THE UNIVERSITY
OF ADELAIDE
AUSTRALIA

Semester 2 2020

Introduction to Statistical Machine Learning
COMPSCI 3314, 7314

Writing Time:	130 mins
Uploading Time:	30 mins
Total Duration:	160 mins

Questions	Time	Marks
Answer all 6 questions	130 mins	100 marks
		100 Total

Basic Concepts of Machine Learning, etc.

Question 1

(a) Identify which of the following classifiers are nonlinear classifiers: (1) RBF kernel SVM (2) 1-nearest neighbour classifier (3) 3-nearest neighbour classifier

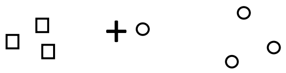
Write down your choice (choices).

[4 marks]

(b) Write down the probabilistic density function of a Gaussian Mixture Model (GMM), that is, the likelihood of a sample x belonging to a GMM (2 points). Explain the relationship between estimating the parameters of a GMM and clustering data with a GMM. In other words, if we already learned the parameters of a GMM, how could we calculate the membership (or equivalently the posterior probability) of a sample belonging to a cluster (4 points)?

[6 marks]

(c) As shown in the following figure, there are 8 data points in the 2-dimensional space. The squares denote training samples belonging to class “1” and circles denote training samples belonging to class “2”. The cross "+" denotes a test sample. What is the predicted category of the test sample if we use a 1-nearest neighbour classifier? and what is the prediction if we use a 3-nearest neighbour classifier? Explain how could we choose k for k-nearest neighbour classifier? (1 point, 2 points and 2 points)



[5 marks]

(d) **True or False:** A classifier with a lower training error on the training set always performs better on the test set.

Briefly explain why.

True or False: One way to avoid overfiting is to use less training data.

Briefly explain why.

True or False: It is a good practice to choose the optimal hyper-parameter of a model on the test set.

Briefly explain why.

[6 marks]

[Total for Question 1: 21 marks]

Support Vector Machines (SVMs) and Kernels

Question 2

(a) Name one benefit of solving the optimization problem of Support Vector Machines in its dual form.

[3 marks]

(b) The primal formulation of the standard Support Vector Machines can be rewritten as:

$$\min_{\mathbf{w},b,\xi} \frac{1}{2} \| \mathbf{w} \|^2 + C \sum_{i=1}^{n} max(0, \xi_i),$$
$$\text{s.t.} \quad y_i(\langle \mathbf{x}_i, \mathbf{w} \rangle + b) \geq 1 - \xi_i, i = 1, \cdots, n,$$

where $max(0, \xi_i) = 0 \; if \; \xi_i < 0; max(0, \xi_i) = \xi_i \; if \; \xi_i \geq 0$.

Explain the reason why, there is no constraint $\xi_i \geq 0$ needed, in the above optimisation problem.

[4 marks]

(c) **True or False:** If we remove the samples that are not support vectors of a Support Vector Machine from the training set and retrain the Support Vector Machine on the remaining samples, we will obtain the same solution as using the original training set.

Briefly explain why.

True or False: If the training samples are linearly separable, using hard-margin support vector machines and soft-margin support vector machines will result in the same solution.

Briefly explain why.

[6 marks]

(d) Suppose we have n kernel functions $K_j(\cdot, \cdot), \quad j = 1, \cdots, n$ such that there are n implicit high-dimensional feature maps $\Phi_j : R^d \rightarrow R^D \;\; j = 1, \cdots, n$ that satisfies $\forall \, \mathbf{x}, \mathbf{z} \in R^d$, $K_j(\mathbf{x}, \mathbf{z}) = \Phi_j(\mathbf{x}) \cdot \Phi_j(\mathbf{z})$, where $\Phi_j(\mathbf{x}) \cdot \Phi_j(\mathbf{z}) = \langle \Phi_j(\mathbf{x}), \Phi_j(\mathbf{z}) \rangle = \sum_{i=1}^{D} \Phi(\mathbf{x})_j^i \Phi(\mathbf{z})_j^i$ is the dot product (a.k.a. inner product) in the D-dimensional space. $\Phi(\cdot)_j^i$ denotes the i-th dimension of the j-th mapped feature.

Is the summation of those kernel functions, that is, $K(\mathbf{x}, \mathbf{z}) = \sum_j^n K_j(\mathbf{x}, \mathbf{z})$, still a valid kernel function?

If yes, prove that. If no, please explain why.

[8 marks]

[Total for Question 2: 21 marks]

Boosting

Question 3

(a) **True or False**: Assume that the weak learners are a finite set of decision stumps, Adaboost cannot achieve zero training error if the training data is not linearly separable.

Briefly explain why.

[4 marks]

(b) **True or False**: Once a weak classifier is picked in a particular round, it will never be chosen in any subsequent round.

Briefly explain why.

[4 marks]

(c) **True or False**: Using Adaboost with a stronger weak classifier, for example, weak classifiers that can achieve higher accuracy when they are used individually, can always lead to better accuracy on the test set.

Briefly explain why.

[4 marks]

(d) **True or False**: Assume that the weak learners are a finite set of decision stumps, normalising features (scaling each dimension of features to ensure that each dimension has zero mean and unit variance) will help Adaboost achieve better predictive accuracy on the test set.

Briefly explain why.

[4 marks]

[Total for Question 3: 16 marks]

Regression

Question 4

(a) Please write down the objective function and solution of the Ridge regression, that is, l_2 norm regularised linear regression, respectively. (2 points and 2 points)

Explain why we need to add the l_2 norm regularisation (2 points)?

[6 marks]

(b) Lasso, that is, l_1 norm regularized linear regression optimises the following objective function

$$\min_{\mathbf{w}} \sum_{i=1}^{N} \|\mathbf{x}_i^T \mathbf{w} - y_i\|_2^2 + \lambda \|\mathbf{w}\|_1, \tag{3}$$

where $\mathbf{x}_i \in \mathbb{R}^d$ is i-th data sample and y_i is the i-th target value. $\mathbf{w}$ is the model parameters. Lasso can be used for selecting important features that are informative in predicting the target values. Please explain why Lasso can be used for feature selection? (3 points) What will happen to the solution if we increase λ in the above optimisation problem. (2 points)

[5 marks]

(c) Please write down the objective function of kernel regression (the kernelized version of linear regression without regularization terms). You need to show the details of your derivation.

[4 marks]

[Total for Question 4: 15 marks]

Dimensionality reduction

Question 5

(a) Name three benefits of performing dimensionality reduction.

[3 marks]

(b) Comparing with Principal Component Analysis (PCA), what is the advantage of using Isomap to perform dimensionality reduction.

[3 marks]

(c) Given 1000 facial images, among them only one image captures a person wearing glass. By using Principal Component Analysis (PCA), we could generate an facial image which is similar to that person but with the glass removed. Describe how to achieve this and why this is possible.

[4 marks]

(d) We want to perform dimensionality reduction with Principal Component Analysis (PCA) to a set of d-dimensional features and choose the target dimensionality to be (1) the same as the original dimension, d, i.e., without reducing the dimensionality but applying the projection matrix, and (2)$d/2$, i.e., reduce the original dimensionality by 50%.

True or False: After applying PCA, the pairwise Euclidean distance between any two data samples will not change if the target dimensionality is d.

True or False: We could always perfectly reconstruct the original features by multiplying the transpose of the projection matrix to the reduced features if the target dimensionality is $d/2$. Note: "perfectly reconstruct" means that there is no difference between the reconstructed feature and the original feature.

Write down your choice and briefly explain why.

[6 marks]

[Total for Question 5: 16 marks]

Neural Networks

Question 6

(a) Why we need to use stochastic gradient descent rather than standard gradient descent to train a convolutional neural network?

[3 marks]

(b) We use the following convolutional neural network to classifiy a set of 32×32 grey scale images (so the input size will be 32×32×1):

1) Layer 1: convolutional layer with the ReLU nonlinear activiation function, 100 5×5 filters with stride 1.

2) Layer 2: 2×2 max-pooling layer

3) Layer 3: convolutional layer with the ReLU nonlinear activiation function, 50 3×3 filters with stride 1.

4) Layer 4: 2×2 max-pooling layer

5) Layer 5: fully-connected layer

6) Layer 6: classfication layer

How many model parameters we need to optimise in the first layer and in the second layer (assume the bias term is used) (4 points and 4 points).

[8 marks]

[Total for Question 6: 11 marks]

End of exam