QUEEN'S UNIVERSITY BELFAST

CSC4007

Exam Time Table Code CSC4007

Calculator permitted

Level 4 Examination for the Degree(s) of Computer Science and related pathways

Advanced Machine Learning

Tuesday, 24th April 2018 2:30 PM - 4:30 PM

Examiners: Dr V. Ngo
Prof. F. Moller
and the internal examiners

Exam Duration TWO Hours

Answer ALL questions

Question 1. Regression

(a) The following function is the cost function for regression:

$$L(\beta)=\sum_{i=1}^{n}\left(y_i-\phi(x_i)^T\beta\right)^2+\lambda\sum_{j=1}^{k}\beta_j^2$$

(i) Use your own words to state the meaning of this cost function and the role of each term?

[4 marks]

(ii) What is the effect on the cost function when λ goes to $+\infty$ and when λ goes to 0 ?

[2 marks]

(b) Table 1 shows the grades of students in their final examinations in maths and physics.

Student	Maths grade	Physics grade
1	80	85
2	40	45
3	60	65
4	90	95
5	70	75

TABLE 1. The grades of students in their examinations in maths and physics.

Based on this data there is a hypothesis that students' grades on these two modules are correlated. Assuming that we use a linear model, $y=ax+b$, to capture the relationship between a grade in maths and one in physics, where *x* denotes the maths grade and *y* denotes the physics grade:

(i) Find optimum values a and b using linear regression to fit the data in Table 1.
[Hint: optimum parameters in linear regression are $\beta=(X^TX)^{-1}X^TY$]

[5 marks]

(ii) What is the square error of the solution found in **i)**?

[1 marks]

(iii) Predict the performance in physics of a student who achieved 30 marks in maths.

[1 marks]

[continued overleaf]

(iv) How would you modify the algorithm so that you can predict the performance of a student in maths when knowing their performance in physics?

[2 marks]

Question 2. Classification

(a) Regression and classification are two main supervised learning problems in machine learning.

(i) Briefly describe the difference between regression and classification tasks. Give an example real-world problem for regression, and one for classification.

[3 marks]

(ii) Describe in detail the *k*-nearest neighbor algorithm.

[3 marks]

(b) Assume an engineer wants to build their own movie recommendation system. Table 2 shows a data set of movies rated by the engineer's parents, together with whether the engineer liked the movies or not.

Movie name	**Mum's rating**	**Dad's rating**	**Like?**
The Lord of the Rings	5	9.5	Yes
Casablanca	9	6	No
A Beautiful Mind	7	7	Yes
Gone with the Wind	7	5	No

TABLE 2. A dataset of movie rating

(i) Formulate the above system as a classification task and describe the problem setting (assigning the correct values to the dataset and input and output variables).

[2 marks]

(ii) Write four discriminative linear feature vectors $\phi(x,y)$ corresponding to the given four data points.

[2 marks]

(iii) Given a linear model with $\beta=[1,1,0,1,0,1]$, use the discriminative linear features from **ii)** to predict ***whether the engineer will like a new movie, "The Good, the Bad and the Ugly", if its ratings are 7.5 (by Mum) and 5.5 (by Dad).***

[3 marks]

[continued overleaf]

(iv) Instead of predicting whether the engineer likes the movie or not, modify the above system in order to predict Dad's rating given Mum' rating and whether the engineer liked the movie or not. Assuming that you can convert categorical data into real values (No = -1, Yes = 1). Speculate as to how this might be achieved.

[2 marks]

Question 3. Kernel Ridge Regression & Support Vector Machine

(a)

(i) Describe what is the kernel trick formally.

[1 mark]

(ii) Given a kernel $k(x,x')=(x_1x'_1+x_2x'_2+1)^2$ between any data points x,x' , where $x=[x_1,x_2]$ and $x'=[x'_1,x'_2]$, compute the equivalent feature ϕ derived from k [Hints: ϕ should satisfy that $k(x,x')=\phi(x)^T\phi(x')$]

[2 marks]

(iii) Figure 1 visualizes a dataset for a binary classification task in 2D (one output is labeled with a square, the other is labeled with a circle).

- Draw the best straight-line (the decision boundary) by your guess to separate these data that might result from a support vector machine (a max-margin method)
- Draw the corresponding positive and negative lines
- Draw a circle around the corresponding support vectors, and justify your solutions of those support vectors.

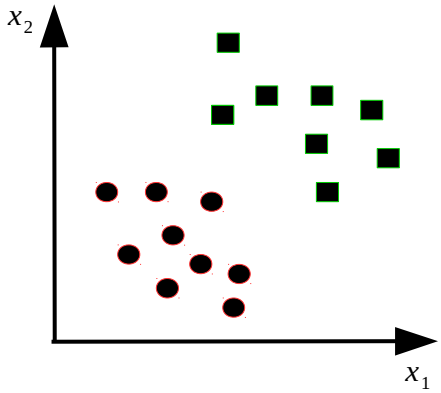

FIGURE 1. A visualization of a binary dataset

[4 marks]

(b) Assume that you want to build a system for vehicle sale prediction for a newly expanded market. You obtain some data in two countries that are Korea and Indonesia, and want to predict the potential sale in Vietnam. The data in Table 3 has inputs in forms of (GDP per capita of that country, population, and the percentage of car ownership).

[continued overleaf]

Country	GDP	Population	Ownership percentage	Vehicle sale in numbers
Korea	30	76	45	1.0
Indonesia	0.4	260	7	1.2
Vietnam	0.2	90	2.5	???

TABLE 3: Vehicle sale corresponding to a country's GDP, population, and ownership percentage.

(i) Assume that we use the linear kernel $k(x,x')=x^T x'+1$, and $\phi(x)$ is the feature derived from $k(x,x')=x^T x'+1$ via the kernel trick. Compute $\phi(x)$ for all three given input points.

[4 marks]

(ii) Use kernel ridge regression to predict the potential vehicle sale for Vietnam, given a linear kernel $k(x,x')=x^T x'+1$, and a regularization parameter $\lambda=0.5$.

[Hints: the formula for kernel ridge regression is $f(x)=\kappa(x)^T(K+\lambda I)^{-1}Y$ where $\kappa(x)$ is a vector as $\kappa(x)=[k(x,x_1),k(x,x_2)]^T$, the matrix K has entries $K_{ij}=k(x_i,x_j)$, Y is a vector of outputs, and I is the identity matrix.]

[4 marks]

Question 4. Unsupervised Learning & Clustering

(a) Supervised learning and unsupervised learning are two main tasks in machine learning.

(i) Explain in your own words what is unsupervised learning.

[1 mark]

(ii) Describe the difference between supervised and unsupervised learning. Give one technique that can do supervised learning and one that can do unsupervised learning.

[3 marks]

(iii) Formally describe the *k*-means clustering algorithm in detail.

[3 marks]

(b) Given 6 data points (x1, x2, x3, x4, x5, x6), where x1 = (1, 1), x2 = (1, 3), x3 = (3, 2), x4 = (5, 4), x5 = (5, 7), x6 = (7, 6).

[continued overleaf]

(i) Using k-means clustering: assuming that the two centroids currently at (2, 1) and (6, 4) (for k=2), what would be the coordinates of these two centroids at the next iteration?

[4 marks]

(ii) What would happen to the centroids (are they still moving?) and the algorithm (should it stop?) if you apply a further iteration of k-means clustering?

[4 marks]