data points as well as the original data points.

(g) Without having access to a separate test dataset, how can we evaluate the accuracy of our binary classifier? **(2 marks)**

We can set aside a portion of the train data as held out data, and evaluate using that portion.

Question 2 Consider a training dataset consisting of four instances $(\boldsymbol{x}_1, 1)$, $(\boldsymbol{x}_2, 1)$, $(\boldsymbol{x}_3, -1)$ $(\boldsymbol{x}_4, -1)$ where $\boldsymbol{x}_1 = (1, 1)^\top$, $\boldsymbol{x}_2 = (-1, 1)^\top$, $\boldsymbol{x}_3 = (-1, -1)^\top$, and $\boldsymbol{x}_4 = (1, -1)^\top$. Here, $\boldsymbol{x}^\top$ denotes the transpose of vector $\boldsymbol{x}$. We would like to train a binary Perceptron to classify the four instances in this dataset. For this question ignore the bias term b in the Perceptron and answer the following.

A. Let us predict an instance $\boldsymbol{x}$ to be positive if $\boldsymbol{w}^\top \boldsymbol{x} \geq 0$, and negative otherwise. Initializing $\boldsymbol{w} = (0, 0)^\top$, show that after observing $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$ in that order the weight vector will be $-\boldsymbol{x}_3 - \boldsymbol{x}_4$. **(6 marks)**

When $\boldsymbol{w} = \boldsymbol{0}$, we have $\boldsymbol{w}^\top \boldsymbol{x}_1 = 0$. Hence, $\boldsymbol{x}_1$ is correctly predicted as positive. Same applies for $\boldsymbol{x}_2$ as well. However, $\boldsymbol{x}_3$ will be misclassified and the weight vector will be updated to $\boldsymbol{w} = \boldsymbol{0} - \boldsymbol{x}_3 = \boldsymbol{x}_3$. Next, $-\boldsymbol{x}_3{}^\top \boldsymbol{x}_4 = 0$ and $\boldsymbol{x}_3$ will be classified incorrectly as positive. Therefore, $\boldsymbol{w} = -\boldsymbol{x}_3 - \boldsymbol{x}_4$.

B. If we present the four instances in the reverse order $(\boldsymbol{x}_4, -1)$, $(\boldsymbol{x}_3, -1)$, $(\boldsymbol{x}_2, 1)$, $(\boldsymbol{x}_1, 1)$, to the Perceptron, what would be the final value of weight vector at the end of the first iteration? **(4 marks)**

$-\boldsymbol{x}_4 - \boldsymbol{x}_3$

C. Normalize each of the four instances $\boldsymbol{x}_1$, $\boldsymbol{x}_2$, $\boldsymbol{x}_3$, and $\boldsymbol{x}_4$ into unit L2 length. **(4 marks)**

All the normalized vectors will have a factor $\frac{1}{\sqrt{2}}$ in front.

D. What would be the final weight vector after observing the four instances if you used the L2 normalized training instances instead of the original (unnormalized) instances to train the Perceptron as you did in the part (A) of above? **(4 marks)**

$-\frac{1}{\sqrt{2}}(\boldsymbol{x}_3 + \boldsymbol{x}_4)$.

E. Now, let us re-assign the target labels for this dataset as follows $(\boldsymbol{x}_1, 1)$, $(\boldsymbol{x}_2, -1)$, $(\boldsymbol{x}_3, 1)$ $(\boldsymbol{x}_4, -1)$. Can we use Perceptron algorithm to linearly classify this revised dataset? Justify your answer. **(4 marks)**

No. The dataset is no longer linearly separable. Answers that either plots the data points in the 2D space or use some other method to show this will receive full marks. If no justification is given, then such answers will receive 2 marks.

F. Describe a method to learn a binary linear classifier for the revised dataset described in part (E) above. **(3 marks)**

Kernalized versions such as using the product of the two features as a third feature will receive full marks.