Queen Mary
University of London

e.g. **BSc Examination by course unit/ BA by Special Regulations/ MSc Examination**

Main Examination period 2017

ECS607U/ECS766P Data Mining ***Duration: 2 hours 30 minutes***

YOU ARE NOT PERMITTED TO READ THE CONTENTS OF THIS QUESTION PAPER UNTIL INSTRUCTED TO DO SO BY AN INVIGILATOR

Answer ALL questions

Calculators are permitted in this examination. Please state on your answer book the name and type of machine used.

Complete all rough workings in the answer book and cross through any work that is not to be assessed.

Possession of unauthorised material at any time when under examination conditions is an assessment offence and can lead to expulsion from QMUL. Check now to ensure you do not have any notes, mobile phones, smartwatches or unauthorised electronic devices on your person. If you do, raise your hand and give them to an invigilator immediately. It is also an offence to have any writing of any kind on your person, including on your body. If you are found to have hidden unauthorised material elsewhere, including toilets and cloakrooms it will be treated as being found in your possession. Unauthorised material found on your mobile phone or other electronic device will be considered the same as being in possession of paper notes. A mobile phone that causes a disruption in the exam is also an assessment offence.

EXAM PAPERS MUST NOT BE REMOVED FROM THE EXAM ROOM

Examiners: Dr. Ioannis Patras and Dr. Antony Constantinou

Question 1

a) Making reference to the performance on training and testing datasets, briefly explain the over and under-fitting pitfalls in supervised learning.

Answer: *Under-fitting: Model is too simple. Performs badly on both train and test data. Over-fitting: Model is too complex. Performs great or perfectly on train, badly on test.*

[7 marks]

b) List three factors that make over-fitting more likely in supervised learning.

Answer: *Any three of: Limited amounts of train data. High dimension of train data. Complex/powerful model (such as KNN/Decision Tree). No regularization or weak regularization.*

[6 marks]

c) Briefly explain how over-fitting is controlled in one of the supervised learning methods that you have studied (e.g., KNN, Linear Regression, etc).

Answer: *Explanation for any model we covered, for example: KNN: Having K>1 smooths the decision boundary of the classifier, reducing overfitting. DT: Pruning excess nodes/branches simplifies the classifier, reducing overfitting. Polynomial Regression: A penalty on the magnitude of the weights results in smaller weights, reducing overfitting.*

[6 marks]

d) Briefly outline the role of validation data in avoiding both over and under-fitting.

Answer: *Too simple (or strongly regularized) model under-fits, and too complex (or weakly regularized) model over-fits. The regularization parameter (e.g.,K, pruning strength, lambda) can be tuned to balance these two for best generalization. This should be done by evaluating the model's performance on validation data for a variety of regularization strengths. The regularization strength that best validation performance is picked.*

[6 marks]

[Q1 total 25 marks]

Question 2

a) Compare and contrast the KNN and MaxEnt (aka Logistic Regression) approaches to classification.

Answer: *Decision Boundary: KNN: Non-linear, Maxent: Linear. Regularization strategy: K>1, MaxEnt penalize high magnitude weights. KNN: Train by storing all the data, MaxEnt: train by finding the line that best separates the data. KNN: Classify by finding the nearest neighbour to test point and using that label. MaxEnt: Classify according to which side of the line test point is one. Test Complexity: O(ND), Test Complexity: O(D).*

[7 marks]

b) The task of learning a non-linear regression model is sometimes expressed as:
$\underset{\mathbf{w}}{\operatorname{argmin}} \sum_i (y_i - \mathbf{w}^T \varphi(\mathbf{x}_i))^2 + \lambda \mathbf{w}^T \mathbf{w}$.

i. Explain each of the symbols in the above equation, explaining what is given as input to the algorithm and what needs to be learned and/or determined.
ii. Briefly outline what this means in English.
iii. How should the value of λ be determined?

Answer:

i. $(\mathbf{x}_i, y_i)$ *is a pair of training datum, y_i is the target and x_i the input.* $\phi(\mathbf{x}_i)$ are the non-linear features. W are the parameters to be estimated. *Lambda is the parameter that trades off the importance of fit to the data (mean squared error), and simple model (sum square of the weights). [4 points]*

ii. Find best (argmin) line (specified by weights w) that minimises the squared error between the training points and prediction (first term) as well as the magnitude of the weights (last term). [4 points]

iii. Lambda should be determined by optimising performance on a validation set, or by cross-validation. [4 points]

[12 marks]

c) Explain one of the greedy feature selection methods, that is, either forward or backward feature selection.

Answer:

Forward feature selection:

i. Start with k =0 features
ii. For each of the remaining D – k features
    a. retrain the classifier/regressor with the k selected features + the current one

b. Select the feature that gives the small cost / highest accuracy. Add it to the set of the selected features. This set has now k+1 elements.

iii. Repeat until the addition of a feature does not improve the cost/accuracy

[6 marks]

[Q2 total 25 marks]

Question 3

a) When assessing a supervised learning method, the resulting accuracy of prediction on test data is obviously very important. List five additional factors that are also useful to assess the learning procedure to decide its suitability for a particular application.

Answer

Any five of: Computational complexity at train time, complexity at test time, memory requirement at train time, memory requirement at test time, robustness to missing data, robustness to outliers, robustness to irrelevant dimensions, resistance to overfitting, which specific classes are mistaken (confusion matrix).

[5 marks]

b) Label the following applications according to their suitability for being treated as a classification or regression problem in supervised learning, or neither.

i. Predicting whether a customer will default on their loan.

ii. Predicting stock market index.

iii. Expected profit on a financial transaction.

iv. Predictive text on a mobile phone.

v. Person recognition at an access gate.

vi. Compressing the number of bytes required to store an image.

vii. Steering system on a self driving car.

viii. Road pedestrian detector in a self driving car.

ix. Deciding suitable sizes for a line of bicycle factory given height of the population.

x. Prediction of fuel consumption for a flight.

Answer

i. Predicting whether a customer will default on their loan. C

ii. Predicting stock market index. R

iii. Expected profit on a financial transaction. R

iv. Predictive text on a mobile phone. C

v. Person recognition at an access gate. C

vi. Compressing the number of bytes required to store an image. N

vii. Steering system on a self driving car. R

viii. Road pedestrian detector in a self driving car. C

ix. Deciding suitable sizes for a line of bicycle factory given height of the population. N

x. Prediction of fuel consumption for a flight. R

[5 marks]

c) Describe how a decision tree is used for classification. How is overfitting avoided during training?

Answer

During testing a datum x is propagated down the tree by making a decision at each node on whether it will follow the left or the right branch. This decision is based on the outcome of a test, typically performed on a single feature of the datum x. When the datum arrives at a leaf note the datum is assigned to the majority class, i.e. the class to which most of the examples in the training set that arrived at the leaf in question belong. [4points]

Overfitting is avoided by pruning (either by merging or by early stopping). [3 points]

[7 marks]

d) Explain the role of the objective function in machine learning. Illustrate your answer by describing the objective function of your favourite learning algorithm.

Answer

The objective function formalizes what is trying to be achieved in a learning problem, and quantifies how well a given model achieves it. It can be thought of as a routine that takes as input the dataset and the model parameters; and returns as output a quantitative measure of how well the model with the specified parameters fits with the data. (E.g., MSE between regression prediction and data points).

Most learning algorithms can be thought of as trying to find the parameters that minimize (or maximize as appropriate) the objective function. In some cases this may literally involve some kind of search over the parameter space (e.g., Decision Tree), in other cases there may be an exact solution (e.g., regression, Naïve Bayes).

[8 marks]

[Q3 total 25 marks]

Question 4

a) A doctor in Brazil can run a test for *Yellow fever*. The test has two possible outcomes: positive and negative. If *Yellow fever* is present, the test comes out positive 90% of the time. Among the population, *Yellow fever* is known to occur in 1% of all people, and on average 10% of people test positive for *Yellow fever*. Juan enters the clinic and tests positive for the disease. What is the probability that he really has Yellow fever? (You may wish to recall that Bayes theorem is: $p(A|B) = \frac{p(B|A)p(A)}{p(B)}$.)

Answer: *p(Y)=0.01, p(!Y)=0.99, p(+)=0.2, p(-)=0.8, p(+|Y)=0.9, p(-|Y)=0.1, p(Y|+)=p(+|Y)p(Y) / (p(+)) = 0.9*0.01/0.1 =0.09*

[7 marks]

b) You are assigned the task of deciding which pedestrian detection algorithm you will adopt in your self driving car company. Two vendors A and B submit pre-trained methods for consideration. They provide equal accuracy. In terms of test-time computational efficiency, applying them on images of very small resolution images reveals that the algorithm offered by vendor A takes 1 minute to run, and the algorithm offered by vendor algorithm B takes 2 minutes to run. The two algorithms are known to have $O_A(ND^2)$ and $O_B(ND)$ computational complexity respectively, for N data instances, and D dimensions. Which of these is more likely to be faster when installed on the car where a very high resolution camera will be used? Why?

Answer: *Although model B is slower so far, it will scale better as the dimension of the data (image resolution) increases, so it is likely to be preferable.*

[6 marks]

c) Learning models can be trained and applied on different computers. Deploying a learned model for test-time execution in an embedded system or mobile application often provides the constraint of limited memory.
   i. Which supervised learning models that you have studied might this constraint rule out, and why?
   ii. Suppose that a particular application needs a non-linear classifier, i.e., linear classifiers like MaxEnt (aka Logistic Regression) are ruled out. What other non-linear classifier might be suitable?

Answer: *This will rule out KNN, because it requires to store the entire training dataset. In this case decision tree could be suitable, due to being potentially compact to store, but also non-linear.*

[6 marks]

d) Suppose you are applying k-means clustering for market-segmentation in ecommerce. A co-worker suggests that you should re-run the clustering algorithm multiple times. Is this a useful thing to do? Why?

Answer: *Yes. This makes sense because K-means converges to a local minimum only. So different runs from different initial conditions/random seeds will give different answers. Repeating runs and taking the best of them increases the chance of finding a good solution.*

[6 marks]

[Q4 total 25 marks]

End of Paper