University of Glasgow

Thursday 9 May 2019
2.00 pm – 4.00 pm
(Duration: 2 hours)

DEGREES OF MSc, MSci, MEng, BEng, BSc,MA and MA (Social Sciences)

Machine Learning (M)

(Answer 3 questions in Section A and all of Section B)

This examination paper is worth a total of 60 marks

The use of a calculator is not permitted in this examination

INSTRUCTIONS TO INVIGILATORS
Please collect all exam question papers and exam answer scripts and retain for school to collect. Candidates must not remove exam question papers.

1. Assume a linear regression model of the form

$$t_n = \sum_{d=1}^{D} w_d\, x_{n,d}, n = 1, \ldots, N$$

(a) Show that the least square solution for estimating model parameters is

$$\mathbf{w} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{t}$$

Please give clear definition of the vector/matrix format variables and start the derivation from writing down the squared loss function.

[6 marks]

(b) The radial basis function (RBF):

$$h_{n,k} = \exp\left(-\frac{\sum_{d=1}^{D}(x_{n,d} - \mu_{d,k})^2}{2s^2}\right), n = 1, \ldots, N;\ k = 1, \ldots, K$$

is a popular choice for converting the original features, $x_{n,d}$, into a new set of K features prior to training. Assume the value of s is given. Describe a procedure for determining the center parameter $\mu_{d,k}$ and K.

[2 marks]

(c) It is also common to normalise the input features by a procedure called *whitening* prior to training. Specifically, $\tilde{x}_{n,d} = \frac{x_{n,d} - \mu_d}{\sigma_d}$, where μ_d and σ_d are the mean and the standard deviation of the *d*th feature. Assuming $x_{n,d}$ follows a Gaussian distribution, what is the distribution of $\tilde{x}_{n,d}$?

[2 marks]

(d) With respect to the functions they can fit, describe the difference between RBF and whitening as pre-processing strategies (you might find a graph useful to answer this question).

[2 marks]

(e) Sketch a graph demonstrating an example of overfitting in linear regression. Also state the type of the transformation that might cause the problem e.g. polynomial basis functions.

[3 marks]

2. Bayesian inference question

(a) Consider the linear regression problem in question 1, the posterior density over **w** can be obtained via Bayes rule:

$$p(\mathbf{w}|\mathbf{t}, \mathbf{X}) = \frac{p(\mathbf{t}|\mathbf{X}, \mathbf{w})p(\mathbf{w})}{p(\mathbf{t}|\mathbf{X})}$$

State the name of each term in this expression and provide a description of its role, or what it could be used for.

[8 marks]

(b) Let $p(\mathbf{t}|\mathbf{X},\mathbf{w}) = \mathcal{N}(\mathbf{Xw}, \sigma^2\mathbf{I})$, what is the corresponding mathematical model relating the target variable and input features. You can choose to use the vector/matrix form **t and X, or the scalar form** t_n and $x_{n,d}$? (i.e. $\mathbf{t}$ = ? or t_n **=?).** Please also state the distribution of any noise in the model.

[2 marks]

(c) When sampling from the distribution $p(\mathbf{w}|\mathbf{t},\mathbf{X})$ using a Metropolis-Hastings sampling scheme with proposal density $q(\mathbf{w}^*|\mathbf{w})$, we compute the following ratio for a proposed sample ($\mathbf{w}^*$):

$$r = \frac{p(\mathbf{w}^*|\mathbf{t},\mathbf{X})}{p(\mathbf{w}|\mathbf{t},\mathbf{X})}\frac{q(\mathbf{w}|\mathbf{w}^*)}{q(\mathbf{w}^*|\mathbf{w})}$$

what happens if the **ratio** is less than 1?

A. The proposed sample is rejected, and the previous sample is added to our set of samples.

B. The proposed sample is accepted with probability equals to r and, if rejected, the previous sample is added again to our set of samples.

C. The proposed sample is accepted and added to our set of samples.

D. The proposed sample is rejected, and nothing is added to our set of samples.

[2 marks]

(d) Describe two advantages of using the Bayesian method over maximum likelihood and one disadvantage.

[3 marks]

3. Classification question

(a) Classification and regression are both supervised learning problems. Describe a way to turn a regression problem into a classification problem. (Please state the differences between the two.)

[3 marks]

(b) The receiver operating characteristic (ROC) curve is a standard way to visualize the performance of classifiers. Which one of the following statements about ROC is NOT true?

A. Points on a ROC curve are created by varying the threshold of a score at which the classifier calls something as belonging to the positive class.

B. Points on a ROC curve are created by choosing ONE optimal value as the threshold of a score at which the classifier calls something as belonging to the positive class

C. The x-axis of ROC curve represents *1- Specificity.* Note: Specificity = $\frac{\text{number of true positives}}{\text{number of true negatives+number of false positives}}$

D. The x-axis of ROC curve represents *False Positive Rate.* Note: $\text{False Positive Rate} = \frac{\text{number of false positives}}{\text{number of true negatives+number of false positives}}$

[2 marks]

(c) Describe a way to compute a **real-valued** score to assign a data point to the positive class for a *k-nearest neighbors* (KNN) classifier.

[2 marks]

(d) Two binary classifiers are used to make predictions for the same set of six test points. These predictions are given below, along with the true labels. In each case, sketch its ROC curve and compute its area under the curve (AUC).

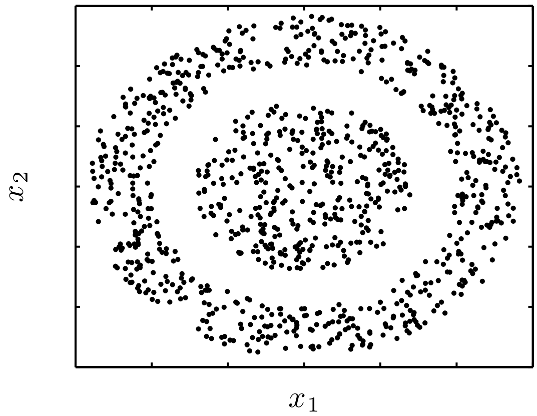
Classifier 1		Classifier 2	
Predicted probability of class 1 (Score of class 1)	True class	Predicted probability of class 1 (Score of class 1)	True class
1	1	0.8	1
0.8	1	0.8	0
0.6	1	0.6	1
0.4	0	0.6	0
0.2	0	0.2	1
0.0	0	0.2	0

[6 marks]

(e) Explain what overfitting is in the context of classification (you might find a graph useful to answer this question).

[2 marks]

4. Consider using the K-means algorithm to perform clustering on the following data.



We want to cluster data in the outer ring to be one cluster and the data in the inner circle as a different cluster.

(a) Outline what would happen if we directly apply K-means to this data. Can it achieve the clustering objective? How will it split/group the data?

[2 marks]

(b) An alternative approach is to use *Kernel K-means.* Explain how kernel could help in this dataset.

[3 marks]

(c) Given that the squared Euclidean distance between data point $\mathrm{x_n}$ and the kth cluster center $\boldsymbol{\mu}_k$is $(\boldsymbol{x}_n - \boldsymbol{\mu}_k)^T(\boldsymbol{x}_n - \boldsymbol{\mu}_k)$, and $\boldsymbol{\mu}_k = \frac{1}{n_k}\sum_{i=1}^{N} z_{ik}\, \boldsymbol{x}_i$, show that the kernelised equivalent is the following:

$$K(\boldsymbol{x}_n, \boldsymbol{x}_n) - \frac{2}{n_k}\sum_{i=1}^{N} z_{ik}\, K(\boldsymbol{x}_n, \boldsymbol{x}_i) + \frac{1}{n_k^2}\sum_{i=1}^{N}\sum_{j=1}^{N} z_{ik}\, z_{jk} K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$

where n_k is the number of objects assigned toμ_k, and $z_{nk}$is 1 if object n is assigned to cluster k and 0 otherwise.

[4 marks]

(d) Which one of the following statements about kernels is NOT correct?

A. One could use the RBF kernel, $K(\boldsymbol{x}_n, \boldsymbol{x}_i) = \exp(-\gamma(\boldsymbol{x}_n - \boldsymbol{x}_i)^T(\boldsymbol{x}_n - \boldsymbol{x}_i))$, to achieve the clustering target.

B. The RBF kernel projects the data onto an infinite dimensional space. The free parameter γ can be estimated with cross-validation.

C. One could use the linear kernel, $K(\boldsymbol{x}_n, \boldsymbol{x}_i) = \boldsymbol{x}_n^T \boldsymbol{x}_i$, to achieve the clustering target.

D. The linear kernel projects the data onto itself, and there is no free parameter to tune.

[2 marks]

(e) Write some pseudo-code to perform kernelised K-means.

[4 marks]

Section B

5. This question is based on '*Efficient L1 Regularized Logistic Regression. Su-In Lee, Honglak Lee, Pieter Abbeel and Andrew Y. Ng. AAAI '06.*

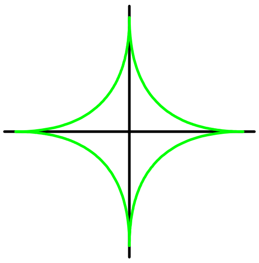
(a) Briefly summarise the contribution of this paper.

[2 marks]

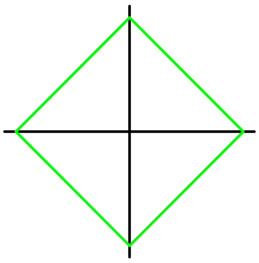
(b) The *L1* regularised logistic regression in the paper aims to estimate unknown parameters by minimizing the following loss function:

$$\min_{\mathbf{w}} \sum_{n=1}^{N} -\log p(y_n = 1|\mathbf{x}_n, \mathbf{w}) + \lambda \sum_{d=1}^{D} |\, w_d|$$

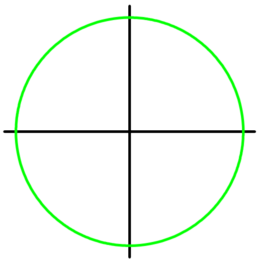
Which one of the following graphs represent the *L1* regularisation term in 2D?



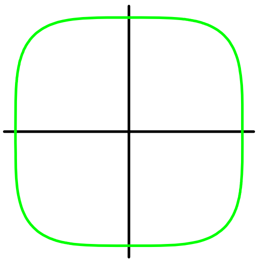
A.



B.



C.



D.

[2 marks]

B

(c) The likelihood function used in the paper is the probability of label being positive (i.e. 1) conditioning on features and parameters:

$$\sum_{n=1}^{N} -\log p(y_n = 1|\mathbf{x}_n, \mathbf{w}) = \sum_{n=1}^{N} \log\left(\frac{1}{1 + \exp(-\mathbf{w}^T\mathbf{x}_n)}\right) = \sum_{n=1}^{N} \log \sigma(\mathbf{w}^T\mathbf{x}_n)$$

This likelihood function is NOT a probability distribution, because it does not directly describe the other possible state, i.e. $y_n = 0$. Which one of the following statements is NOT true about the distribution of y_n

A. The distribution of y_n can be written as

$$p(y_n|\mathbf{x}_n, \mathbf{w}) = \sigma(\mathbf{w}^T\mathbf{x}_n)^{y_n}(1 - \sigma(\mathbf{w}^T\mathbf{x}_n))^{1-y_n}$$

B. The probability of $y_n = 0$ can be written as

$$p(y_n = 0|\mathbf{x}_n, \mathbf{w}) = \frac{1}{1 + \exp(\mathbf{w}^T\mathbf{x}_n)}$$

C. The summation in the likelihood is a result of assuming independence between data points.

D. Minimising the likelihood has a closed-form solution

[2 marks]

(d) Another popular regularisation method is the L2 regularisation, $\lambda \sum_{d=1}^{D} w_d^2$. Describe the different in effect of the two approaches on parameter estimation and their benefits.

[4 marks]

(e) Describe an approach to estimate λ

[2 marks]

(f) Unsupervised learning approaches can also be used to reduce the dimensionality of features. Give an example of such method and explain how it reduce the number of features.

[3 marks]