UNIVERSITY OF SOUTHAMPTON COMP3008W1

SEMESTER 2 EXAMINATION 2008/2009

MACHINE LEARNING

Duration: 120 mins

Answer THREE *questions out of* FOUR

This examination is worth 70%. The coursework was worth 30%.

University approved calculators MAY be used.

 Number of pages 12

Question 1

(a) Explain what you understand by *supervised learning*, *unsupervised learning* and *novelty detection*. Briefly describe one potential application of each of these.

Supervised learning—problem definition includes targets (real numbers/vectors) for an interpolation/regression problem; class labels for a classification problem; several example applications—voice recognition, medical diagnostics, etc.

Unsupervised learning is essentially cluster analysis; examples include the analysis of marketing data, genome experimental data, etc.

Novelty detection is closely related to unsupervised learning, involves the estimation of probability densities and the use of these to detect abnormal behavior—bank transaction data, health monitoring, complex engine monitoring, etc.

(6 marks)

(b) A linear regression model given by

$$f = \boldsymbol{w}^t \boldsymbol{x}$$

is to be estimated from N items of data, $\{\boldsymbol{x}_n, f_n\}_{n=1}^N$, where the usual offset term w_0 is ignored. Show how minimising the average squared error leads to a closed form solution for the parameter vector $\boldsymbol{w}$. Derive your answer in the form of a pseudo inverse of a matrix.

Linear regression—derivation bookwork as discussed in class; the answer should have a correct expression for squared error summed over all data; a data matrix Y ("design matrix") where the input data vectors are stacked as rows, writing of the error in matrix algebra form, differentiating the error in matrix form, leading to the pseudo inverse.

(7 marks)

(c) How would you modify your solution above to obtain an *online* algorithm for estimating $\boldsymbol{w}$.

Gradient derived in previous section used to update—gradient descent solution; answer should note that the gradient is obtained by summing over all the data, if, instead, we use a noisy estimate (corresponding to a single data), we get an on line algorithm.

(4 marks)

(d) Comment on the convergence properties of the online algorithm.

Convergence properties—answer should have two properties: (a) that the choice of step size strikes a balance between speed of convergence and oscillation; (b) with the on-line algorithm, convergence will be noisy, not smooth.

(3 marks)

(e) Show how, by choosing a suitable substitute for the squared error in the regression problem, the *perceptron* algorithm for classification may be derived.

For perceptron derivation, the answer should accurately define the cost function as sum over all misclassified examples and the negative of the scalar product between data and parameters. No marks if either of these is wrong.

(7 marks)

(f) Comment on the convergence properties of the perceptron algorithm.

Perceptron will converge to a solution if the data is linearly separable, and will oscillate about an approximate least squares solution when there is no linear separability.

(3 marks)

(g) What is its main limitation in solving pattern classification problems?

Perceptron will not perform well if the data are not linearly separable.

(3 marks)

Question 2

(a) Bayes rule for conditional probabilities, as commonly used in statistical pattern classification problems, is

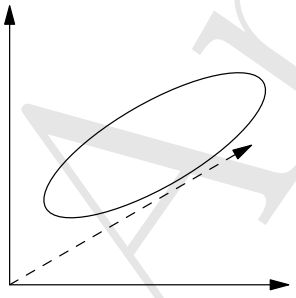
$$P\left[A|\boldsymbol{x}\right] = \frac{P\left[A\right]\, p\left(\boldsymbol{x}|A\right)}{p\left(\boldsymbol{x}\right)}$$

With reference to a practical problem of your choice, explain the different terms in the above expression.

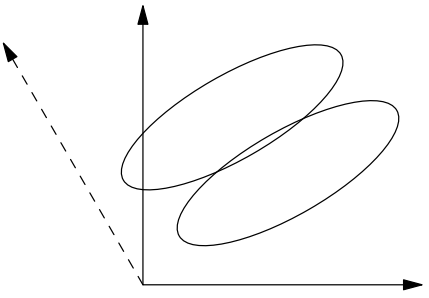
Answer should explain prior probability, likelihood model for the features conditioned on each class and that the denominator term is a normalizing constant obtained by summing over all the classes. If the example is a voice classification problem, the prior probabilities may come from a language model (frequency of sounds in the language), if it is a medical diagnostics problem prior knowledge of prevalence of a disease—look in answer for the comment that the prior is obtained with no reference to the measurement. Classification is done by comparing the posteriors of different classes. (4 marks for straight description of the terms and 6 for associating clearly the terms with a real example—students were encouraged to think along these lines in class.)

(10 marks)

(b) Explain, using two dimensional sketches, the difference between *Principal Component Analysis* and *Fisher Linear Discriminant Analysis*. Briefly describe a potential application of each of these.



PCA preserves data variance



Discriminant analysis looks for a direction to project so that the projected data is maximally separated

PCA is applicable in a data reduction (low bit rate coding, for example) situation; it is also used as a feature reduction method in some applications

LDA (with Fisher criterion) is meant for classification problems; any example of a discrimination problem gets full marks.

(10 marks)

(c) A two dimensional two-class pattern classification problem is defined by the following data:

- Class A:
  $\begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 1 \\ 2.2 \end{bmatrix}, \begin{bmatrix} 1 \\ 1.8 \end{bmatrix}, \begin{bmatrix} 0.8 \\ 2 \end{bmatrix}, \& \begin{bmatrix} 1.2 \\ 2 \end{bmatrix}$
- Class B:
  $\begin{bmatrix} 3 \\ 1 \end{bmatrix}, \begin{bmatrix} 2.7 \\ 1 \end{bmatrix}, \begin{bmatrix} 3.3 \\ 1 \end{bmatrix}, \begin{bmatrix} 3 \\ 0.7 \end{bmatrix}, \& \begin{bmatrix} 3 \\ 1.3 \end{bmatrix}$

Assuming the data are distributed according to Gaussian probability densities, derive an expression for the Bayes optimum class boundary.

Draw a neat sketch of the data and the class boundary.

Illustrate on your sketch the class boundary of a *nearest-neighbour* decision rule.

For this problem, the data have been carefully chosen to give isotropic distributions, but the variances are different (so are the class means). Substituting the expression for Gaussian into Bayes formula (note: denominator cancels, students may assume equal prior without loss of marks, but carrying through the prior doesn't change the results); class boundary will be quadratic (because the covariances being different, the quadratic terms will not cancel). Full marks for the correct derivation (even if there are algebraic errors), as long as the student spots that the shape of the class boundary will be a parabola in 2D. [10 marks]

Nearest neighbour class boundary will be piecewise linear, segments that bisect adjacent data points across the two classes, forming an approximation to the parabola. [3 marks]

(13 marks)

Question 3

(a) What is the difference between learning error and generalisation error?

(Simple opener testing basic understanding.)

The learning error is the error on the training data, while the generalisation error is the error obtained on unseen data. For many machine learning algorithms the generalisation error is significantly higher than the learning error.

(2 marks)

(b) Explain how you can accurately estimate the generalisation error given a limited data set.

(Test knowledge of practical data handling issues.)

Generalisation can be measured, by reserving some of the training data for measuring generalisation performance. This data must not be used in any way for training. When data is scarce, then cross-validation can be used, where a small proportion of the data is kept out for testing. To get better accuracy, this is done several times over—so called, *k-way cross-validation*—each time a different set of data is taken out of the training data and used for testing. The learning machine has to be retrained from scratch each time. The extreme case is *leave one out cross-validation* where a single data-point is left out of the training data and used for measuring generalisation performance. This is repeated for every member of the data set.

(6 marks)

(c) Why are regularisation terms added to the error function?

(Test understanding of generalisation theory.)

Complicated learning machines are often necessary to learn difficult data-sets. However, given a limited amount of data, they are liable to over-fit the training data. A regularisation term biases the machine to find "simpler" solutions which are more likely to represent the underlying rule being learned and so will have better generalisation performance.

(6 marks)

(d) A weight decay term has the form $\lambda \sum_i w_i^2$. Show how adding such a term modifies the update rule for the weights and hence explain why it is known as a weight decay term.

(Test ability to manipulate mathematical formula necessary for understanding machine learning.)

We typically train a learning machine by minimising an error term and a regularisation term

$$E(\boldsymbol{w}) = \sum_{k=1}^{P} \left(F(\boldsymbol{x}_k; \boldsymbol{w}) - y_k\right)^2 + \lambda \sum_i w_i^2$$

were we have a (multi-)set of data $\{(\boldsymbol{x}_k, y_k)\}_{k=1}^{P}$ where $\boldsymbol{x}_k$ are the input features and y_k are the targets. $F(\boldsymbol{x}_k; \boldsymbol{w})$ is the output of the learning machine. To minimise the weights we move in a direction which minimises the error (i.e. in the direction of the negative gradient)

$$\begin{aligned}\boldsymbol{\nabla} E(\boldsymbol{w}) &= 2 \sum_{k=1}^{P} \boldsymbol{\nabla} F(\boldsymbol{x}_k; \boldsymbol{w}) \left(F(\boldsymbol{x}_k; \boldsymbol{w}) - y_k\right) + 2\,\lambda\,\boldsymbol{w} \\ &= 2 \sum_{k=1}^{P} \boldsymbol{\nabla} F(\boldsymbol{x}_k; \boldsymbol{w}) \delta_k + 2\,\lambda\,\boldsymbol{w}\end{aligned}$$

where $\delta_k = F(\boldsymbol{x}_k; \boldsymbol{w}) - y_k$ is the error between the predicted and true result. Thus we update the weights

$$\begin{aligned}\Delta \boldsymbol{w} = \boldsymbol{w}' - \boldsymbol{w} &= -\eta\,\boldsymbol{\nabla} E(\boldsymbol{w}) \\ &= -2\,\eta \sum_{k=1}^{P} \boldsymbol{\nabla} F(\boldsymbol{x}_k; \boldsymbol{w})\,\delta_k - 2\,\eta\,\lambda\,\boldsymbol{w}\end{aligned}$$

where $\boldsymbol{w}'$ are the updated weights and η is a learning weight. Solving for $\boldsymbol{w}'$ we find

$$\boldsymbol{w}' = (1 - 2\,\eta\,\lambda)\,\boldsymbol{w} - 2\,\eta \sum_{k=1}^{P} \boldsymbol{\nabla} F(\boldsymbol{x}_k; \boldsymbol{w})\,\delta_k.$$

Thus, at each learning step the previously learned weights are reduced by a factor $1 - 2\,\eta\,\lambda$, hence the term weight decay.

(10 marks)

(e) Explain the drawback of using a weight decay term and explain how an SVM avoids the need for such a term.

(Test deeper understanding of the challenge of obtaining good generalisation performance.)

When we add a regularisation term we modify the error function. This would be optimal if the regularisation term exactly encoded our prior beliefs. This is rarely the case (more often the regularisation term is chosen to be convenient to implement). We have to determine the regularisation parameter λ which is either

an *ad hoc* choice or has to be determined using data which could otherwise be used for training.

In SVMs we minimise the learning error for a very powerful learning machine, but we choose the learning machine, which in a sense is the simplest (or most robust) machine that is consistent with the data.

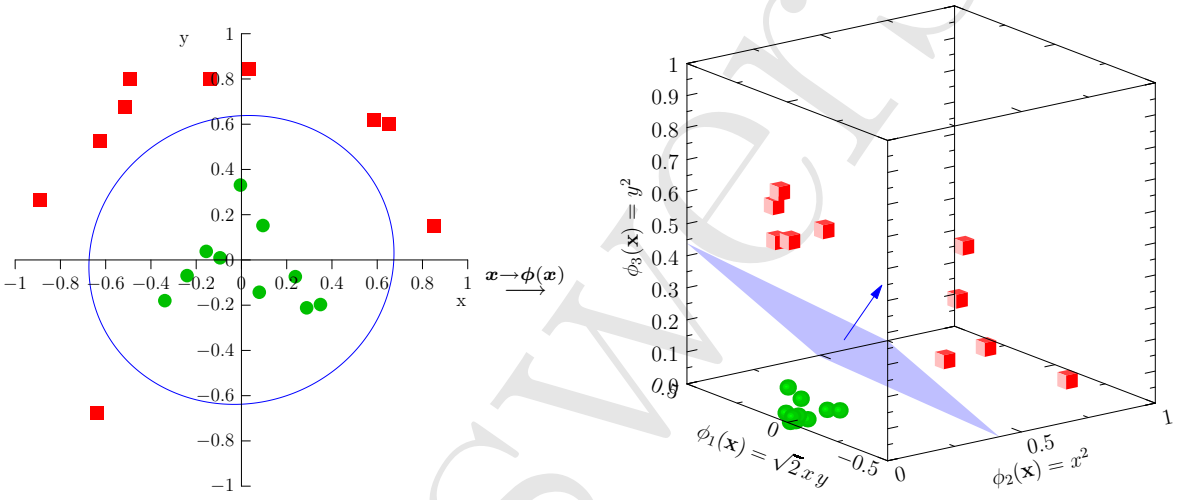
(9 marks)

Question 4

(a) Explain how an SVM can be made to separate linearly separable data. Provide a schematic sketch of how this is done.

(Tests understanding of the kernel trick in the context of SVMs.)

To separate linearly separable data we project the data into a higher dimensional space using the kernel trick. In this space the data is now linearly separable.



(5 marks)

(b) Mercer's theorem states that

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \lambda_i \, \psi_i(\boldsymbol{x}) \, \psi_i(\boldsymbol{y}).$$

Show that if the eigenvalues λ_i are non-negative (i.e. $\lambda_i \geq 0$) then for any real function $f(x)$

$$\int f(\boldsymbol{x}) \, K(\boldsymbol{x}, \boldsymbol{y}) \, f(\boldsymbol{y}) \, \mathrm{d}\,\boldsymbol{x} \, \mathrm{d}\,\boldsymbol{y} \geq 0.$$

(The rest of this question tests knowledge about positive seim-definiteness. It tests student's ability to put together and understand a set of proofs.)

TURN OVER

$$\int f(\boldsymbol{x})\, K(\boldsymbol{x},\boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\,\boldsymbol{x}\, \mathrm{d}\,\boldsymbol{y} = \int f(\boldsymbol{x}) \sum_i \phi_i(\boldsymbol{x})\, \phi_i(\boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\,\boldsymbol{x}\, \mathrm{d}\,\boldsymbol{y}$$
$$= \sum_i \left(\int f(\boldsymbol{x})\, \phi_i(\boldsymbol{x})\, \mathrm{d}\,\boldsymbol{x} \right)^2 \geq 0$$

(5 marks)

(c) Explain why positive semi-definiteness is an important property of kernels used in SVMs.

(Test key understanding).

This ensures that the fixed point maximises the margin (i.e. that the maximum margin solution does not occur when some parameters become infinite.)

(5 marks)

(d) Show that if $K_1(\boldsymbol{x},\boldsymbol{y})$ and $K_2(\boldsymbol{x},\boldsymbol{y})$ are positive semi-definite kernels then so is $K_3(\boldsymbol{x},\boldsymbol{y}) = K_1(\boldsymbol{x},\boldsymbol{y}) + K_2(\boldsymbol{x},\boldsymbol{y})$.

(A simple proof)

We require the quadratic form Q to be non-negative

$$Q = \int f(\boldsymbol{x})\, K_3(\boldsymbol{x},\boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\,\boldsymbol{x}\, \mathrm{d}\,\boldsymbol{y}$$
$$= \int f(\boldsymbol{x})\, \left(K_1(\boldsymbol{x},\boldsymbol{y}) + K_2(\boldsymbol{x},\boldsymbol{y}) \right) f(\boldsymbol{y})\, \mathrm{d}\,\boldsymbol{x}\, \mathrm{d}\,\boldsymbol{y}$$
$$= \int f(\boldsymbol{x})\, K_1(\boldsymbol{x},\boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\,\boldsymbol{x}\, \mathrm{d}\,\boldsymbol{y} + \int f(\boldsymbol{x})\, K_2(\boldsymbol{x},\boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\,\boldsymbol{x}\, \mathrm{d}\,\boldsymbol{y} \geq 0$$

which follows form Mercer's theorem.

(5 marks)

(e) Using the fact that any positive semi-definite kernel can be decomposed as

$$K(\boldsymbol{x},\boldsymbol{y}) = \sum_i \phi_i(\boldsymbol{x})\, \phi_i(\boldsymbol{y})$$

show that the product of two kernel functions is positive semi-definite.

(This is a more complicated proof.)

- **If $K_1(\boldsymbol{x}, \boldsymbol{y})$ and $K_2(\boldsymbol{x}, \boldsymbol{y})$ are valid kernels then so is $K_3(\boldsymbol{x}, \boldsymbol{y}) = K_1(\boldsymbol{x}, \boldsymbol{y})\, K_2(\boldsymbol{x}, \boldsymbol{y})$**
- **Writing $K_1(\boldsymbol{x}, \boldsymbol{y}) = \sum_i \phi_i^1(\boldsymbol{x})\, \phi_i^1(\boldsymbol{y})$ and $K_2(\boldsymbol{x}, \boldsymbol{y}) = \sum_j \phi_j^2(\boldsymbol{x})\, \phi_j^2(\boldsymbol{y})$ then**

$$K_3(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i,j} \phi_{ij}^3(\boldsymbol{x})\, \phi_{ij}^3(\boldsymbol{y})$$

where $\phi_{ij}^3(\boldsymbol{x}) = \phi_i^1(\boldsymbol{x})\, \phi_j^2(\boldsymbol{x})$. In other words,

$$\int f(\boldsymbol{x})\, K_3(\boldsymbol{x}, \boldsymbol{y})\, f(\boldsymbol{y})\, \mathrm{d}\,\boldsymbol{x}\, \mathrm{d}\,\boldsymbol{y} = \sum_{i,j} \left(\int \phi_i^1(\boldsymbol{x})\, \phi_j^2(\boldsymbol{x})\, f(x)\, \mathrm{d}\, x \right)^2 \geq 0.$$

(5 marks)

(f) Using the previous results show that the exponential of a positive semi-definite kernel function is also positive semi-definite.

(Test ability to put together the pieces.)

- **If $K(\boldsymbol{x}, \boldsymbol{y})$ is a valid kernel so is $K^n(\boldsymbol{x}, \boldsymbol{y})$ (by induction)**
- **and $\exp(K(\boldsymbol{x}, \boldsymbol{y}))$ is also a valid kernel since**

$$\mathrm{e}^{K(\boldsymbol{x}, \boldsymbol{y})} = \sum_i \frac{1}{i!} K^i(\boldsymbol{x}, \boldsymbol{y}) = 1 + K(\boldsymbol{x}, \boldsymbol{y}) + \frac{1}{2} K^2(\boldsymbol{x}, \boldsymbol{y}) + \cdots$$

but each term in the sum is a kernel

(4 marks)

(g) Prove that the Gaussian kernel is positive semi-definite.

(and finally. . .)

- **Now $\boldsymbol{x}^\mathsf{T}\boldsymbol{y}$ is a valid kernel because it is of the form $\sum_i \phi_i(\boldsymbol{x})\, \phi_i(\boldsymbol{y})$ where $\phi_i(\boldsymbol{x}) = x_i$**
- **Thus $\exp(\boldsymbol{x}^\mathsf{T}\boldsymbol{y})$ is a valid kernel**
- **Since $\exp(\boldsymbol{x}^\mathsf{T}\boldsymbol{x}/2)$ and $\exp(\boldsymbol{y}^\mathsf{T}\boldsymbol{y}/2)$ are positive**

$$\begin{aligned} \frac{\mathrm{e}^{\boldsymbol{x}^\mathsf{T}\boldsymbol{y}}}{\mathrm{e}^{\boldsymbol{x}^\mathsf{T}\boldsymbol{x}/2}\, \mathrm{e}^{\boldsymbol{y}^\mathsf{T}\boldsymbol{y}/2}} &= \mathrm{e}^{-\boldsymbol{x}^\mathsf{T}\boldsymbol{x}/2 + \boldsymbol{x}^\mathsf{T}\boldsymbol{y} - \boldsymbol{y}^\mathsf{T}\boldsymbol{y}/2} \\ &= \mathrm{e}^{-\|\boldsymbol{x}-\boldsymbol{y}\|^2/2} \end{aligned}$$

is a valid kernel

(4 marks)