ELL409: Machine Intelligence and Learning

Minor Test I, Form: $\boxed{\text{A}}$

Maximum marks: 20

(Answer all questions on this question paper. 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, −1 for each incorrect choice).

1. Suppose your model is demonstrating overfitting on a given training set. Which of the following might possibly be helpful in reducing the overfitting? (Assume that it is practically possible to do all of these in the given setting.)

   [✓ (a)] Increasing the amount of training data.

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

   (c) Increasing the model complexity.

   [✓ (d)] Reducing the noise in the training data.

2. Four different people are assessing regularised linear regression models via cross-validation. They come to you and make the following respective claims about certain experiments they've done. Which of these claims are definitely incorrect? (Here λ refers to the regularisation parameter as usual.)

   (a) 'I increased λ and the model started underfitting the data, whilst the validation error went up'.

   [✓ (b)] 'I decreased λ and the model started overfitting the data, whilst the training error went up'.

   (c) 'I decreased λ and the model started overfitting the data, whilst the validation error went up'.

   [✓ (d)] 'I increased λ and the model started underfitting the data, whilst the training error went down'.

3. When doing MAP estimation of the parameters of a linear regression model (assuming that the optimisation can be done exactly), increasing the value of the prior precision α

   [✓ (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.

4. When comparing multiple regularised machine learning models for a given task, which of the following are NOT a suitable approach to select the best one (in terms of its ability to generalise to unseen data)? (Here λ refers to the regularisation parameter as usual.)

   [✓ (a)] Pick the one with lowest error on a separate test set, with λ having been chosen so as to minimise training error.

   [✓ (b)] Pick the one with lowest error on a separate test set, with λ having been chosen so as to minimise error on this test set.

   (c) Pick the one with lowest error on a separate test set, with λ having been chosen so as to minimise cross-validation error on the training set.

   [✓ (d)] Pick the one with lowest cross-validation error on the training set, with λ having been chosen so as to minimise cross-validation error on the training set.

5. Which of the following are meaningful only if one takes a Bayesian view of probability?

   (a) Probabilistic curve-fitting with a Gaussian noise model.

   [✓ (b)] Probabilistic curve-fitting with the model parameters treated as random variables.

   (c) Probabilistic curve-fitting to obtain a predictive distribution for test data.

   [✓ (d)] The idea of taking an expectation over the predictions of multiple models/curves.

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.

6. 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; 7:12; 7:49; 7:54; 8:01; 8:07. Let us denote by k the time interval (in minutes) between the arrival of two successive trains; we will assume here 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. First, let us consider a generic data set: 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]

$$\mathcal{L}(\lambda, \sigma) = \prod_{n=1}^{N} \left[\frac{1}{\sqrt{2\pi}\,\sigma} e^{-(k_n-\lambda)^2/2\sigma^2} \right]$$

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

$$\frac{\partial \log \mathcal{L}}{\partial \lambda} = \sum_{n=1}^{N} \frac{k_n - \lambda}{\sigma^2}$$

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

$$\hat{\lambda}_{ML} = \frac{1}{N} \sum_{n=1}^{N} k_n$$

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

$$\frac{\partial \log L}{\partial \sigma} = \frac{-N}{\sigma} + \sum_{n=1}^{N} \frac{(k_n - \lambda)^2}{\sigma^3}$$

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

$$\hat{\sigma}_{ML} = \sqrt{\frac{1}{N} \sum_{n=1}^{N} (k_n - \hat{\lambda}_{ML})^2}$$

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

$$\hat{\lambda}_{ML} = 12.2 \text{ min.}$$

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

$$\hat{\sigma}_{ML} = \sqrt{\frac{1}{5}(6.2^2 + 24.8^2 + 7.2^2 + 5.2^2 + 6.2^2)} = 12.4 \text{ min.}$$

(h) 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 (generic) data, α, β, and σ (which we assume to be fixed). **[1.5]**

$$\tilde{L}(\lambda) \triangleq p(\lambda | \underline{k}) = \frac{\left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma} e^{-(k_n - \lambda)^2 / 2\sigma^2}\right] \frac{1}{\sqrt{2\pi}\beta} e^{-\frac{(\lambda - \alpha)^2}{2\beta^2}}}{p(\underline{k})}$$

/ const, not dependent on λ

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

$$\frac{\partial \log \tilde{L}}{\partial \lambda} = \sum_{n=1}^{N} \frac{k_n - \lambda}{\sigma^2} - \frac{\lambda - \alpha}{\beta^2}$$

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

$$\hat{\lambda}_{MAP} = \frac{\beta^2 \sum_{n=1}^{N} k_n + \sigma^2 \alpha}{\beta^2 N + \sigma^2}$$

(k) 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 (j).) **[1]**

$\alpha = 5$ (prior being added as wtd. 'virtual data pt.')

(l) Now 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]**

$\hat{\lambda}_{MAP} = 7.6$ min.

(m) Is this better than the maximum likelihood estimate obtained in part (f)? Why? **[1]**

Yes, it would appear to be better, as it is closer to 4 of the 5 observed intervals. The ML estimate seems to overfit to one outlier (the 37-minute interval).