ELL784: Introduction to Machine Learning

Re-Minor Test, Form: A

Maximum marks: 20

(Answer all questions on this question paper (check that you have all 5 pages). Use the answer script only for working; it will not be graded. Read all section-specific instructions carefully.)

Name: ______________________________

Entry Number: ________________________

Section 1. Multiple choice questions

Instructions: Each question may have any number of correct choices. Clearly mark (tick) all choices you believe to be correct (1 mark for each correct choice, −0.5 for each incorrect choice).

1. Consider the following possible choices of error function in training a neural network for classification: cross-entropy error (I), classification error (II), and sum-of-squares error (III). Which of the following are true?

   (a) (II) is problematic because it's non-differentiable, but either of (I) or (III) should give the same result.

   (b) Any of the three could be easily used for backpropagation, but (I) is preferred because it corresponds to maximising the likelihood of the data.

   (c) (II) is problematic because it's non-differentiable; (III) is preferred to (I) because the latter corresponds to an inappropriate noise model.

   (d) (II) is problematic because it's non-differentiable; (I) is preferred to (III) because the former corresponds to maximising the likelihood of the data.

   (e) Any of the three could be easily used for backpropagation, but (III) is preferred because it corresponds to maximising the likelihood of the data.

2. You are given a labeled binary classification data set with N data points and D features. Suppose that $N < D$. Which of the following kinds of feature transformation do you think would generally make sense in such a setting, prior to or as part of training a classifier?

   (a) Using a sparse autoencoder to obtain more fine-grained features

   (b) Using the kernel trick to implicitly map to a higher-dimensional feature space

   (c) Using a standard neural network with a non-linear activation function in the hidden layer(s)

   (d) No feature transformation should be necessary, a simple linear classifier trained on the raw features is likely to work

3. Suppose your model is demonstrating high bias across different training sets. Which of the following are valid ways to try and reduce the bias?

   (a) Decrease the amount of training data in each training set.

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

   (c) Increase the model complexity.

   (d) Increase the noise in the training data.

4. Which of the following might be valid reasons for preferring a neural network over an SVM?

   (a) An neural net effectively applies a non-linear transformation on the input space; an SVM cannot.

   (b) The model size (number of parameters) for a neural net is fixed in advance, whereas for an SVM it depends on the number of support vectors and hence on the training data.

   (c) A neural net should not get stuck in local minima, unlike an SVM.

   (d) Neural nets extend more naturally to multi-class classification than SVMs.

5. You are training an RBF SVM with the following parameters: C (slack penalty) and σ (where σ^2 is the variance of the RBF kernel). How should you tweak the parameters when you find the model to be underfitting?

   (a) Increase C and/or reduce σ

   (b) Reduce C and/or increase σ

   (c) Reduce C and/or reduce σ

   (d) Increase C and/or increase σ

   (e) Increase C only (σ has no predictable effect on underfitting)

6. For an RBF SVM with a particular pair of randomly chosen values of the hyperparameters C and σ (where σ^2 is the variance of the RBF kernel), which of the following tests can be taken to indicate that the chosen values likely correspond to underfitting?

   (a) When I increase C the training and validation accuracies both increase.

   (b) When I decrease C the training accuracy reduces, but validation accuracy increases.

   (c) When I decrease C the training and validation accuracies are both reduced.

   (d) When I decrease σ the number of support vectors increases.

   (e) When I decrease σ the training accuracy reduces, but validation accuracy increases.

   (f) When I increase σ the training accuracy increases, but validation accuracy reduces.

Section 2. Numerical/Short-answer questions

Instructions: Please write *only the final answers* on this question paper, in the space provided for each item. The provided answer script should be used for all working, but will not be graded. However, in case of any doubt regarding your answers, we may refer to the answer script to check your working. So please try to write out your working as clearly as possible.

7. One Sunday morning, the arrival times of six successive north-bound trains at the Hauz Khas Metro station were observed to be as follows: 7:06; 7:13; 7:47; 7:53; 8:01; 8:08. Let us denote by k the time interval (in minutes) between the arrival of two successive trains; we will assume that this follows a *Gaussian distribution*, i.e.,

$$p(k|\lambda,\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(k-\lambda)^2}{2\sigma^2}\right\}.$$

(a) The expected value of k under the Gaussian distribution is λ; i.e., in our case, λ corresponds to the expected time interval between successive arrivals. Given a sequence of independent observations of time intervals, $\{k_1, k_2, ..., k_N\}$, write down the likelihood function for this data under the Gaussian distribution. **[1]**

(b) Give the partial derivative of the log likelihood with respect to λ. **[0.5]**

(c) Give a general expression for the maximum likelihood estimate for λ. **[0.5]**

(d) Give the partial derivative of the log likelihood with respect to σ, the standard deviation parameter. **[0.5]**

(e) Give a general expression for the maximum likelihood estimate for σ. **[0.5]**

(f) Using the Metro data above, give the maximum likelihood estimate for the expected time interval at Hauz Khas station, based on the five observations. **[0.5]**

(g) Similarly, use the data to give the maximum likelihood estimate for the standard deviation of the time intervals at Hauz Khas. **[0.5]**

(h) Comment on the meaningfulness of the two estimates just obtained. **[1]**

(i) Suppose we now treat λ as a random variable and put a prior distribution on it.
We may use another Gaussian for this:

$$p(\lambda|\alpha, \beta) = \frac{1}{\sqrt{2\pi}\beta} \exp\left\{-\frac{(\lambda-\alpha)^2}{2\beta^2}\right\}.$$

Here the α and β are 'hyperparameters' to be chosen. Use this prior and the likelihood function from part (a) to obtain the posterior distribution of λ, as a function of the observed data, α, β, and σ (which we assume to be fixed). **[2]**

(j) Give the partial derivative of the log posterior with respect to λ. **[0.5]**

(k) Give a general expression for the maximum a posteriori (MAP) estimate for λ. **[0.5]**

(l) The parameter β above specifies the spread or variance of the prior distribution. Suppose we set $\beta = \sigma/3$. Suppose also that you have heard somewhere that the time interval between successive north-bound trains at Hauz Khas station is about 5 minutes. What would be a reasonable choice for α, in this case? (Look at the MAP expression derived in part (k).) **[1]**

(m) Use your choice of α, along with $\beta = \sigma/3$ and the Metro data given above, to give the MAP estimate for the waiting time at Hauz Khas. **[0.5]**

(n) Comment on the difference between this and the maximum likelihood estimate obtained in part (f). **[1.5]**