ELL784/EEL709: Introduction to Machine Learning

Minor Test I, Form: A (**please write this Form ID on the cover page of your answer script**)

Maximum marks: 20

Section 1. Multiple choice questions

Each question may have any number of correct answers, including zero. List all choices you believe to be correct (1 mark for each correct answer, -0.5 for each incorrect answer). No justification is required.

1. Consider a binary classification problem. Suppose I have trained a model on a linearly separable training set, and now I get a new labeled data point which is correctly classified by the model, and far away from the decision boundary. If I now add this new point to my earlier training set and re-train, in which cases is the learnt decision boundary likely to change?
   (a) When my model is a perceptron.
   (b) When my model is logistic regression.
   (c) When my model is an SVM.
   (d) When my model is Gaussian discriminant analysis.
2. When doing least-squares regression with regularisation (assuming that the optimisation can be done exactly), increasing the value of the regularisation parameter λ
   (a) will never decrease the training error.
   (b) will never increase the training error.
   (c) will never decrease the testing error.
   (d) will never increase the testing error.
   (e) may either increase or decrease the training error.
   (f) may either increase or decrease the testing error.
3. Which of the following points would Bayesians and frequentists disagree on?
   (a) The use of a non-Gaussian noise model in probabilistic regression.
   (b) The use of probabilistic modelling for regression.
   (c) The use of prior distributions on the parameters in a probabilistic model.
   (d) The use of class priors in Gaussian Discriminant Analysis.
   (e) The idea of assuming a probability distribution over models.
4. Regarding bias and variance, which of the follwing statements are true? (Here 'high' and 'low' are relative to the ideal model.)
   (a) Models which overfit have a high bias.
   (b) Models which overfit have a low bias.
   (c) Models which underfit have a high variance.
   (d) Models which underfit have a low variance.
5. Which of the following are characteristics of data sampled from a Gaussian distribution?
   (a) The sample mean systematically underestimates the true mean.
   (b) The sample variance systematically overestimates the true variance.
   (c) Both the sample mean and variance are unbiased estimators of the true values.

6. Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting?

   (a) Increase the amount of training data.

   (b) Improve the optimisation algorithm being used for error minimisation.

   (c) Decrease the model complexity.

   (d) Reduce the noise in the training data.

7. You are reviewing papers for the World's Fanciest Machine Learning Conference, and you see submissions with the following claims. Which ones would you consider accepting?

   (a) My method achieves a training error lower than all previous methods!

   (b) My method achieves a test error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise test error.)

   (c) My method achieves a test error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise cross-validaton error.)

   (d) My method achieves a cross-validation error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise cross-validaton error.)

Section 2. Numerical questions

Please show all steps in your working fully and clearly, except where indicated otherwise.

8. The Indian Railways have been trialling 2 different machine learning methods which attempt to predict whether a train will arrive at its final destination on time or not, using a number of input features corresponding to weather conditions, train priorities, ongoing repair works etc. (for this purpose, 'on time' is defined as no more than 10 minutes after its scheduled time). The methods have been tested on a common set of 500 train runs, and the results are as follows:

	Actually on time	**Actually late**
Method 1 predicted on time	131	155
Method 1 predicted late	19	195
Method 2 predicted on time	82	72
Method 2 predicted late	68	278

Suppose we set up a simple probabilistic model for this as follows: θ is the prior probability of a train being late; p is the probability of a late prediction from Method 1 if the train is on time (also called the *False Positive Rate (FPR)*); and q is the probability of a late prediction from Method 1 if the train is in fact late (also called the *True Positive Rate (TPR)*).

(a) Write down the joint likelihood of the data for Method 1, as a function of the three model parameters θ, p, and q. Obtain maximum likelihood estimates for each of these parameters. **[4]**

(b) Suppose the loss matrix for this prediction task is defined as follows:

	Actually on time	**Actually late**
Predicted on time	0	1
Predicted late	K	0

Using the parameter estimates computed above, obtain the expected loss for Method 1 as a function of K. **[2]**

(c) Obtain the expected loss for Method 2 as well (you can compute its FPR and TPR directly, without doing the maximum likelihood derivations again); which is the preferable method? What is the critical value of K at which this preference changes? **[4]**

Answer Key for Exam A

Section 1. Multiple choice questions

Each question may have any number of correct answers, including zero. List all choices you believe to be correct (1 mark for each correct answer, -0.5 for each incorrect answer). No justification is required.

1. Consider a binary classification problem. Suppose I have trained a model on a linearly separable training set, and now I get a new labeled data point which is correctly classified by the model, and far away from the decision boundary. If I now add this new point to my earlier training set and re-train, in which cases is the learnt decision boundary likely to change?

   (a) When my model is a perceptron.
   [(b)] When my model is logistic regression.
   (c) When my model is an SVM.
   [(d)] When my model is Gaussian discriminant analysis.

2. When doing least-squares regression with regularisation (assuming that the optimisation can be done exactly), increasing the value of the regularisation parameter λ

   [(a)] will never decrease the training error.
   (b) will never increase the training error.
   (c) will never decrease the testing error.
   (d) will never increase the testing error.
   (e) may either increase or decrease the training error.
   [(f)] may either increase or decrease the testing error.

3. Which of the following points would Bayesians and frequentists disagree on?

   (a) The use of a non-Gaussian noise model in probabilistic regression.
   (b) The use of probabilistic modelling for regression.
   [(c)] The use of prior distributions on the parameters in a probabilistic model.
   (d) The use of class priors in Gaussian Discriminant Analysis.
   [(e)] The idea of assuming a probability distribution over models.

4. Regarding bias and variance, which of the follwing statements are true? (Here 'high' and 'low' are relative to the ideal model.)

   (a) Models which overfit have a high bias.
   [(b)] Models which overfit have a low bias.
   (c) Models which underfit have a high variance.
   [(d)] Models which underfit have a low variance.

5. Which of the following are characteristics of data sampled from a Gaussian distribution?

   (a) The sample mean systematically underestimates the true mean.
   (b) The sample variance systematically overestimates the true variance.
   (c) Both the sample mean and variance are unbiased estimators of the true values.

6. Suppose your model is overfitting. Which of the following is NOT a valid way to try and reduce the overfitting?

   (a) Increase the amount of training data.

   (b) Improve the optimisation algorithm being used for error minimisation.

   (c) Decrease the model complexity.

   (d) Reduce the noise in the training data.

7. You are reviewing papers for the World's Fanciest Machine Learning Conference, and you see submissions with the following claims. Which ones would you consider accepting?

   (a) My method achieves a training error lower than all previous methods!

   (b) My method achieves a test error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise test error.)

   (c) My method achieves a test error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise cross-validaton error.)

   (d) My method achieves a cross-validation error lower than all previous methods! (Footnote: When regularisation parameter λ is chosen so as to minimise cross-validaton error.)

Section 2. Numerical questions

Please show all steps in your working fully and clearly, except where indicated otherwise.

8. The Indian Railways have been trialling 2 different machine learning methods which attempt to predict whether a train will arrive at its final destination on time or not, using a number of input features corresponding to weather conditions, train priorities, ongoing repair works etc. (for this purpose, 'on time' is defined as no more than 10 minutes after its scheduled time). The methods have been tested on a common set of 500 train runs, and the results are as follows:

	Actually on time	**Actually late**
Method 1 predicted on time	131	155
Method 1 predicted late	19	195
Method 2 predicted on time	82	72
Method 2 predicted late	68	278

Suppose we set up a simple probabilistic model for this as follows: θ is the prior probability of a train being late; p is the probability of a late prediction from Method 1 if the train is on time (also called the *False Positive Rate (FPR)*); and q is the probability of a late prediction from Method 1 if the train is in fact late (also called the *True Positive Rate (TPR)*).

(a) Write down the joint likelihood of the data for Method 1, as a function of the three model parameters θ, p, and q. Obtain maximum likelihood estimates for each of these parameters. **[4]**

$$\mathcal{L}(\theta, p, q) = \theta^{350}(1-\theta)^{150}p^{19}(1-p)^{131}q^{195}(1-q)^{155}$$

(Compute partial derivatives with respect to all parameters and set to 0 to get ML estimates):

$$\hat{\theta}_{ML} = \frac{350}{500}$$
$$\hat{p}_{ML} = \frac{19}{150}$$
$$\hat{q}_{ML} = \frac{195}{350}$$

(b) Suppose the loss matrix for this prediction task is defined as follows:

	Actually on time	**Actually late**
Predicted on time	0	1
Predicted late	K	0

Using the parameter estimates computed above, obtain the expected loss for Method 1 as a function of K. **[2]**

$$\begin{aligned}\mathbb{E}[L] &= \theta(1-q) + K(1-\theta)p \\ &= \frac{350}{500} \times \frac{155}{350} + K\frac{150}{500} \times \frac{19}{150} \\ &= \frac{155}{500} + K\frac{19}{500} \\ &= \frac{155 + 19K}{500}\end{aligned}$$

(c) Obtain the expected loss for Method 2 as well (you can compute its FPR and TPR directly, without doing the maximum likelihood derivations again); which is the preferable method? What is the critical value of K at which this preference changes? **[4]**

For Method 2:

$$\begin{aligned}\hat{p}_{ML} &= \frac{68}{150} \\ \hat{q}_{ML} &= \frac{278}{350} \\ \mathbb{E}[L] &= \frac{350}{500} \times \frac{72}{350} + K\frac{150}{500} \times \frac{68}{150} \\ &= \frac{72}{500} + K\frac{68}{500} \\ &= \frac{72 + 68K}{500}\end{aligned}$$

The preferable method is the one with the lower expected loss, which depends on the value of K. Let the critical value be K_C, then we have

$$\begin{aligned}155 + 19K_C &= 72 + 68K_C \\ 83 &= 49K_C \\ K_C &= \frac{83}{49}\end{aligned}$$

If $K > K_C$, then Method 1 is preferable; if $K < K_C$, then Method 2 is preferable.