UNIVERSITY OF SOUTHAMPTON COMP3008W1

SEMESTER 2 EXAMINATION 2012/2013

MACHINE LEARNING

Duration: 120 mins

You must enter your Student ID and your ISS login ID (as a cross-check) on this page. You must not write your name anywhere on the paper.

Student ID:	
ISS ID:	

Question	Marks
1	
2	
3	
4	
Total	

Answer all parts of the question in section A (20 marks) and TWO questions from section B (25 marks each)

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

University approved calculators MAY be used.

Each answer must be completely contained within the box under the corresponding question. No credit will be given for answers presented elsewhere.

You are advised to write using a soft pencil so that you may readily correct mistakes with an eraser.

You may use a blue book for scratch—it will be discarded without being looked at.

Section A

Question A 1

(a) Explain what the **bias** and **variance** terms in the expected generalisation error is and explain the bias variance dilemma. *(6 marks)*

(i) The bias is the generalisation error of the average machine

(ii) The variance measures the expected variation from the average machine due to the fluctuations caused by finite training set

(iii) The bias variance dilemma is that a simple machine is likely to have a high bias but low variance while a complex machine will have a low bias but high variance

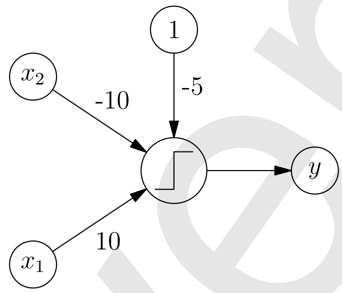
---

(b) Show that the kernel function $K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{\phi}^{\mathsf{T}}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{y})$, where $\boldsymbol{\phi}(\boldsymbol{x})$ is a vector equal to $(x_1^2, x_2^2, x_3^2, \sqrt{2}\, x_1 x_2, \sqrt{2}\, x_1 x_3, \sqrt{2}\, x_2 x_3)$, can be written as $(\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y})^2$ is $\boldsymbol{x}$ and $\boldsymbol{y}$ are vectors of length 3. *(4 marks)* $\boxed{\overline{6}}$

$$\begin{aligned}\boldsymbol{\phi}^{\mathsf{T}}(\boldsymbol{x})\boldsymbol{\phi}(\boldsymbol{y}) &= x_1^2 y_1^2 + x_2^2 y_2^2 + x_3^2 y_3^2 + 2 x_1 x_2 y_1 y_2 + 2\, x_1 x_3 y_1 y_3 + 2 x_2 x_3 y_2 y_3 \\ &= (x_1 y_1 + x_2 y_2 + x_3 y_2)^2 = (\boldsymbol{x}^{\mathsf{T}}\boldsymbol{y})^2\end{aligned}$$

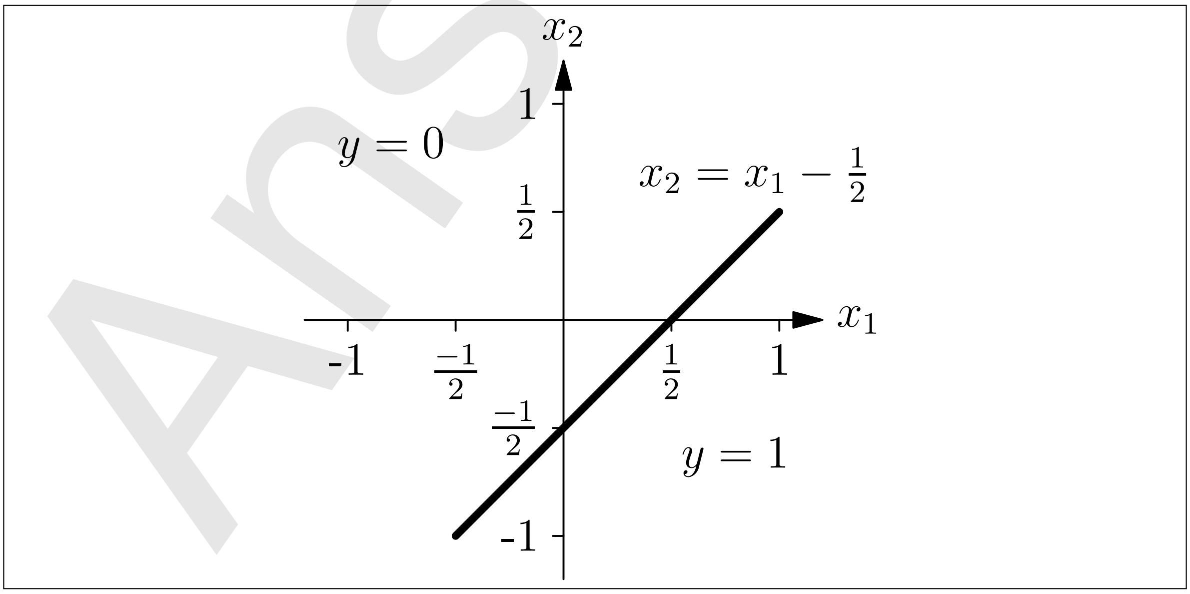
$\boxed{\overline{4}}$

(c) The figure below shows a step perceptron with an output $\Theta(V)$ which is equal to 1 if $V > 0$ and 0 otherwise. Write down the formula that describes the response of the perceptron and draw the separating surface in the input space, indicating the response y.



(5 marks)

$$y = \Theta(10x_1 - 10x_2 - 5)$$



$\overline{5}$

(d) Give (high level) pseudo code for k-means clustering. *(5 marks)*

(i) Choose k

(ii) Randomly partition the input patterns into k groups, $\mathcal{C}_i$

(iii) Do until no change

- **i. Calculate the mean of the partition $\boldsymbol{\mu}_i = \frac{1}{|\mathcal{C}_i|}\sum_{k\in\mathcal{C}_i} \boldsymbol{x}_k$**
- **ii. For each input pattern**
  - **A. For each class calculate distance $\|\boldsymbol{x}_k - \boldsymbol{\mu}_i\|$**
  - **B. Assign pattern to nearest centre**

end for

end do

End of question 1

Q1: (a) $\frac{}{6}$	(b) $\frac{}{4}$	(c) $\frac{}{5}$	(d) $\frac{}{5}$	Total $\frac{}{20}$

$\boxed{\frac{}{5}}$

Section B

Question B 2

(a) Show that the squared error of a linear perceptron with data $(\boldsymbol{x}_k, y_k)$ for $k = 1, 2, \ldots, P$ can be written as $\|\mathbf{X}^\mathsf{T}\boldsymbol{w} - \boldsymbol{y}\|^2$ where $\mathbf{X}$ is the usual matrix of input patterns and $\boldsymbol{y}$ is a vector of target values. *(5 marks)*

When $\mathbf{X} = (\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_P)$ then the k^{th} component of $\mathbf{X}^\mathsf{T}w$ is $\boldsymbol{x}_k\boldsymbol{w}$. Thus, writing out the length of a vector component-wise

$$\|\mathbf{X}^\mathsf{T}\boldsymbol{w} - \boldsymbol{y}\|^2 = \sum_{k=1}^{P} \left(\boldsymbol{x}_k^\mathsf{T}\boldsymbol{w} - y_k\right)^2$$

which is just the squared error

$\overline{5}$

(b) Write down the cost you would minimise in vector form if you include a weight decay regularisation term with a regularisation parameter ν. *(3 marks)*

$$C(\boldsymbol{w}) = \|\mathbf{X}^\mathsf{T}\boldsymbol{w} - \boldsymbol{y}\|^2 + \nu\|\boldsymbol{w}\|^2$$

$\overline{3}$

(c) Obtain an equation for the weights that minimise this cost (show your working). *(8 marks)*

We can write the cost as

$$\begin{aligned} C(\boldsymbol{w}) &= (\mathbf{X}^{\mathsf{T}}\boldsymbol{w} - \boldsymbol{y})^{\mathsf{T}}(\mathbf{X}^{\mathsf{T}}\boldsymbol{w} - \boldsymbol{y}) + \nu\|\boldsymbol{w}\|^2 \\ &= \boldsymbol{w}^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}\boldsymbol{w} - 2\boldsymbol{w}^{\mathsf{T}}\mathbf{X}\boldsymbol{y} + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} + \nu\boldsymbol{w}^{\mathsf{T}}\boldsymbol{w} \\ &= \boldsymbol{w}^{\mathsf{T}}\left(\mathbf{X}\mathbf{X}^{\mathsf{T}} + \nu\mathbf{I}\right)\boldsymbol{w} - 2\boldsymbol{w}^{\mathsf{T}}\mathbf{X}\boldsymbol{y} + \boldsymbol{y}^{\mathsf{T}}\boldsymbol{y} \end{aligned}$$

Setting the gradient of the cost to zero

$$\boldsymbol{\nabla}C(\boldsymbol{w}) = 2\left(\mathbf{X}\mathbf{X}^{\mathsf{T}} + \nu\right)\boldsymbol{w} - 2\mathbf{X}\boldsymbol{y} = 0$$

or

$$\boldsymbol{w} = \left(\mathbf{X}\mathbf{X}^{\mathsf{T}} + \nu\mathbf{I}\right)^{-1}\mathbf{X}\boldsymbol{y}$$

$\frac{\quad}{8}$

(d) Explain why adding the regularisation term guarantees that the problem is never ill-posed and makes the solution better conditioned. *(9 marks)*

Let v_i be an eigenvector of $\mathbf{X}\mathbf{X}^{\mathsf{T}}$ so that $\mathbf{X}\mathbf{X}^{\mathsf{T}}v_i = \lambda_i v_i$. Then

$$\boldsymbol{v}_i^{\mathsf{T}}\mathbf{X}\mathbf{X}^{\mathsf{T}}\boldsymbol{v}_i = \lambda_i\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{v}_i \qquad\qquad \boldsymbol{u}_i^{\mathsf{T}}\boldsymbol{u}_i = \lambda_i\boldsymbol{v}_i^{\mathsf{T}}\boldsymbol{v}_i$$

where $u_i = \mathbf{X}^{\mathsf{T}}v_i$. Thus

$$\lambda_i = \frac{\|\boldsymbol{u}_i\|^2}{\|\boldsymbol{v}_i\|^2} \geq 0.$$

Now

$$\left(\mathbf{X}\mathbf{X}^{\mathsf{T}} + \nu\mathbf{I}\right)\boldsymbol{v}_i = (\lambda_i + \nu)\boldsymbol{v}_i$$

So matrix $\mathbf{X}\mathbf{X}^{\mathsf{T}} + \nu\mathbf{I}$ has eigenvalues $\lambda_i + \nu$ which are strictly greater than zero. Thus, the inverse is defined (so the problem is no longer ill-posed). Further the condition number is improved as

$$\frac{\lambda_{max} + \nu}{\lambda_{min} + \nu} \leq \frac{\lambda_{max}}{\lambda_{min}}.$$

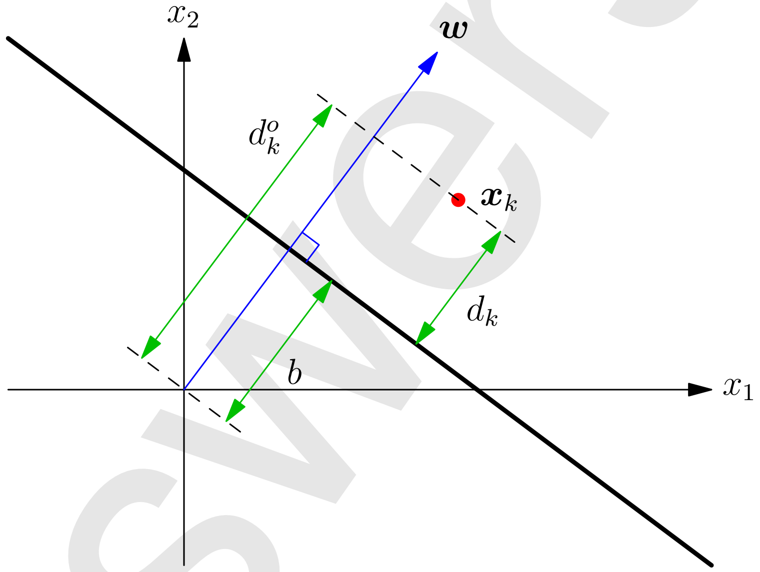
$\frac{\quad}{9}$

End of question 2

Q2: (a) $\frac{\quad}{5}$ (b) $\frac{\quad}{3}$ (c) $\frac{\quad}{8}$ (d) $\frac{\quad}{9}$ Total $\frac{\quad}{25}$

Question B 3

(a) Write down a formula for the minimum distance d_0^k between $\boldsymbol{x}_k$ and a hyperplane through the origin perpendicular to $\boldsymbol{w}$, and the minimum distance d_k from $\boldsymbol{x}_k$ to the hyperplane perpendicular to $\boldsymbol{w}$ displaced by b. *(5 marks)*



$$d_k^o = \frac{\boldsymbol{x}_k^T \boldsymbol{w}}{\|\boldsymbol{w}\|}$$

$$d_k = \frac{\boldsymbol{x}_k^T \boldsymbol{w}}{\|\boldsymbol{w}\|} - b$$

$\overline{5}$

(b) Depending on the category $y_k \in \{-1, 1\}$, write down the condition for a data point to be at least a distance m above (or below if $y_k = -1$) the hyperplane shown in part (a). *(3 marks)*

$$y_k \left(\frac{\boldsymbol{x}_k^T \boldsymbol{w}}{\|\boldsymbol{w}\|} - b \right) \geq m$$

$\overline{3}$

(c) Define $\boldsymbol{w'} = \boldsymbol{w}/(m\|\boldsymbol{w}\|)$ and $b' = b/m$, rewrite the condition above and explain why minimising $\|\boldsymbol{w}'\|^2$ is equivalent to maximising the margin m.
(3 marks)

$$y_k \left(\boldsymbol{x}_k^T \boldsymbol{w}' - b' \right) \geq 1$$

$\|\boldsymbol{w}'\|^2 = 1/m^2$**, thus minimising $\|\boldsymbol{w}'\|$ is equivalent to maximising w'**

$\overline{3}$

(d) Write down a Lagrangian for finding the maximal margin hyperplane for an SVM given data $(\boldsymbol{x}_k, y_k)$ for $k = 1, 2, \ldots, P$. *(3 marks)*

$$\mathcal{L}(\boldsymbol{w}', b', \boldsymbol{\alpha}) = \frac{\|\boldsymbol{w}'\|^2}{2} - \sum_{k=1}^{P} \alpha_k \left(y_k \left(\boldsymbol{x}_k^T \boldsymbol{w}' - b' \right) - 1 \right)$$

$\overline{3}$

(e) Write down the optimisation condition for the Lagrangian (i.e. what are you maximising or minimising with respect to) and what are the conditions on the Lagrange multipliers. *(3 marks)*

(i) Optimisation condition

$$\min_{\boldsymbol{w}', b'} \max_{\boldsymbol{\alpha}} \mathcal{L}(\boldsymbol{w}', b', \boldsymbol{\alpha})$$

$\overline{3}$

(ii) $\alpha_k \geq 0$ **for** $k = 1, 2, \ldots, P$

(f) Find the weight vector $\boldsymbol{w}'$ and threshold b' which minimises the Lagrangian and by substituting the result back into the Lagrangian find the dual form for optimisation problem. *(8 marks)*

Setting the derivatives with respect to w' to 0 we obtain

$$\boldsymbol{\nabla}\mathcal{L} = \boldsymbol{w}' - \sum_{k=1}^{P} \alpha_k y_k \boldsymbol{x}_k = 0$$

or

$$\boldsymbol{w}' = \sum_{k=1}^{P} \alpha_k y_k \boldsymbol{x}_k$$

also

$$\frac{\partial \mathcal{L}}{\partial b} = \sum_{k=1}^{P} \alpha_k y_k = 0$$

substituting back into the Lagrangian

$$\mathcal{L} = -\frac{1}{2} \sum_{k,l=1}^{P} \alpha_k \alpha_l y_k y_l \boldsymbol{x}_k^T \boldsymbol{x_l} + \sum_{k,l=1}^{P} \alpha_k$$

The dual optimisation problem is

$$\max_{\boldsymbol{\alpha}} -\frac{1}{2} \sum_{k,l=1}^{P} \alpha_k \alpha_l y_k y_l \boldsymbol{x}_k^T \boldsymbol{x_l} - \sum_{k,l=1}^{P} \alpha_k$$

$\overline{8}$

subject to

$$\alpha_k \geq 0 \quad \forall k = 1, \ldots, P \qquad\qquad \sum_{k=1}^{P} \alpha_k y_k = 0$$

End of question 3

Q3: (a) $\frac{}{5}$ (b) $\frac{}{3}$ (c) $\frac{}{3}$ (d) $\frac{}{3}$ (e) $\frac{}{3}$ (f) $\frac{}{8}$ Total $\frac{}{25}$

• Do not write in this space •

Question B 4

(a) Describe Bayes' rule giving a description of all the parts. *(5 marks)*

Given a hypothesis or parameters for a model, θ and data $\mathcal{D}$ Bayes' rule is

$$p(\boldsymbol{\theta}|\mathcal{D}) = \frac{p(\mathcal{D}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{p(\mathcal{D})}$$

where $p(\boldsymbol{\theta}|\mathcal{D})$ is the posterior which gives the updated probability distribution for the parameters θ of our model. $p(\mathcal{D}|\boldsymbol{\theta})$ is the likelihood of the data given the parameters. $p(\boldsymbol{\theta})$ is the prior probability expressing our prior belief of the parameters. Finally, $p(\mathcal{D})$ is a normalisation term, sometimes known as the evidence. $\overline{5}$

(b) Show that if the likelihood of observing n events is given by the Poisson distribution $p(n|\mu) = \mu^n \mathrm{e}^{-\mu}/n!$, then the Gamma distribution $p(\mu) = \mu^{a_0-1}\mathrm{e}^{-b_0\mu}$ is a conjugate distribution and compute the updated parameters of the posterior. *(5 marks)*

The posterior is proportional to

$$p(\mu|n) \propto p(n|\mu)p(\mu) = \frac{1}{n!}\mu^n \mathrm{e}^{-\mu}\mu^{a_0-1}\mathrm{e}^{-b_0\mu} \propto \mu^{a_0-1+n}\mathrm{e}^{-(b_0+1)\mu}$$

which is of the form of a Gamma distribution (hence conjugate). The updated parameters of the Gamma distribution are

$$a_1 = a_0 + n \qquad\qquad b_1 = b_0 + 1$$

$\overline{5}$

(c) Describe the MAP solution and explain its advantages and disadvantages over computing the posterior. *(5 marks)*

In the MAP solution we find the solution which maximises the posterior, or the log-posterior. The evidence is irrelevant as it is just a normalisation, thus

$$\begin{aligned}\boldsymbol{\theta}_{MAP} &= \operatorname{argmax} \log\left(p(\boldsymbol{\theta}|\mathcal{D})\right)\\ &= \operatorname{argmax}\left(\log\left(p(\mathcal{D}|\boldsymbol{\theta})\right) + \log\left(p(\boldsymbol{\theta})\right)\right)\end{aligned}$$

It is easier to compute than the full posterior as it does not involve any normalisation (which can be very difficult to compute). It also does not require describing a full distribution. However, it does not provide a full probabilistic solution which can be very misleading if the posterior is not sharply peaked. $\boxed{\overline{5}}$

(d) What is the naive Bayes assumption and explain how you would use it to implement a spam filter? *(10 marks)*

The naive Bayes assumption is that the all the data is conditionally independent, so if $\mathcal{D} = (d_i | i = 1, \ldots, n)$ then

$$p(\mathcal{D}|\boldsymbol{\theta}) = \prod_{i=1}^{n} p(d_i|\boldsymbol{\theta}).$$

To implement a spam filter we can treat all the words in the email as independent of each other. Given an email $\langle w_1, w_2, \ldots, w_n \rangle$ we can compute the probability of it being spam as

$$p(spam|\mathcal{D}) = \frac{\prod_{i=1}^{n} p(w_i|spam)\, p(spam)}{p(\mathcal{D})}$$

where $p(spam)$ is the empirically measured frequency of spam emails. To compute the likelihood we use a database of spam and non spam emails

$$p(w_i|spam) = \frac{\textbf{\# of occurances of } w_i \textbf{ in spam database}}{\textbf{\# of words in spam database}}$$

(we might include pseudo counts to make this more robust). The probability of the data is

$$p(\mathcal{D}) = p(\mathcal{D}|spam)\, p(spam) + p(\mathcal{D}|\neg spam)\, p(\neg spam)$$

We use exactly the same procedure to compute $p(\mathcal{D}|\neg spam)$ as we did to compute $p(\mathcal{D}|spam)$ (i.e. independence assumption and word count). $\boxed{\frac{\quad}{10}}$

End of question 4

Q4: (a) $\frac{\quad}{5}$ (b) $\frac{\quad}{5}$ (c) $\frac{\quad}{5}$ (d) $\frac{\quad}{10}$ Total $\frac{\quad}{25}$

END OF PAPER