Supervised Learning Coursework 2, GI01/M055, 2017/2016

Answer all questions.

Marks for each part of each question are indicated in square brackets

1. This question concerns Bayesian inference and Gaussian Process regression.

   a. What are the two sources of randomness in the Bayesian analysis of learning?

   [4 marks]

   b. Given training data $\mathbf{X}$, how is the posterior distribution defined in terms of the two sources identified in a).

   [3 marks]

   c. In the case of Gaussian Process (GP) regression describe the two sources of randomness identified in a).

   [3 + 2 marks]

   d. For Gaussian Process regression what type of distribution is the posterior?

   [3 marks]

   e. Given your answer to d), describe the form of the marginal distribution on a test example. Give expressions for the mean and variance in terms of the Covariance matrix $\mathbf{C}$ on the training data and the vector $\mathbf{c}$ of Covariance evaluations between the training data and test example x. Explain any other symbols in the expressions.

   [8 marks]

   f. Explain why full Bayesian inference and maximum a posteriori (MAP) inference give the same prediction in the case of GP regression, illustrating your answer with an example where this equivalence would fail.

   [5 marks]

[Total 28 marks]

2. This question concerns Bayesian inference.

a. Consider the Beta distribution over the interval $[0, 1]$ parametrised by integers α and β given by

$$P_{\alpha,\beta}(p) \propto p^{\alpha-1}(1-p)^{\beta-1},$$

where the normalising constant is the Beta function

$$B(\alpha, \beta) = \int_0^1 p^{\alpha-1}(1-p)^{\beta-1}dp.$$

The distribution has mean $\alpha/(\alpha+\beta)$ and mode (maximal value) $(\alpha-1)/(\alpha+\beta-2)$ provided both α and β are greater than 1.

What is the form of this distribution when $\alpha = \beta = 1$? Verify the expression for the mean in this case.

[3 marks]

b. Consider a finite set $X = \{1, \ldots, n\}$ and hypotheses

$$H = \{h : i \in X \mapsto [0, 1]\} = [0, 1]^X$$

with a prior distribution for h given by

$$P(h) = \prod_{i=1}^{n} P_{1,1}(h(i))$$

and noise model $P(y = 1|h, (i, y)) = h(i)$ and $P(y = 0|h, (i, y)) = 1 - h(i)$, with $P(y \notin \{0, 1\}) = 0$. Compute the posterior distribution after processing the example $(j, 1)$ twice and the example $(j, 0)$ once.

[10 marks]

c. Suppose a training set contains 10 instances that involve input j, 3 of which have output 1. Compute the MAP classification $h_{\text{map}}(j)$ and the posterior Bayesian average

$$h_{\text{avg}}(j) = \int_0^1 h(j)dP_{\text{post}}(h)$$

[5 marks]

[Total: 18 marks]

3. This question concerns overfitting and regularisation.

   a. What is meant by overfitting? Illustrate your answer with a graph showing the typical training and test performance observed as the complexity of a function class of learners increases.

   [5 marks]

   b. Given a set H of functions that can be output by a learner and training data $\mathbf{X}$, what is meant by Empirical Risk Minimisation (ERM)?

   [3 marks]

   c. Describe how Structural Risk Minimisation (SRM) addresses the problem of overfitting when applying ERM.

   [6 marks]

   d. Explain how regularisation addresses the problem of overfitting in a different way to that described in c). Describe the role of the regularisation parameter.

   [6 marks]

   e. The Support Vector Machine (SVM) optimisation is given by

$$\begin{aligned} \min_{\mathbf{w},b,\gamma,\xi} \quad & \|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \\ \text{subject to} \quad & y_i\left\langle \mathbf{w},\mathbf{x}_i\right\rangle \geq 1-\xi_i, \xi_i \geq 0, \\ & i = 1,\ldots,m. \end{aligned}$$

   Which part of the optimisation corresponds to the regulariser? What is the loss function incorporated into the optimisation?

   [5 marks]

   f. What component of the optimisation in e) is changed to create Linear Programming boosting? What assumptions about the classification problem would make such a change advisable?

   [5 marks]

   [Total 30 marks]

4. a. Assume that labelled training and test data $(\mathbf{x}, y) \in \mathbb{R}^d \times \{-1, 1\}$ are generated identically and independently (i.i.d.) according to a distribution $\mathcal{D}$. Define the generalisation error $\mathrm{err}(c)$ of a classifier

$$c : \mathbb{R}^d \longrightarrow \{-1, 1\}.$$

[4 marks]

b. Consider a fixed learning algorithm $\mathcal{A}(S)$ that returns a classification function $c = \mathcal{A}(S)$ from a training set S. As random training sets are drawn of a fixed size m, the generalisation error of the function $\mathcal{A}(S)$ is a random variable. Consider its distribution and suppose that it has 0.95 percentile 0.15 and mean 0.08. Use this distribution to illustrate the difference between expected and high confidence generalisation bounds when the confidence parameter $\delta = 0.05$.

[6 marks]

c. Suppose that we have a sequence of hypothesis spaces

$$\mathcal{H}_1 \subseteq \mathcal{H}_2 \subseteq \cdots \subseteq \mathcal{H}_K$$

of classifiers together with bounds for each $\mathcal{H}_i$, $i = 1, \ldots, K$, $c \in \mathcal{H}_i$

$$\mathrm{err}(c) \leq \hat{\mathrm{err}}(c) + \sqrt{\frac{\mathcal{C}(\mathcal{H}_i) + \ln \frac{1}{\delta}}{m}}$$

where $\hat{err}(c)$ is the empirical (training set) error and $\mathcal{C}(\mathcal{H}_i)$ some measure of complexity of the space $\mathcal{H}_i$.

Explain what is meant by structural risk minimisation and how it resolves the trade-off between complexity and empirical error.

[7 marks]

d. Using the notation from part 4.c., give the generalisation bound that applies to the function selected by structural risk minimisation indicating what technique is used to obtain the bound.

[7 marks]

[Total: 24 marks]

END OF COURSEWORK