UNIVERSITY OF SOUTHAMPTON COMP3008W1

SEMESTER 2 EXAMINATION 2011/2012

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.

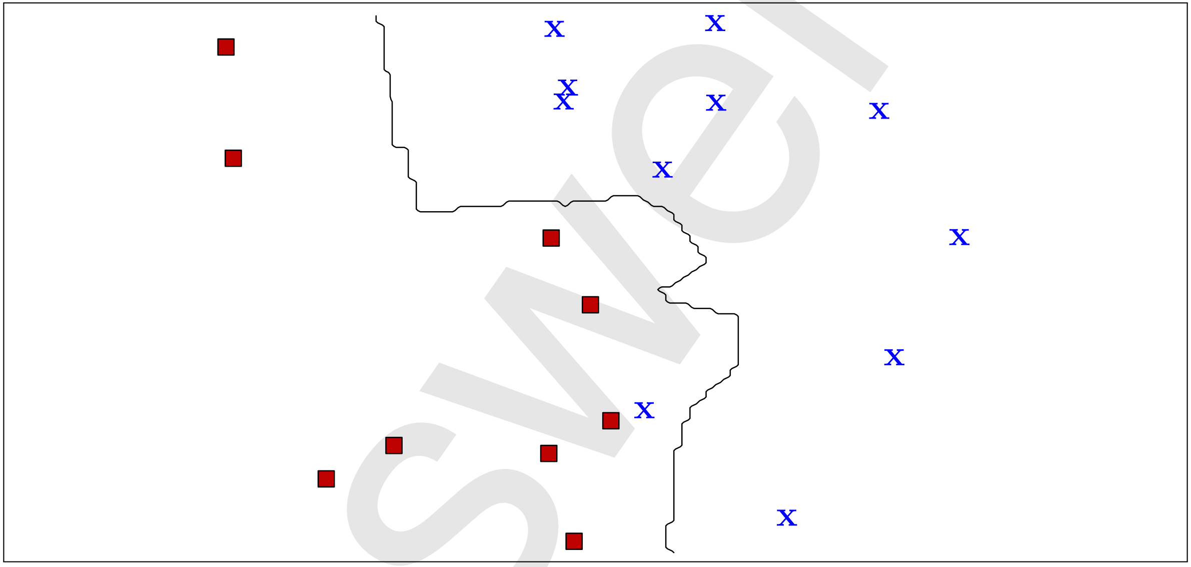
You may use a blue book for scratch—it will be discarded without being looked at.

Number of pages 12

Section A

Question A 1

(a) Roughly sketch the dividing curve for a 3-nearest-neighbours classifier for the data shown below. *(5 marks)*

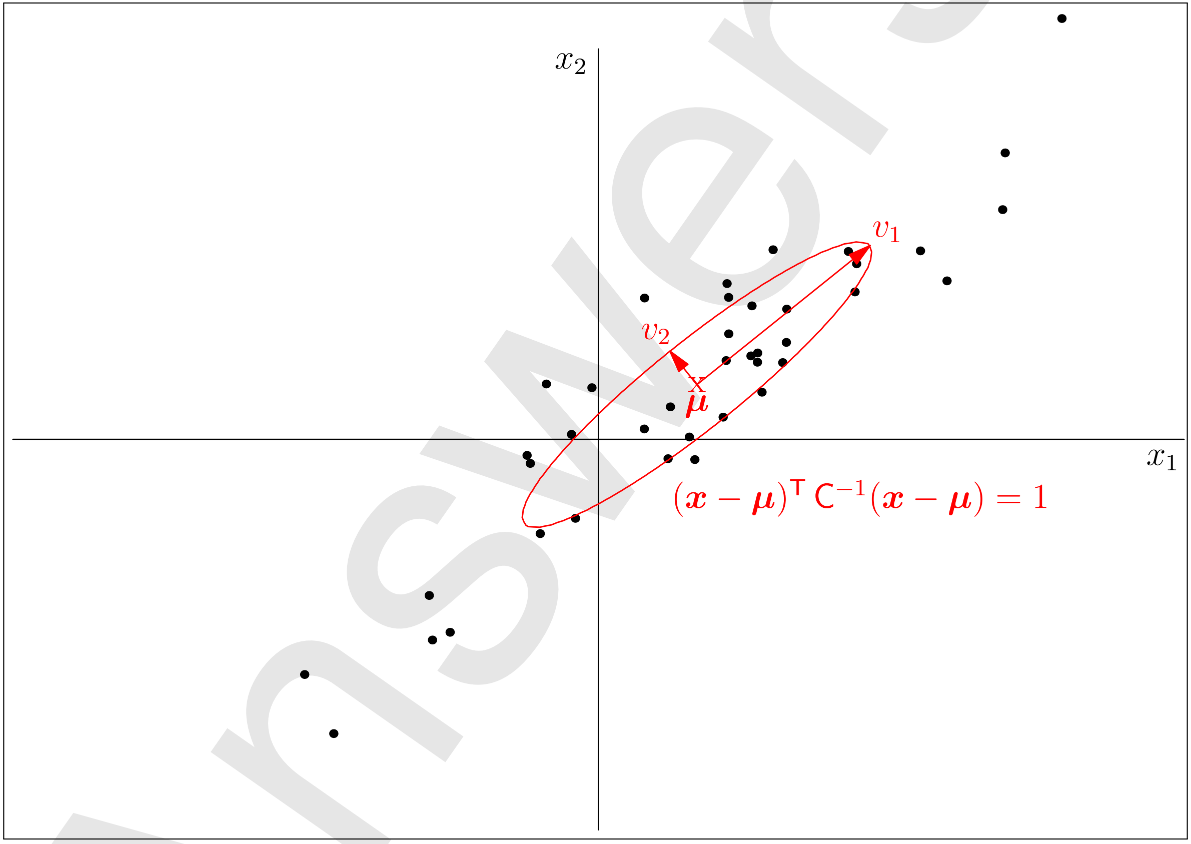


(b) Explain why the kernel trick allows an SVM to separate data that is not linearly separable. *(5 marks)*

(Test knowledge and understanding of kernels.)

The kernel trick projects data into and extended feature space (the space of eigenfunctions of the kernel). Although an SVM finds a linear separating plane in this extended space, as the extended features are typically a non-linear function of the original features this corresponds to finding a non-linear separating surface in the original space.

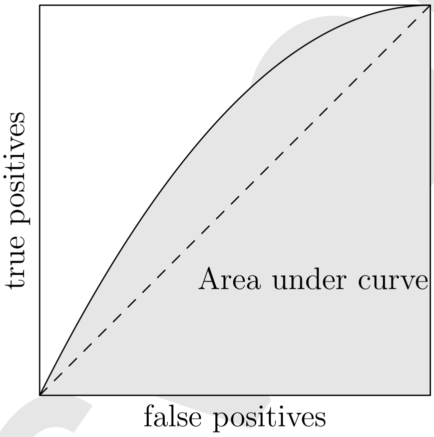
(c) Below we show a set of two-dimensional data points. Show the approximate position of the mean, $\boldsymbol{\mu}$ and sketch the contour $(\boldsymbol{x}-\boldsymbol{\mu})^{\mathsf{T}}\,\mathbf{C}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = 1$, where $\mathbf{C}$ is the covariance matrix. Also show the direction of the leading eigenvector (principle component), $\boldsymbol{v}_1$, and the eigenvector with the second largest eigenvalue, $\boldsymbol{v}_2$. *(5 marks)*



• Do not write in this space •

(d) Explain when you would use an ROC curve and sketch a typical curve, labelling the axes. *(5 marks)*

A ROC curve is used to evaluate the performance of a binary classify with a variable threshold. The ROC shows how you can play off false negative rate versus the true positive rate.



End of question 1

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

Section B

Question B 2

(a) Show that a multi-layer perceptron (MLP) with linear nodes is no more powerful than a single-layer perceptron. *(5 marks)*

(Fairly simple introductory question.)

The response of a multi-layer linear perceptron is equal to

$$\sum_{i\in\textbf{Hidden Units}} \bar{w}_i \sum_j w_i^j\, x_j = \sum_j x_j \sum_{i\in\textbf{Hidden Units}} \bar{w}_i\, w_i^j = \sum_j \tilde{w}_j x_j$$

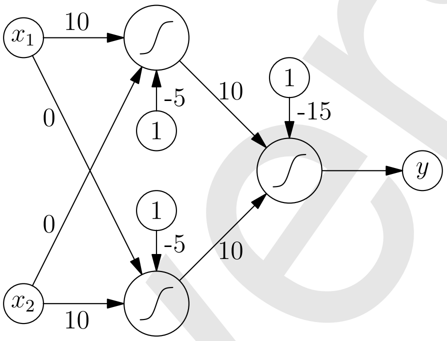
where $\tilde{w}_j = \sum_{i\in\textbf{Hidden Units}} \bar{w}_i\, w_i^j$. This is just the response of a single layer perceptron with weights $\tilde{\boldsymbol{w}}$

(b) Describe how to train a multi-layer perceptron. *(8 marks)*

(This is more a book work question, although it covers a technically challenging topic.)

A multi-layer perceptron is trained by iteratively reducing the learning error. This is achieved by computing the gradient of the learning error. The direction of steepest descent is the direction of the negative gradient. We can either do a gradient descent or use a more sophisticated (Newton or pseudo-Newton method) such as Levenberg-Marquard. The computation of the gradient is known as back-propagation. For the logistic perceptron and tanh perceptron the derivative can be computed in terms of the response function itself. As a consequence the gradient can be computed efficiently.

(c) For the multi-layer perceptron shown, write the response function $f(\boldsymbol{x}|\boldsymbol{w})$ explicitly as a function of the input $\boldsymbol{x} = (x_1, x_2)$. Compute the response for the inputs given in the table. Finally sketch the line where $f(\boldsymbol{x}|\boldsymbol{w}) = 1/2$. Approximate values are sufficient.

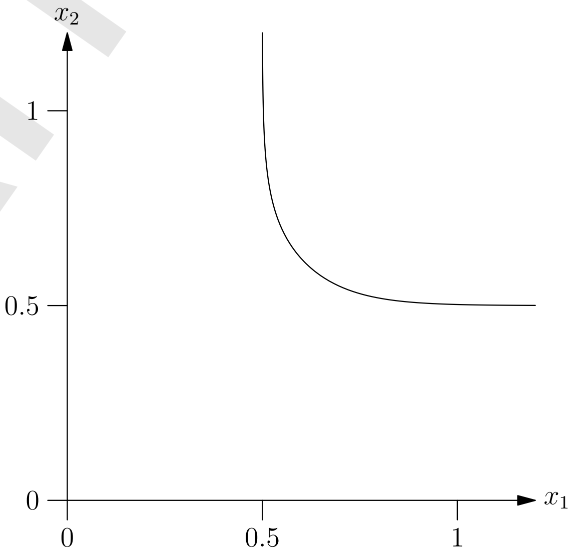


Assume that the output of the nodes are given by $g(V) = \frac{1}{1+e^{-V}}$. *(12 marks)*

(Test understanding of MLPs. Students have not seen this example.)

$$f(\mathbf{x}|\mathbf{w}) = g\Big(10\, g(10\, x_1 - 5) + 10\, g(10\, x_2 - 5) - 15\Big)$$

$\mathbf{x}$	$(0,0)$	$(0,0.5)$	$(0,1)$	$(0.5,0)$	$(0.5,0.5)$	$(0.5,1)$	$(1,0)$	$(1,0.5)$	$(1,1)$
$f(\mathbf{x}\|\mathbf{w})$	3.5e-07	4.9e-05	0.0067	4.9e-05	0.0067	0.48	0.0067	0.48	0.99



End of question 2

Q2: (a) $\frac{\quad}{5}$ (b) $\frac{\quad}{8}$ (c) $\frac{\quad}{12}$ Total $\frac{\quad}{25}$

- Do not write in this space •

Question B 3

(a) Consider a finite set of hypotheses $\mathcal{H}$ for a binary classification task. If $h \in \mathcal{H}$ has an error rate ϵ, calculate the probability that it will make no error on P randomly selected (i.e. independent) patterns. *(5 marks)*

(This is a conceptually challenging part of the course, although the mathematics is not hard.)

The probability of correctly classifying one pattern is $1-\epsilon$. As the patterns are independent the probability of correctly classifying all of them is $(1-\epsilon)^P$

(b) Explain why the probability of any hypothesis with an error rate greater than ϵ will correctly classifying P patterns is bounded by $|\mathcal{H}|\, \mathrm{e}^{-\epsilon P}$. *(5 marks)*

The number of hypothesis with an error greater than ϵ is bounded above by the total number of hypotheses $|\mathcal{H}|$. The probability of any of them correctly classifying all the patterns is bounded by $(1-\epsilon)^P$. But, $1-\epsilon < \mathrm{e}^{-\epsilon}$ so the probability that any one of them correctly classifies all P patterns is strictly less than or equal to $|\mathcal{H}|\, \mathrm{e}^{-\epsilon P}$.

(c) Obtain a bound on the number of patterns required to ensure that a consistent learner (i.e. a machine that finds a hypothesis which is consistent with all the input patterns) will have an error less than ϵ with a probability of, at least, δ. *(5 marks)*

The probability of a consistent learner returning a hypothesis with an error greater than ϵ is $|\mathcal{H}|\, \mathrm{e}^{-\epsilon P}$. We require this probability to be less than δ. Taking logarithms (which does not change the inequality as it is monotonic), we find

$$\log(|\mathcal{H}|) - \epsilon\, P < \log(\delta)$$

or

$$P > \frac{1}{\epsilon}\left(\log(|\mathcal{H}|) + \log\left(\frac{1}{\delta}\right)\right).$$

(d) Given a hypothesis space of with $|\mathcal{H}| = 10^{10}$ hypotheses, how many patterns do you need to guarantee an error rate less that 0.1% with a probability of at least 99.99%? *(5 marks)*

(This is just substituting numbers into the formula, although it require understanding what these numbers mean.)

We have $\epsilon = 0.001$ and $\delta = 10^{-4}$, thus

$$P > 10^3 \left(\log\left(10^{10}\right) + \log\left(10^4\right)\right) = 14\,000\ \log(10) = 32\,236$$

(e) Explain what the VC-dimension is and why it is needed? *(5 marks)*

The VC-dimension measures the capacity of a learning machine. That is the number of linearly separable patterns at which a learning machine cannot achieve all possible dichotomies of the inputs. This is used to obtain similar bounds to the given above, but for machines with continuous parameters so that total number of possible hypotheses is infinite.

End of question 3

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

• Do not write in this space •

Question B 4

(a) A patient comes to the doctor with a symptom S. The doctor knows that all patients with a rare disease A have this symptom. However, 0.5% of patients with a common disease B also exhibit this symptom. If the probability of disease A is 10^{-6} and the probability of disease B is 10^{-3} what is the probability of the patient having disease A? (Show your working.) *(10 marks)*

(A straightforward application of Bayes rule, but requires the student to understand how to apply it.)

We have been given the following likelihoods

$$\mathbb{P}(S|A) = 1, \qquad \mathbb{P}(S|B) = 0.005,$$

In addition we have the priors

$$\mathbb{P}(A) = 10^{-6} \qquad \mathbb{P}(B) = 10^{-3}$$

Now

$$\begin{aligned}\mathbb{P}(S) &= \mathbb{P}(S, A) + \mathbb{P}(S, B) = \mathbb{P}(S|A)\,\mathbb{P}(A) + \mathbb{P}(S|B)\,\mathbb{P}(B)\\ &= 10^{-6} + 0.005 \times 10^{-3} = 6 \times 10^{-6}\end{aligned}$$

Thus

$$\mathbb{P}(D|S) = \frac{\mathbb{P}(S|A)\ \mathbb{P}(A)}{\mathbb{P}(S)} = \frac{10^{-6}}{6 \times 10^{-6}} = \frac{1}{6} = 0.16667$$

(b) Explain how to set up a collaborative filter recommender system as a matrix completion problem. Describe an approximation in terms of the mean rating for users and items. Explain how to approximate the residual matrix using a low rank approximation. Show how we can introduces priors into the problem and outline how we can find a MAP solution to the problem. *(15 marks)*

(Students have been shown this, but it is conceptually challenging involving many steps and different concepts.)

In collaborative filtering we have a set of users and items. The users have each recommended some of the items but not all. We can think of all possible user-item pairs as elements of a matrix where we know a few elements and are trying to deduce all the others. Let U is the set of users and I is the set of items. The we can define the $|U| \times |I|$ matrix R with elements R_{ui}. We can approximate the unknown ratings by

$$R_{ui} \approx \bar{R}_u + \bar{R}_i - \bar{R}$$

where $\bar{R}_u$ is the mean rating of user u, $\bar{R}_i$ is the mean rating of item i and $\bar{R}$ is the overall mean rating. To obtain a more accurate rating we can estimate the residue matrix $\tilde{\mathbf{R}}$ with elements $\tilde{R}_{ui} = R_{ui} - \bar{R}_u - \bar{R}_i + \bar{R}$ using a low rank approximation.

$$\tilde{\mathbf{R}} = \mathbf{A}\mathbf{B}^{\mathsf{T}}$$

where A is an $|U| \times K$ matrix and B is a $|I| \times K$ matrix. K is the rank of the approximation with K typically much small than either $|U|$ or $|I|$. We can introduce normal priors on the elements of A and B

$$f(A_{uk}) = \mathcal{N}(0, \sigma_A^2) \qquad\qquad f(B_{uk}) = \mathcal{N}(0, \sigma_B^2)$$

where we choose σ_A^2 and σ_B^2 to explain the typical variance in the elements of $\tilde{\mathbf{R}}$. If we assume that the low rank approximation has normally distributed error then

$$\tilde{R}_{ui} - \sum_{k=1}^{K} A_{uk}\, B_{ik} \sim \mathcal{N}(0, \sigma^2)$$

where σ^2 is a parameter encoding the expected error in the approximation. Under these assumptions the log-posterior is given by

$$-\frac{1}{2\sigma^2} \sum_{(u,i)\in\mathcal{D}} \left(\tilde{R}_{ui} - \sum_{k=1}^{K} A_{uk}\, B_{ik} \right) - \frac{1}{2\sigma_A^2} \sum_{k=1}^{K} \sum_{u\in\mathcal{U}} A_{uk}^2 - \frac{1}{2\sigma_B^2} \sum_{k=1}^{K} \sum_{i\in\mathcal{I}} B_{ik}^2 + \textbf{const}$$

The MAP solution is given by the maximising this quantity. This can be done iteratively starting from keeping A fixed and optimising with respect to B. We can then switch to keeping B and optimising with respect to A. Switching backwards and forwards we rapidly converge to a solution.

End of question 4

Q4: (a) $\frac{}{10}$ (b) $\frac{}{15}$ Total $\frac{}{25}$

END OF PAPER