UNIVERSITY OF SOUTHAMPTON COMP3008W1

SEMESTER 2 EXAMINATION 2006/2007

MACHINE LEARNING

Duration: 120 mins

Answer ALL *questions from section A (20 marks)*
and ONE *question from section B (25 marks)*
and ONE *question from section C (25 marks).*

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

Calculators without text storage MAY be used.

 Number of pages 11

Section A

Question 1

(a) Explain what is meant when a problem is under-constrained in the context of machine learning. How are under-constrained problems solved?

Test understanding of mathematical setting of ML.

A problem is under-constrained if there are more variables than equations. In machine learning this typically happens when the learning machine has more free parameters than there are data points (although not necessarily).

In under-constrained systems there are many possible solutions. In ML we overcome these problems by choosing a 'simple solution'. One way this can be accomplished is by introducing a regularisation term which provides a bias towards simple solutions.

(3 marks)

(b) Explain what is meant when a problem is over-constrained in the context of machine learning. How are over-constrained systems solved?

Follow on from previous question.

A problem is over-constrained if there are more equations than variables. In machine learning this typically happens when the learning machine has insufficient flexibility to correctly predict all the training data.

The problem is solved by minimising an error function which finds a machine that explains the training data as best it can.

(3 marks)

(c) Describe the perceptron learning algorithm for the step perceptron.

Tests knowledge of classic ML algorithm.

Starting from data $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{P}$ **with** $y_k \in \{-1, 1\}$

(i) Initialise: $\boldsymbol{w} = 0$**,** $b = 0$ **and** $R = \max_k \|\boldsymbol{x}_k\|^2$

(ii) Repeat

- **for** $k = 1$ **to** P
  - **if** $\left(y_k(\boldsymbol{w}^{\mathsf{T}}\boldsymbol{x}_k - b) \leq 0\right)$ **then**
    - $\boldsymbol{w} \leftarrow \boldsymbol{w} + y_k\boldsymbol{x}_k$
    - $b \leftarrow b - y_k\, R$

  end if

end for

Until no mistakes

(4 marks)

(d) Steve's questions *(10 marks)*

Section B

Question 2

(a) Write down the mean squared training error, $E(\boldsymbol{w}|\mathcal{D})$, for a leaning machine $f(\boldsymbol{x}|\boldsymbol{w})$ given a finite data set $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^{P}$

Start with easy question

$$E(\boldsymbol{w}|\mathcal{D}) = \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_k, y_k) \in \mathcal{D}} \big(f(\boldsymbol{x}_k|\boldsymbol{w}) - y_k\big)^2$$

(1 marks)

(b) The bias-variance dilemma is a theoretical analysis for understanding the causes of generalisation error. In this approach we consider the generalisation error when averaged over infinitely many randomly chosen data sets $\mathcal{D}$. Write down an equation for the generalisation error in terms of an average over training sets.

This tests a key concept in formalising the bias-variance dilemma.

$$E_g(\boldsymbol{w}) = \left\langle \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_k, y_k) \in \mathcal{D}} \big(f(\boldsymbol{x}_k|\boldsymbol{w}) - y_k\big)^2 \right\rangle_{\mathcal{D}}$$

where $\langle \cdots \rangle_{\mathcal{D}}$ denotes the average of data sets.

(2 marks)

(c) Denoting the set of weights that minimise the training error for a data set by $\boldsymbol{w}(\mathcal{D})$ write down an expression for the response of the average machine, $f_{av}(\boldsymbol{x})$ (i.e. the mean response averaged over machines trained on all data sets).

Still fairly simple, although quite abstract.

The mean response averaged over machines trained on all data sets is given by

$$f_{av}(\boldsymbol{x}) = \langle f(\boldsymbol{x}|\boldsymbol{w}(\mathcal{D}))\rangle_{\mathcal{D}}$$

(2 marks)

(d) Show the generalisation error can be written as a bias term and variance term. (Hint, add and subtract the average machine response to the generalisation error, then expand the square.) Explain what these two terms means and explain the dilemma.

The students have seen this derivation. However, the analysis is both mathematically reasonably challenging and requires sophistication to understand.

Starting from the generalisation error

$$\begin{aligned} E_g(\boldsymbol{w}) &= \left\langle \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_k, y_k)\in\mathcal{D}} \big(f(\boldsymbol{x}_k|\boldsymbol{w}(\mathcal{D})) - y_k\big)^2 \right\rangle_{\mathcal{D}} \\ &= \left\langle \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_k, y_k)\in\mathcal{D}} \big(f(\boldsymbol{x}_k|\boldsymbol{w}(\mathcal{D})) - f_{av}(\boldsymbol{x}) + f_{av}(\boldsymbol{x}) - y_k\big)^2 \right\rangle_{\mathcal{D}} \end{aligned}$$

We expand the squared term as

$$\big(f(\boldsymbol{x}_k|\boldsymbol{w}) - f_{av}(\boldsymbol{x}) + f_{av}(\boldsymbol{x}) - y_k\big)^2 = \big(f(\boldsymbol{x}_k|\boldsymbol{w}(\mathcal{D})) - f_{av}(\boldsymbol{x})\big)^2 + 2\big(f(\boldsymbol{x}_k|\boldsymbol{w}(\mathcal{D})) - f_{av}(\boldsymbol{x})\big)\big(f_{av}(\boldsymbol{x}) + \big(f_{av}(\boldsymbol{x}) - y_k\big)^2$$

Substituting this back we get three terms

$$E_g(\boldsymbol{w}) = B + V + C$$

where

$$B = \left\langle \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_k, y_k)\in\mathcal{D}} \big(f_{av}(\boldsymbol{x}) - y_k\big)^2 \right\rangle = \big(f_{av}(\boldsymbol{x}) - y_k\big)^2$$

is the bias and measures how far the average term is from the targets

$$V = \left\langle \frac{1}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_k, y_k)\in\mathcal{D}} \big(f(\boldsymbol{x}_k|\boldsymbol{w}(\mathcal{D})) - f_{av}(\boldsymbol{x})\big)^2 \right\rangle$$

is the variance and measures the fluctuations in the response of the machines trained on different training sets. Finally,

$$C = \left\langle \frac{2}{|\mathcal{D}|} \sum_{(\boldsymbol{x}_k, y_k) \in \mathcal{D}} \left(f_{av}(\boldsymbol{x}) - y_k \right) \left(f(\boldsymbol{x}_k | \boldsymbol{w}(\mathcal{D})) - f_{av}(\boldsymbol{x}) \right) \right\rangle = 0$$

is the cross term which cancels because the only term which depends on the data set is $f(\boldsymbol{x}_k|\boldsymbol{w}(\mathcal{D}))$. However the average of this quantity is equal to $f_{av}(\boldsymbol{x})$.

The bias measures the limitation of the machine to learn the underlying function. Simpler machines will typically have larger biases. The variance measures the sensitivity of the machine to a particular data set. The more complex the learning machine usually the larger the variance. Thus the dilemma is to obtain a machine that is powerful enough to have a small learning error but not too powerful to have a large variance.

(17 marks)

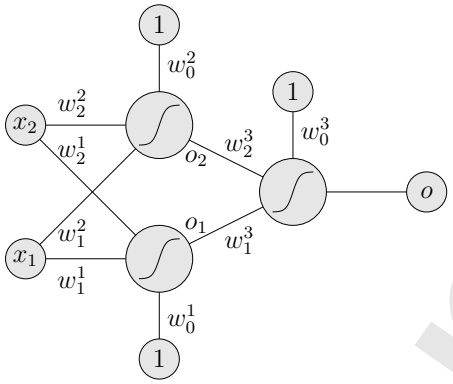
(e) Explain why regularisation terms can improve generalisation performance in the context of the bias-variance dilemma. Why do regularisation terms allow more complex machine to be used.

Testing ability to put together the big picture.

Regularisation terms reduce the variance by favouring simple machines. Although they do this at the expense of "learning the wrong thing". In consequence they can lead to a worse bias, but a much improved variance. Usually complex machines would tend to badly over-fit the data leading to a large variance. Using a regularisation term allows you to use a more complex machine as ti will strongly reduce the variance. As the machine is complex it can also achieve a low bias.

(3 marks)

Question 3 Consider the multilayer perceptron shown below



where the response function for the nodes are some squashing function $g(x)$.

(a) Write down a formula describing the output, $o = f(\boldsymbol{x}|\boldsymbol{w})$, of this network shown above.

Test understanding of formulating ML in terms of equations.

$$o = f(\boldsymbol{x}|\boldsymbol{w}) = g(w_1^3 g(w_1^1 x_1 + w_2^1 x_2 + w_0^1) + w_2^3 g(w_1^2 x_1 + w_2^2 x_2 + w_0^2) + w_0^3)$$

(2 marks)

(b) Write down an expression for the mean squared error and find the derivative with respect to w_1^3 and w_1^1.

Test mathematical manipulation skills needed to train MLPs.

The mean squared error given a data set $\mathcal{D} = \{(\boldsymbol{x}_k, y_k)\}_{k=1}^P$ is

$$E(\boldsymbol{w}|\mathcal{D}) = \frac{1}{P}\sum_{k=1}^{P}\left(f(\boldsymbol{x}_k|\boldsymbol{w}) - y_k\right)^2$$

where $f(\boldsymbol{x}|\boldsymbol{w})$ is the response of the network for an input $\boldsymbol{x}$. The derivative with respect to w_1^3 is

$$\frac{\partial E(\boldsymbol{w}|\mathcal{D})}{\partial w_1^3} = \frac{2}{P}\sum_{k=1}^{P}\left(f(\boldsymbol{x}_k|\boldsymbol{w}) - y_k\right) g'(w_1^3 o_1 + w_2^3 o_2 + w_0^3) o_1$$

where

$$o_1 = g(w_1^1 x_1 + w_2^1 x_2 + w_0^1) \qquad o_2 = g(w_1^2 x_1 + w_2^2 x_2 + w_0^2)$$

The derivative with respect to w_1^3 is

$$\frac{\partial E(\boldsymbol{w}|\mathcal{D})}{\partial w_1^1} = \frac{2}{P}\sum_{k=1}^{P}(f(\boldsymbol{x}_k|\boldsymbol{w}) - y_k)\, g'(w_1^3 o_1 + w_2^3 o_2 + w_0^3) g'(w_1^1 x_1 + w_2^1 x_2 + w_0^1) x_1$$

(5 marks)

(c) Assuming the following weights $w_0^1 = -10$, $w_1^1 = 0$, $w_2^1 = 20$, $w_0^2 = 30$, $w_1^2 = 0$, $w_2^2 = -20$, $w_0^3 = -30$, $w_1^3 = 20$, and $w_2^2 = 20$. Calculate the output o given inputs, (x_1, x_2) equal to $(0, 0)$, $(-1, 1)$, $(1, 1)$, and $(0, 2)$ for a logistic perceptrons $g(x) = 1/(1 - \mathrm{e}^{-x})$, (you may assume $g(x) = 0$ for $x \leq -10$ and $g(x) = 1$ for $x \geq 10$).

Can students plug in numbers

The response is now

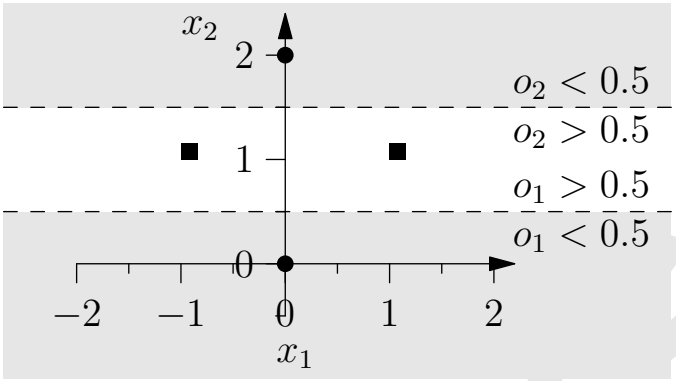
$$o = g(20g(20x_2 - 10) + 20g(30 - 20x_2) - 30)$$

(x_1, x_2)	$o_1 = g(20x_2 - 10)$	$o_2 = g(30 - 20x_2)$	$o = g(20o_1 + 20o_2 - 30)$
$(0, 0)$	**0**	**1**	**0**
$(-1, 1)$	**1**	**1**	**1**
$(1, 1)$	**1**	**1**	**1**
$(0, 2)$	**1**	**0**	**0**

(8 marks)

(d) Draw the separating planes for the two hidden nodes in input space (i.e. when their input fields equal zero). Which features are relevant to solving this problem?

Test deep understanding of how MLPs work.



For this problem input x_1 is entirely irrelevant to the classification. We could solve the problem using only input x_2.

(8 marks)

(e) Why could this problem not be solved using a single layer perceptron?

Test understanding of linear separability

This problem is not linearly separable (it is an XOR problem on its side).

(2 marks)

Section C

Question 4

(25 marks)

Question 5

(25 marks)

END OF PAPER