THE UNIVERSITY OF ADELAIDE
AUSTRALIA

Practise Exam

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

Overview of Machine Learning, etc.

Question 1

(a) Cross-validation is a method to (Choose the best single answer from multiple choices):

(A) Remove the curse of dimensionality

(B) Assess how the results of a machine learning model will generalise to an unseen data set

(C) Remove noises or outliers from a data set

[2 marks]

(b) Kernel Principal Component Analysis is a method for (Choose the best single answer from multiple choices):

(A) Classification

(B) Reduction of the dimensionality

(C) Probability estimation

(D) Regression

[2 marks]

(c) Which of the following statements is best practice in Machine Learning for building a real system? (Choose the best single answer from multiple choices)

(A) Use all the data available for training to obtain optimal performance

(B) Use all the data available for testing the performance of your algorithm

(C) Split the training data into two separate sets. Use the first subset for training and perform cross-validation solely on the second subset

(D) Perform cross-validation on training, validation and testing sets

[3 marks]

(d) Which of the following statements about Machine Learning is **False**? (Choose the best single answer from multiple choices)

(A) Machine learning algorithms often suffer from the curse of dimensionality

(B) Machine learning algorithms cannot generalise to the data that are not observed during training of the algorithm

(C) Machine learning algorithms are typically sensitive to noise

(D) Machine learning algorithms typically perform better in terms of testing accuracy when more training data become available

[3 marks]

(e) Which of the following statements is (are) true? (Select all the correct ones)

(A) Gaussian mixture model (GMM) is a supervised learning method.

(B) With the correct step size and batch size, stochastic gradient descent always converges to the global minimum of the objective function for a support vector machine with polynomial kernels if such as minimum exists.

(C) With the correct step size and batch size, stochastic gradient decent always converges to the global minimum of the objective function for a 10-layer convolutional neural network if such a minimum exists.

(D) For $k = 1$, k-nearest neighbour (kNN) classifiers can achieve 100% accuracy on the training set, therefore implying that choosing $k = 1$ in kNN tends to produce the best model.

(E) If one wants to train a classifier to categorise an input image into one of 1,000 categories. There are over 10^6 labelled images available for training. For this problem, a deep convolutional neural network with 50 layers is more likely to work better than a support vector machine classifier.

[4 marks]

(f) Which of the following modification to the Gaussian mixture models will make it *most similar* to k-means? (Choose the best single answer from multiple choices)

(A) Restrict each covariance matrix Σ_i to have all off-diagonal entries being zeros.

(B) Restrict Σ_i to take the form $\alpha \mathbf{I}$. Here α a positive real-valued scalar shared by all clusters and $\mathbf{I}$ is the identity matrix.

(C) Restrict each Σ_i to take the from $\alpha_i \mathbf{I}$. Here α_i is a positive real-valued scalar and $\mathbf{I}$ is the identity matrix.

[4 marks]

[Total for Question 1: 18 marks]

Support Vector Machines (SVMs) and Kernels

Question 2

Let $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ be the training data for a binary classification problem, where $\mathbf{x}_i \in \mathbb{R}^d$ and $y_i \in \{-1, 1\}$. Let $\mathbf{w} \in \mathbb{R}^d$ be the parameter vector, $b \in \mathbb{R}$ be the offset, ξ_i be the slack variable for $i = 1, ..., n$.

Here the notation $\langle \mathbf{p}, \mathbf{q} \rangle = \mathbf{p} \cdot \mathbf{q}$ calculates the inner product of two vectors.

(a) What is wrong with the following primal form of the soft margin SVMs?

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

[2 marks]

(b) After fixing the problem in the above form, what is the estimated $\mathbf{w}$ if $C = 0$?

[2 marks]

(c) The dual form of the soft margin SVMs is given below. How to modify it (slightly) to make it become the dual form for the hard margin SVMs?

$$\max_{\alpha} \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \langle \mathbf{x}_i, \mathbf{x}_j \rangle$$
$$\text{s.t.} \quad 0 \leq \alpha_i \leq C, i = 1, \cdots, n$$
$$\sum_{i=1}^n \alpha_i y_i = 0$$

[2 marks]

(d) Express b using the dual variables and the training data.

[3 marks]

(e) A RBF kernel corresponds to lifting to a feature space with how many dimensions?

[3 marks]

(f) Let $\mathbf{u} = [\mathbf{w}; b]$ and $\mathbf{z} = [\mathbf{x}; 1]$. We can rewrite $(\langle \mathbf{w}, \mathbf{x} \rangle + b)$ as $\langle \mathbf{u}, \mathbf{z} \rangle$. This means if we augment the training data $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ to $\{(\mathbf{z}_i, y_i)\}_{i=1}^n$, where $\mathbf{z}_i = [\mathbf{x}_i; 1]$, we only need to learn one parameter $\mathbf{u}$ instead of two parameters $\mathbf{w}$ and b.

1. Please write down the primal form of the soft margin SVMs using decision function $\text{sign}[\langle \mathbf{u}, \mathbf{z} \rangle]$.

2. Is the new primal form equivalent to the old primal form? In other words, if we train two SVMs (standard SVM and this new re-parameterised SVM), in general, will we obtain exactly the same classification function?
3. Please prove your answer for above question (*i.e.* using derivation to show why or why not equivalent).

[6 marks]

(g) Suppose that we have a kernel $K(\cdot,\cdot)$ such that there is an implicit high-dimensional feature map $\Phi : \mathbb{R}^d \to \mathbb{R}^D$ that satisfies $\forall \mathbf{x}, \mathbf{z} \in \mathbb{R}^d$, $K(\mathbf{x}, \mathbf{z}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{z}) \rangle = \Phi(\mathbf{x})^\top \Phi(\mathbf{z}) = \sum_{i=1}^{D} \Phi(\mathbf{x})_i \Phi(\mathbf{z})_i$ is the inner product in the D-dimensional space.

Show how to compute the squared ℓ_2 distance in the D-dimensional space:

$$||\Phi(\mathbf{x}) - \Phi(\mathbf{z})||^2 = \sum_{i=1}^{D} (\Phi(\mathbf{x})_i - \Phi(\mathbf{z})_i)^2$$

without explicitly calculating the values in the D-dimensional vectors. You are asked to provide a formal proof.

[6 marks]

[Total for Question 2: 24 marks]

Boosting

Question 3

(a) **True or False**

AdaBoost must give zero training error regardless of the type of weak classifiers it uses, provided enough iterations are performed.

[3 marks]

(b) The AdaBoost algorithm has two drawbacks. Answer the following questions regarding these.

(I) Show mathematically why a weak learner with $< 50\%$ predictive accuracy presents a problem to AdaBoost.

(II) AdaBoost is susceptible to outliers. Suggest a simple heuristic that may alleviate this.

[6 marks]

(c) Assume that the weak learners are a finite set of decision stumps. We then train a AdaBoost classifier. Can the boosting algorithm select the same weak classifier more than once? Explain.

[5 marks]

[Total for Question 3: 14 marks]

Neural Networks

Question 4

(a) Which one statement is true about neural networks? (Select the single best answer)

(A) We always train neural networks by optimising a convex cost function.

(B) Neural networks are more robust to outliers than support vector machines.

(C) Neural networks always output values between 0 and 1.

(D) A neural network with a large number of parameters often can better use big training data than support vector machines.

[3 marks]

(b) Which of the following statements is (are) true about neural networks?

(A) The training time depends on the size of the network as well as the training data.

(B) The perceptron is a single layer recurrent neural network.

(C) In image processing, compared with fully connected networks, usually convolutional networks are preferred.

(D) Neural network cannot be used for solving Regression problems.

[3 marks]

(c) **True or False:** The training strategy "back propagation" in neural networks is essentially to calculate gradients of the network parameters using the chain rule.

[3 marks]

(d) Training a convolutional neural network for speech recognition, one finds that performance on the training set is very good while the performance on the validation set is unacceptably low. A reasonable fix might be to: (Select the single best answer)

(A) Decrease the weight decay

(B) Reduce the training set size

(C) Reduce the number of layers and neurons

(D) Increase the number of layers and neurons

[4 marks]

(e) In neural networks, nonlinear activation functions such as sigmoid, tanh, and ReLU

(Select the single best answer)

(A) speed up the gradient calculation in back propagation as compared to linear units.

(B) help to learn nonlinear decision boundaries

(C) always output values between 0 and 1
(D) are applied only to the output units

[4 marks]

[Total for Question 4: 17 marks]

Regression

Question 5

(a) Linear regression solves the following optimisation problem,

$$\min_{\mathbf{w}} \|X\mathbf{w} - \mathbf{y}\|^2,$$

where X is the data matrix (each row is a data vector), $\mathbf{y}$ is the label column vector, and $\mathbf{w}$ is the parameter vector to be learned.

Note that we have omitted the offset parameter b here.

Write down the solution of linear regression (i.e., the optimal $\mathbf{w}$).

[3 marks]

(b) Write down the formulation of support vector regression (both Primal and Dual).

[3 marks]

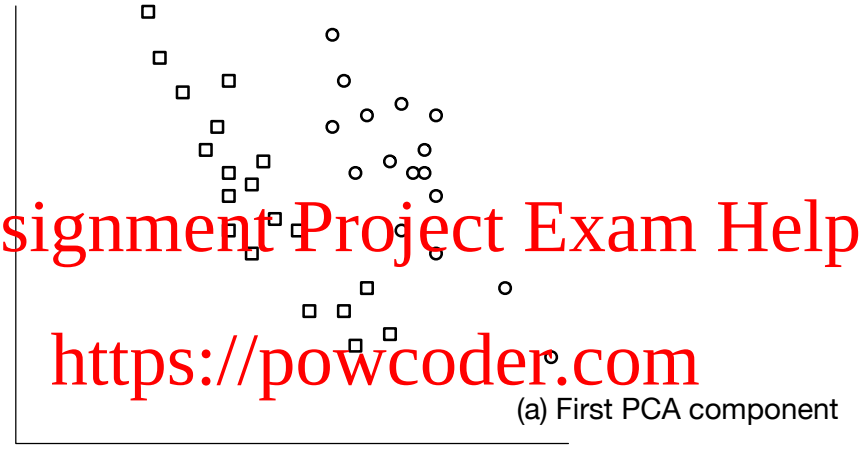
[Total for Question 5: 6 marks]

Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)

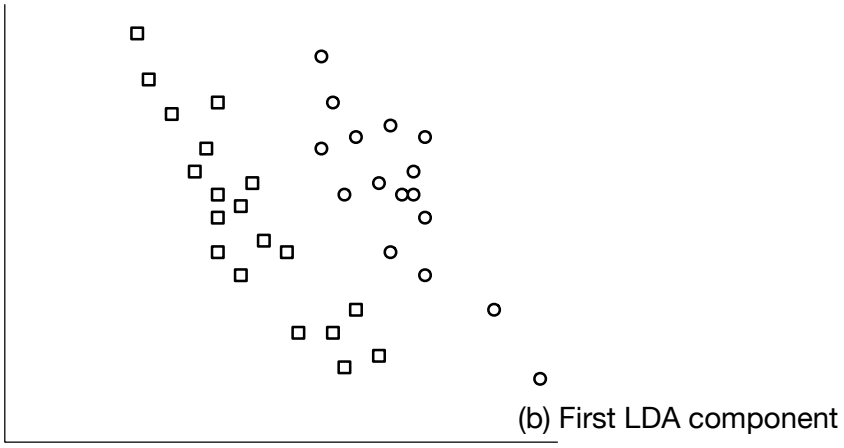
Question 6

In this problem two linear dimenionality reduction methods will be discussed. They are principal component analysis (PCA) and linear discriminant analysis (LDA).

(a) LDA reduces the dimensionality given labels by maximising the overall interclass variance relative to intraclass variance. Plot the directions (roughly) of the first PCA and LDA components in the following figures respectively. In the figures, squares and circles represent two different classes of data points.



(a) First PCA component



(b) First LDA component

[5 marks]

(b) In a supervised binary classification task we have a total of 15 features, among which only 4 are useful for predicting the target variable, the other features are pure random noise with very high variance. What complicates matters even worse is that the 4 features when considered individually show no predictive power, and only

work when considered together.

Consider each of the following dimension reduction (or classification techniques) and indicate whether it may be able to successfully identify the relevant dimensions (Yes or No). Briefly explain why.

(I) Principle Component Analysis

(II) Linear Discriminant Analysis

(III) AdaBoost with decision stumps as the weak learner

(IV) AdaBoost with a linear support vector machine as the weak learner

[16 marks]

[Total for Question 6: 21 marks]