35

ELL784: Introduction to Machine Learning

Minor Test I, Form: A

Maximum marks: 20 (+2 Extra credit)

(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, −0.5 for each incorrect choice).

1. Which of the following would be incompatible with a frequentist (non-Bayesian) view of probability?

   (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 idea of assuming a probability distribution over models.

2. Four different people are doing bias-variance estimates on regularised linear regression models. They come to you and make the following 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 variance went down'.

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

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

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

3. 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 via gradient descent, initialising the parameters to those of the original model, in which cases will the learnt decision boundary remain *exactly* the same?

   [✓ (a)] When my model is a perceptron.

   (b) When my model is logistic regression.

   (c) When my model is Fisher's linear discriminant.

   (d) When my model is a linear discriminant trained via least squares.

4. Suppose your model is demonstrating high variance across different training sets. Which of the following is NOT a valid way to try and reduce the variance?

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

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

   (c) Decrease the model complexity.

   (d) Reduce the noise in the training data.

5. When comparing multiple regularised machine learning models for a given task, which of the following are reasonable ways to pick 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 training error, 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 training error.

   (c) 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.

   [✓ (d)] 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.

   (e) 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.

6. 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 noise precision β

   (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.

7. 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 underestimates the true variance.

   (c) Both the sample mean and variance are unbiased estimators of the true values.

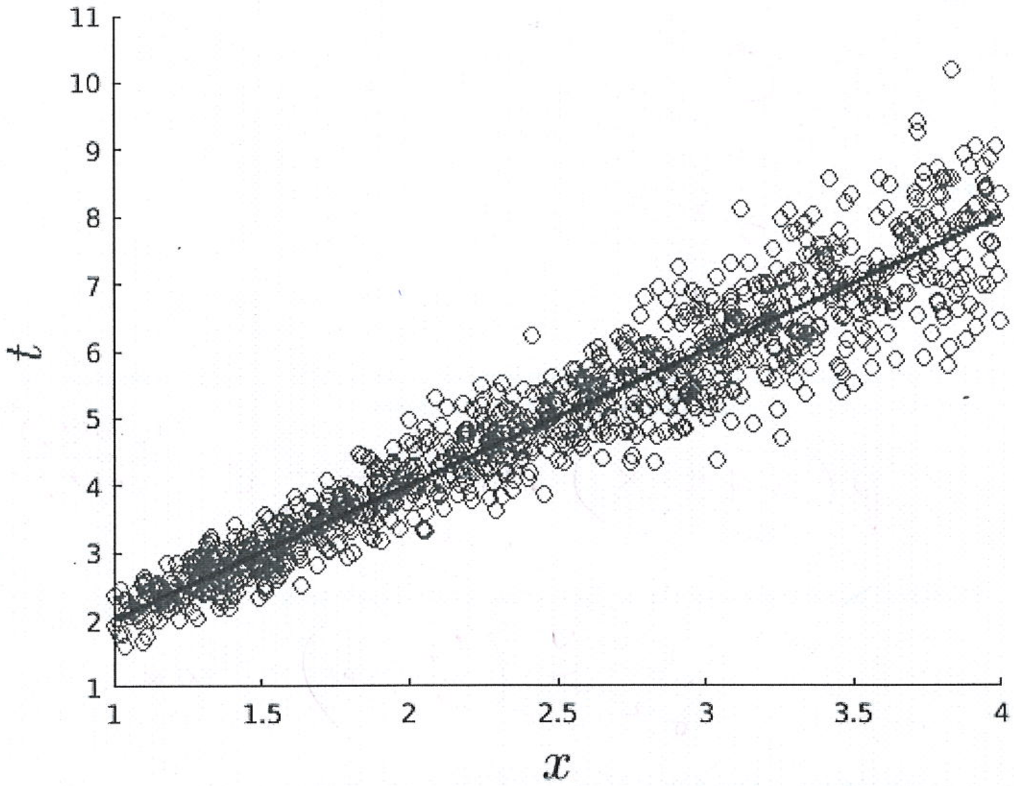
Section 2. Numerical questions

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.

8. Here we explore a discriminative regression model where the noise variance is a function of the input (variance increases as a function of input). Specifically

$$t = wx + \epsilon,$$

where the noise ϵ is normally distributed with mean 0 and standard deviation σx. The value of σ is assumed known and the input x is restricted to the interval $[1, 4]$. We can write the model more compactly as $t \sim \mathcal{N}(wx, \sigma^2 x^2)$. If we let x vary within $[1, 4]$ and sample outputs t from this model with some w, the regression plot might look like the figure on top of the next page.



Given this model, please answer the below.

(Some potentially useful relations: if $z \sim \mathcal{N}(\mu, \sigma^2)$, then $az \sim \mathcal{N}(a\mu, \sigma^2 a^2)$ for fixed a. If $z_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $z_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$ and they are independent, then $Var(z_1 + z_2) = \sigma_1^2 + \sigma_2^2$.)

(a) Suppose we have N training points and targets $\{(x_1, t_1), (x_2, t_2), ..., (x_N, t_N)\}$, where each x_n is chosen at random from $[1, 4]$ and the corresponding t_n is subsequently sampled from $t_n \sim \mathcal{N}(wx_n, \sigma^2 x_n^2)$. Give the likelihood of this data, as a function of w: **[1]**

$$\mathcal{L}(w) = \prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma x_n} e^{-\frac{(t_n - wx_n)^2}{2\sigma^2 x_n^2}}$$

(b) Give the derivative of the log likelihood with respect to w: **[0.5]**

$$\frac{1}{\sigma^2} \sum_{n=1}^{N} \left(\frac{t_n}{x_n} - w \right)$$

(c) Give the maximum likelihood estimate for w as a function of the training data: **[1]**

$$\hat{w}_{ML} = \frac{1}{N} \sum_{n=1}^{N} \frac{t_n}{x_n}$$

(d) Give the bias (i.e., the difference between the expected value of the estimator and the actual value of the parameter) and variance of the estimator for w just obtained, as a function of N and σ^2 for fixed inputs $x_1, ..., x_N$.

Bias: **[1]**

$$0$$

Variance: **[1.5]**

$$\frac{\sigma^2}{N}$$

(e) Now supposing I put a prior distribution on w: $w \sim \mathcal{N}(0, \alpha^{-1})$, for some fixed α. Give the posterior distribution for w, given the same data set as above: **[1]**

$$p(w \mid \underline{t}, \underline{x}) \propto \left[\prod_{n=1}^{N} \frac{1}{\sqrt{2\pi}\sigma x_n} e^{-\frac{(t_n - w x_n)^2}{2\sigma^2 x_n^2}} \right] \frac{\sqrt{\alpha}}{\sqrt{2\pi}} e^{-\frac{w^2 \alpha}{2}}$$

(f) Give the derivative of the log posterior with respect to w: **[0.5]**

$$\frac{1}{\sigma^2} \sum_{n=1}^{N} \left(\frac{t_n}{x_n} - w \right) - w\alpha$$

(g) Give the maximum a posteriori (MAP) estimate of w: **[1]**

$$\hat{w}_{MAP} = \frac{1}{N + \sigma^2 \alpha} \sum_{n=1}^{N} \frac{t_n}{x_n}$$

(h) Give the bias and variance of this MAP estimator.

Bias: **[1]**

$$\left(\frac{N}{N + \sigma^2 \alpha} - 1 \right) w$$

Variance: **[1.5]**

$$\frac{N\sigma^2}{(N + \sigma^2 \alpha)^2}$$

(i) (*Extra credit*) What is the role of α in controlling the bias-variance tradeoff? **[1]**

$\alpha \uparrow \Rightarrow$ Bias ↑ Variance ↓

(j) (*Extra credit*) Give the distribution of t/x for a fixed (constant) x: **[1]**

$$p\left(\frac{t}{x} \mid x\right) = \mathcal{N}\left(\frac{t}{x} \mid w, \sigma^2\right)$$