The University Of Sheffield.

Data Provided: None

DEPARTMENT OF COMPUTER SCIENCE **Autumn Semester 2013–2014**

MACHINE LEARNING AND ADAPTIVE INTELLIGENCE **2 hours**

Answer THREE of the four questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

Registration number from U-Card (9 digits) — to be completed by student

1. This question concerns general concepts in machine learning.

   a) Overfitting is a common problem in machine learning, in both regression and classification.

      (i) Explain the problem, when it might occur, and how you can measure the extent of overfitting. [10%]

      (ii) With reference to the regression or classification models you have studied, describe two different regularisation techniques for addressing the overfitting problem. Provide a sentence or two on each, explaining how they limit overfitting. [15%]

   b) Model training for probabilistic models often involves taking point estimates for the model parameters, such as the *maximum a posteriori* (MAP) estimate.

      (i) Define the objective the MAP is optimising, making reference to Bayes' rule. [10%]

      (ii) *Bayesian inference* is an alternative inference technique which also makes use of a prior. Outline the Bayesian inference technique, and contrast it with the use of point estimates. [20%]

      (iii) Why might you choose to use Bayesian inference instead of a point estimate, or vice-versa? In what circumstances would Bayesian inference be preferable? [15%]

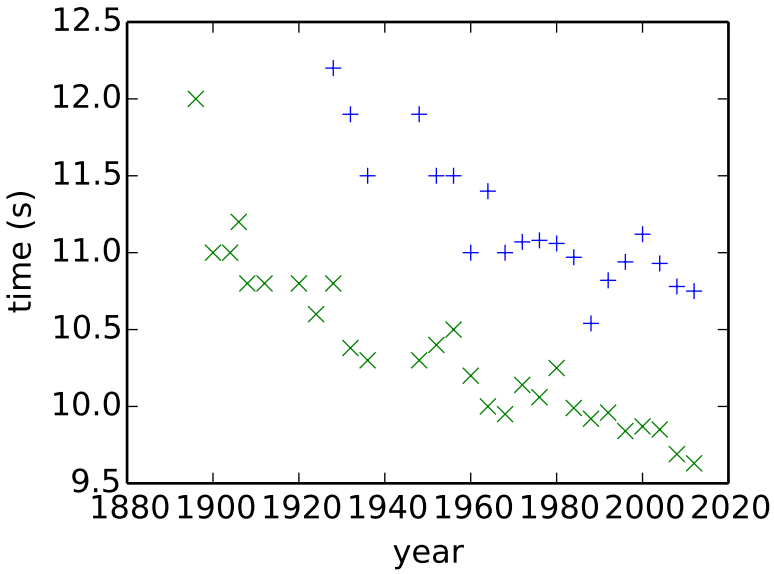
   c) Several pairs of distributions are said to be *conjugate*, such as the Binomial and Beta; Multinomial and Dirichlet; and Normal (mean) and Normal. Explain the notion of *conjugacy*, and how this might be practically important in a classification or regression scenario. [15%]

   d) Logistic regression is a probabilistic model for binary classification. It defines the probability of class 1 as

$$p(\mathcal{C}_1|\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^\top \mathbf{x})}$$

   and $p(\mathcal{C}_1|\mathbf{x}) = 1 - p(\mathcal{C}_2|\mathbf{x})$. Show how this gives rise to a linear discriminant function. You may want to start by formulating the log-odds ratio for predicting $\mathcal{C}_1$ versus $\mathcal{C}_2$. [15%]

2. This question is based on classifying the gender of Olympic 100m sprint winners. Shown below is a plot of the winning times in each year for the women's and men's events.



The women's results are shown with plus symbols (+) and the men's results with crosses (x). We would now like to develop a classifier to predict the gender (t = male or female) automatically based on a two dimensional data point, $\mathbf{x}$ = (year, winning time). Note that we are seeking to model this as a classification dataset, **not** regression as used for your class work.

a) We decide to model this data using a linear binary classifier.

(i) Draw a rough diagram to illustrate a decision boundary that you might hope to learn from this data with a linear classifier. Label the regions corresponding to the two classes, $\mathcal{C}_1$ = female, $\mathcal{C}_2$ = male. [10%]

(ii) Is a linear classifier an appropriate choice for this data? You should consider both *interpolation* and *extrapolation* settings, and explain these terms in your answer. [15%]

b) Basis functions can be used to develop non-linear models. Give an example of a basis function that would be appropriate for this dataset, and explain why. Based on your choice, state the basis vector $\phi(\mathbf{x})$ used to represent a data-point $\mathbf{x}$. [15%]

c) Various techniques can be used for *validation*, such as using a fixed held-out validation set, or cross-validation.

(i) Explain the purpose of validation. Define fixed held-out validation and cross-validation. [10%]

(ii) Imagine we were to perform k-fold cross-validation on this data. We do so by assigning the data points from the 100m dataset based on the year of each race, such that each 20 year period forms a fold. Explain why this form of evaluation might not give a reliable estimate of the generalisation error, and how you might fix this. [15%]

d) Kernel methods extend the basis functions technique to allow for richer and more flexible data representations.

(i) Explain what is meant by a *kernel function*, and how they relate to basis functions. [10%]

(ii) Using a linear perceptron with basis function ϕ, the perceptron update takes the form

$$\mathbf{w} \leftarrow \mathbf{w} + \eta t_i \phi(\mathbf{x}_i)$$

after an error is made on the i^{th} training instance. Show how the weights can be represented as a linear combination of the training samples,

$$\mathbf{w} = \sum_i \alpha_i t_i \phi(\mathbf{x}_i)$$

and show the dual form of the update rule, in terms of α. [15%]

(iii) Using the above reparameterisation, derive the kernel perceptron. This requires you to prove that the perceptron discriminant function $y(\mathbf{x}) = \mathbf{w}^T \phi(\mathbf{x})$ can be expressed such that the basis functions occur solely as inner products $\phi(\mathbf{x}_i)^T \phi(\mathbf{x}_j)$. Justify why this is important. [10%]

3. This question concerns regression and maximum likelihood fits of regression models with basis functions.

   a) What role do the basis functions play in a regression model? [10%]

   b) The polynomial basis with degree d computed for a one dimensional input has the form

   $$\boldsymbol{\phi}(x_i) = \begin{bmatrix} 1 & x_i & x_i^2 & x_i^3 & \dots & x_i^d \end{bmatrix}^\top .$$

   Give a disadvantage of the polynomial basis. Suggest a potential fix for this disadvantage and propose an alternative basis. [20%]

   c) The likelihood of a single data point in a regression model is given by,

   $$p(y_i|\mathbf{w}, \mathbf{x}_i, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(y_i - \boldsymbol{\phi}(\mathbf{x}_i)^\top \mathbf{w})^2}{2\sigma^2}\right).$$

   Assuming that each data point is independent and identically distributed, derive a suitable *error function* that should be minimized to recover $\mathbf{w}$ and σ^2. Explain your reasoning at each step. [25%]

   d) Now show that this error function is minimized with respect to the vector $\mathbf{w}$, at the following point,

   $$\mathbf{w}^* = \left[\boldsymbol{\Phi}^\top \boldsymbol{\Phi}\right]^{-1} \boldsymbol{\Phi}^\top \mathbf{y}$$

   where $\boldsymbol{\Phi}$ is a *design matrix* containing all the basis vectors, and $\mathbf{y}$ is a vector of regression targets. [30%]

   e) What problem will arise as the number of basis functions we use increases to become larger than the number of the data points we are given? How can we perform a regression in this case? [15%]

4. This question deals with Bayesian approaches to machine learning problems.

   a) Machine learning deals with data. What do we need to combine with the data in order to make predictions? [10%]

   b) Bayes' Rule relates four terms: the *likelihood*, the *prior*, the *posterior* and the *marginal likelihood* or *evidence*.

      (i) Describe the role of each of these terms when modelling data. [20%]

      (ii) Write down the relationship between these four terms as given by Bayes' rule. [10%]

   c) In a regression problem we are given a vector of real valued targets, $\mathbf{y}$, consisting of N observations $y_1 \dots y_N$ which are associated with multidimensional inputs $\mathbf{x}_1 \dots \mathbf{x}_N$. We assume a linear relationship between y_i and $\mathbf{x}_i$ where the data is corrupted by independent Gaussian noise giving a likelihood of the form

$$p(\mathbf{y}|\mathbf{x}, \mathbf{w}, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{N}(y_i - \mathbf{w}^\top \mathbf{x}_i)^2\right).$$

   Consider the following Gaussian prior density for the k dimensional vector of parameters, $\mathbf{w}$,

$$p(\mathbf{w}) = \frac{1}{(2\pi\alpha)^{\frac{k}{2}}} \exp\left(-\frac{1}{2\alpha}\mathbf{w}^\top \mathbf{w}\right)$$

      (i) This prior and likelihood can be combined to form the posterior. Explain why the resulting posterior will be Gaussian distributed. [10%]

      (ii) Show that the covariance of the posterior density for $\mathbf{w}$ is given by

$$\mathbf{C}_w = \left[\frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{X} + \frac{1}{\alpha}\mathbf{I}\right]^{-1},$$

      where $\mathbf{X}$ is a *design matrix* of the input data. [35%]

      (iii) Show that the mean of the posterior density for $\mathbf{w}$ is given by

$$\boldsymbol{\mu}_w = \mathbf{C}_w \frac{1}{\sigma^2}\mathbf{X}^\top \mathbf{y}.$$

      [15%]

END OF QUESTION PAPER