PAPER CODE NO. **COMP527**	EXAMINER : Dr. Danushka Bollegala DEPARTMENT : Computer Science Tel. No. 0151 7954283

UNIVERSITY OF LIVERPOOL

Resit Examinations 2016/17

Data Mining and Visualisation

TIME ALLOWED : Two and a Half Hours

INSTRUCTIONS TO CANDIDATES

Answer FOUR questions.

If you attempt to answer more questions than the required number of questions (in any section), the marks awarded for the excess questions answered will be discarded (starting with your lowest mark).

Question 1

A. State two techniques used in Support Vector Machines to classify linearly non-separable datasets. **(2 marks)**

(a) Consider the quadratic kernel given by $k(\boldsymbol{x}, \boldsymbol{y}) = (\boldsymbol{x}^\top \boldsymbol{y} + 1)^2$ for two vectors $\boldsymbol{x} = (x_1, x_2)^\top$ and $\boldsymbol{y} = (y_1, y_2)^\top$. Show that the quadratic kernel can be written as the inner-product between two 6-dimensional vectors, each containing information only from one of $\boldsymbol{x}$ and $\boldsymbol{y}$. **(4 marks)**

(b) Explain why a quadratic kernel might be able to learn a decision hyperplane for a non-linearly separable dataset. **(5 marks)**

B. Consider three data points $\boldsymbol{x}_1 = (1, 0)$, $\boldsymbol{x}_2 = (3, -1)$, and $\boldsymbol{x}_3 = (3, 1)$. Here, $\boldsymbol{x}_2$ and $\boldsymbol{x}_3$ are labelled positively (+1), whereas $\boldsymbol{x}_1$ is negative (-1). Answer the following questions related to training a linear-kernel support vector machine on this dataset.

(a) Assuming that the decision hyperplane is defined by a weight vector $\boldsymbol{w} = (w_1, w_2)^\top$ and a bias term b, show that the prediction score y of a test instance $\boldsymbol{z} = (z_1, z_2)$ is given by $y = w_1 z_1 + w_2 z_2 + b$. **(3 marks)**

(b) Write the inequality constraints that must be satisfied by the three support vectors $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$. **(3 marks)**

(c) By solving the inequality constraints you derived in (b), compute w_1, w_2 and b. **(3 marks)**

(d) Plot the decision hyperplane alongside with the support vectors. **(3 marks)**

(e) Predict the class label of a test data point $(4, 1)$. **(2 marks)**

Question 2 We would like to use the Perceptron algorithm to learn a linear classifier $y = \boldsymbol{w}^\top \boldsymbol{x} + b$, defined by a weight vector $\boldsymbol{w} \in \mathbb{R}^d$ and a bias $b \in \mathbb{R}$ from a training dataset consisting of four instances, $\{(t_n, \boldsymbol{x}_n)\}_{n=1}^4$. Here, $\boldsymbol{x}_1 = (-1, 0)^\top$, $\boldsymbol{x}_2 = (1, 0)^\top$, $\boldsymbol{x}_3 = (1, 1)^\top$, and $\boldsymbol{x}_4 = (-1, 1)^\top$, and the labels are $t_1 = -1$, $t_2 = +1$, $t_3 = +1$, and $t_4 = -1$. We predict an instance $\boldsymbol{x}$ as positive if $\boldsymbol{w}^\top \boldsymbol{x} + b > 0$, and negative otherwise. The initial values of the weight vector and the bias are set respectively to $\boldsymbol{w}^{(0)} = (0, 0)^\top$ and $b = 0$. We visit the training instances in the order $\boldsymbol{x}_1, \boldsymbol{x}_2, \boldsymbol{x}_3$, and $\boldsymbol{x}_4$. Answer the following questions.

A. Plot the dataset in the two-dimensional space. **(2 marks)**

B. Write the perceptron update rule for a misclassified instance $(t, \boldsymbol{x})$. **(3 marks)**

C. What will be the values of the weight vector and the bias after observing the instance $\boldsymbol{x}_1$. **(3 marks)**

D. What will be values of the weight vector and the bias after observing $\boldsymbol{x}_2$. **(4 marks)**

E. What will be the values of the weight vector and the bias after observing $\boldsymbol{x}_3$. **(4 marks)**

F. What will be the values of the weight vector and bias after observing $\boldsymbol{x}_4$. **(4 marks)**

G. If we re-assign the labels as $t_1 = -1, t_2 = +1, t_3 = -1, t_4 = 1$, show that there does not exist a weight vector $\boldsymbol{w} = (w_1, w_2)^\top$ and a bias b that can classify all four instances correctly. **(5 marks)**

Question 3 Consider a two-dimensional dataset consisting of five instances $x_0 = (0, 0)$, $x_1 = (1, 1)$, $x_2 = (-1, 1)$, $x_3 = (-1, -1)$, and $x_4 = (1, -1)$. We would like to cluster this dataset into two clusters using the k-means clustering algorithm. Answer the following questions.

A. Write the within cluster sum of squares objective function for a set of K clusters $S_1, S_2, \ldots, S_K$. **(3 marks)**

B. Explain why it is important to randomly initialise the k-means algorithm and run for multiple times. **(4 marks)**

C. Let us assume the two initial cluster centres to be x_1 and x_2. Find the two clusters produced by the k-means algorithm at convergence. **(3 marks)**

D. Compute the value of the within cluster sum of squares objective for the set of clusters obtained in **(C)**. **(2 marks)**

E. Let us assume the two initial cluster centres to be x_0 and x_2. Find the two clusters produced by the k-means algorithm at convergence. **(3 marks)**

F. Compute the value of the within cluster sum of squares objective for the set of clusters obtained in **(E)**. **(2 marks)**

G. Let us assume the two initial cluster centres to be x_4 and x_2. Find the two clusters produced by the k-means algorithm at convergence. **(3 marks)**

H. Compute the value of the within cluster sum of squares objective for the set of clusters obtained in **(G)**. **(2 marks)**

I. Based on your results above, which set of clusters should we select? Justify your answer. **(3 marks)**

Question 4 Let us assume that we must develop a sentiment classifier to predict sentiment of user reviews about products for an online e-commerce portal. Answer the following questions.

A. Providing examples, explain what is meant by the term *stop word* in the context of text mining. **(3 marks)**

B. State a benefit of removing stop words when training a classifier. **(2 marks)**

C. Providing examples, explain what is meant by the term *part-of-speech tagging* in the context of text mining. **(3 marks)**

D. Why would it be useful to use part-of-speech tags when training a sentiment classifier? **(2 marks)**

E. Why would it be good to use bigrams in addition to unigrams when representing user reviews for training a sentiment classifier. **(4 marks)**

F. If the user reviews are rated from 1 to 5 stars in a discrete ordinal scale, where higher the value more positive the sentiment, how would you assign labels to the reviews such that you can train a binary sentiment classifier? **(3 marks)**

G. If our dataset contains identical copies (duplicate reviews), will it affect the performance of a naive Bayes classifier? Explain your answer. **(4 marks)**

H. We would like to apply singular value decomposition to reduce the size of the feature vectors representing the reviews. Assuming that you do not have a separate validation dataset, how can you determine the optimal dimensionality for the singular value decomposition? **(4 marks)**

Question 5 Consider the multi-layer feedforward neural network shown in Figure 1. This neural network has 3 inputs x_1, x_2 and x_3 connected to a hidden layer consisting of two nodes h_1 and h_2. The weight of the edge connecting x_i to h_j is w_{ji}. The two hidden nodes are connected to the output node o. The weight of the edge connecting the hidden node h_i to the output node o is u_i. The activation functions at hidden and output layers is set to sigmoid function defined as follows:

$$\sigma(\theta) = \frac{1}{1 + \exp(-\theta)}$$

Moreover, squared error is used as the loss function at the output node, and is defined as,

$$E(o, t) = \frac{1}{2}(o - t)^2,$$

where t is the target output.

Answer the following questions.

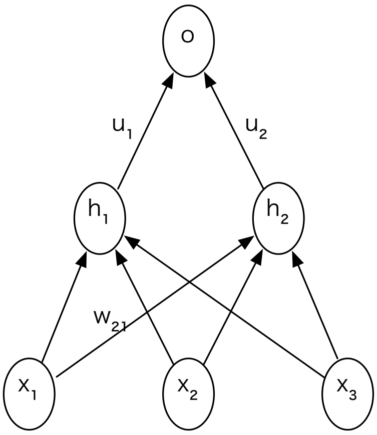

Figure 1: A multi-layer feedforward neural network.

A. State one advantage and one disadvantage of using a single layer neural network vs. multi-layer neural network. **(2 marks)**.

B. Using the symbols defined in Figure 1, compute the activation at h_1. **(3 marks)**

C. Compute the gradient of the loss with respect to the output o. **(3 marks)**

D. Compute the gradient of the loss w.r.t u_1. **(3 marks)**

E. Compute the gradient of the loss with respect to w_{12}. **(4 marks)**

F. Using the Stochastic Gradient Descent rule, write the update rule for w_{12}. **(4 marks)**

G. Using the update rule derived in **(F)**, explain why we should scale the initial values of the weights such that the activation at the output node does not fall in the saturated regions of the logistic sigmoid function. **(3 marks)**

H. If we would like to learn a sparse neural network where most of the edge weights are set to zero, how can we modify the loss function. **(3 marks)**