UNIVERSITY OF SOUTHAMPTON COMP3008W1

SEMESTER 2 EXAMINATION 2007/2008

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 two-class pattern classification problem, with classes denoted ω_1 and ω_2, is defined by a *scalar* feature x. The class conditional densities $p\,(x \,|\, \omega_1)$ and $p\,(x \,|\, \omega_2)$ are Gaussian, with means of -0.5 and 0.5 respectively. The variances of the two densities are equal to 1.0. Prior probabilities are $P\,[\omega_1] = 0.7$ and $P\,[\omega_2] = 0.3$ respectively.

(a) Taking a real-world example of your choice, explain what the different terms in the above problem statement mean.

Whatever example chosen to explain, identify in answer that class conditional densities are modelled by gathering some data, extracting features for the training data and fitting some parametric density to it. Prior probabilities come from some other information about the problem domain – a language model in speech sound classification etc.

(8 marks)

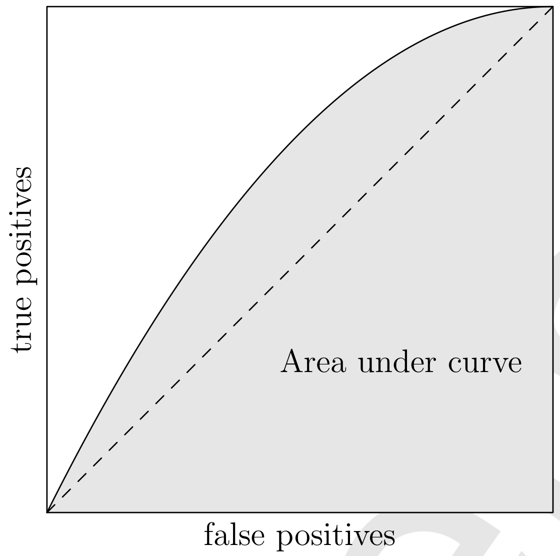
(b) Derive the Bayes optimal decision rule for this problem.

Derivation should substitute Gaussian expression into Bayes rule and show that the variance terms cancel out. Optimal classifier will be a threshold, look in answer for a shift away from the mean of the apriori more frequent class. This was taught in class for the general case – features vectors, but never with a scalar feature, question tests if they can map the algebra to a simpler situation than covered in class).

(8 marks)

(c) What is a Receiver Operating Characteristics (ROC) curve? Sketch the ROC curve for the problem above. Carefully identify how *true positive* and *false positive* rates are calculated.

ROC curve plots true positives against false negatives for varying classification thresholds.



(8 marks)

(d) Explain, without derivation, how the Bayes optimal classifier for the above problem will differ, if the variances of the two densities were different from each other.

When the variances are not the same, the optimal classifier is quadratic; in the case of a one dimensional problem, such as this one, there will be two decision thresholds: class one of the measurement is bigger than the first *and* lower than the second. And vice versa for the other class.

(9 marks)

Question 2

(a) Explain clearly how a linear regression problem may be solved using an *on-line* gradient descent algorithm contrasting it with the corresponding *batch* solution. How does this differ from the *perceptron*?

Least squares solution is given by pseudo-inverse using notation used in class $a = \left(Y^{\mathsf{T}} Y\right)^{-1} Y^{\mathsf{T}} f$ for the linear regression problem, identify that the error as a function of the unknowns is quadratic; identify that the gradient point towards the solution (preferably with a sketch). Write the gradient descent algorithm with batch update, identifying that the full gradient is computed as a sum over all training data. On-line algorithms use a noisy estimate of the gradient, without summing over all the data. Look in answer for how the error might decrease during training – smooth for batch update and noisy for on-line update.

This differs from perceptron because perceptron update is *only* with data that is misclassified; when the data is correctly classified perceptron algorithm makes no update – hence error correcting learning.

Full marks if the explanation covers all of the above points.

(16 marks)

(b) The equation for a multivariate Gaussian density is given by the formula

$$p(\boldsymbol{x}) = \frac{1}{(2\pi)^{d/2}\, |\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{m})^{\mathsf{T}}\, \boldsymbol{\Sigma}^{-1}\, (\boldsymbol{x}-\boldsymbol{m})\right\}$$

(i) Briefly explain what $\boldsymbol{m}$ and $\boldsymbol{\Sigma}$ are, and how they may be estimated from a sample of data $\{\boldsymbol{x}_1,\, \boldsymbol{x}_2,\, \dots \boldsymbol{x}_N\}$

(ii) Compute the eigenvalues and eigenvectors of $\boldsymbol{\Sigma}$ when it takes the values

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 2 & 1 \\ 1 & 2 \end{pmatrix} \text{ and } \begin{pmatrix} 2 & -1 \\ -1 & 2 \end{pmatrix}.$$

(iii) Sketch bivariate densities having zero means and $\boldsymbol{\Sigma}$ taking the values above. Briefly explain how computing the eigenvalues was helpful.

Mean and covariance; estimated from a sample of data as

$$\boldsymbol{m} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{x}_i \text{ and } \boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N} (\boldsymbol{x}-\boldsymbol{m})\,(\boldsymbol{x}-\boldsymbol{m})^{\mathsf{T}}$$

Eigenvalues for $\begin{pmatrix} \mathbf{1} & \mathbf{0} \\ \mathbf{0} & \mathbf{1} \end{pmatrix}$ **one and any vector is an eigenvector.**

Eigenvalues for $\begin{pmatrix} \mathbf{2} & \mathbf{1} \\ \mathbf{1} & \mathbf{2} \end{pmatrix}$ **are three and one, and the corresponding eigenvectors are** $\begin{pmatrix} \mathbf{1} \\ \mathbf{1} \end{pmatrix}$ **and** $\begin{pmatrix} \mathbf{1} \\ \mathbf{-1} \end{pmatrix}$**.**

Eigenvalues for $\begin{pmatrix} \mathbf{2} & \mathbf{-1} \\ \mathbf{-1} & \mathbf{2} \end{pmatrix}$ **are three and one, and the corresponding eigenvectors are** $\begin{pmatrix} \mathbf{1} \\ \mathbf{-1} \end{pmatrix}$ **and** $\begin{pmatrix} \mathbf{1} \\ \mathbf{1} \end{pmatrix}$**.**

Sketches: isotropic covariance should be circular and the other two elliptical with major axes in the direction of the eigenvector corresponding to the largest eigenvalue.

Answer for how eigenvectors helped – extra credit for identifying the major axes with principal components.

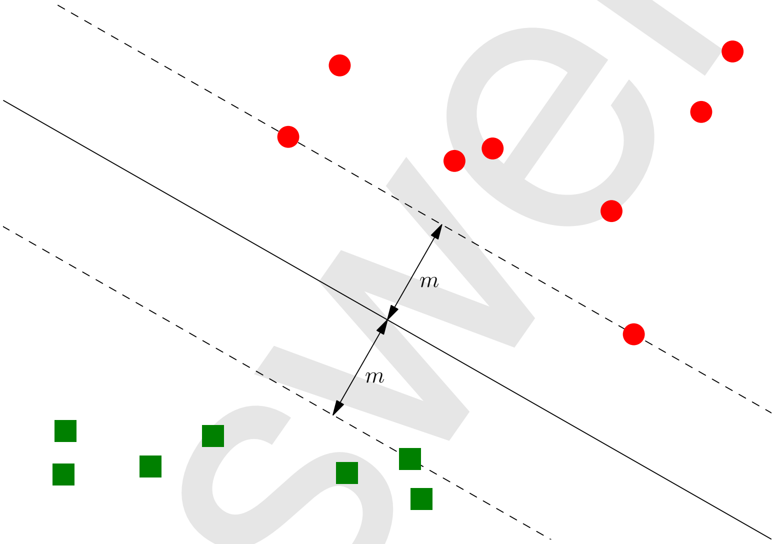
(17 marks)

Question 3

(a) Define the maximum stability perceptron and sketch a diagram showing where the dividing hyperplane would lie.

Easy question to test basic understanding.

The maximum stability perceptron is a perceptron where the dividing plane is chosen so as to maximise the margin from the dividing plane to any of the data points.



(6 marks)

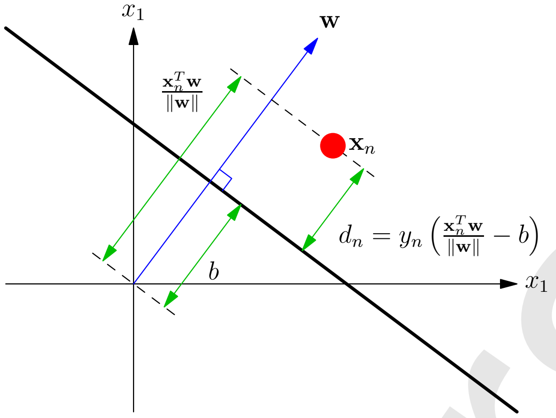
(b) Derive a quadratic programming formulation for finding the maximum-margin hyperplane

Harder derivation although covered in lectures.

Given a data set $\mathcal{D} = \{\boldsymbol{x}_n, y_n\}_{n=1}^{N}$ where $y_n = \{-1, 1\}$ we can write the distance to the hyperplane of data point n as

$$d_n = y_n \left(\frac{\boldsymbol{x}^\mathsf{T} \boldsymbol{w}}{\|\boldsymbol{w}\|} - b \right)$$

where the hyperplane is defined by the vector w and the threshold b.


x_1
$\mathbf{w}$
$\frac{\mathbf{x}_n^T \mathbf{w}}{\|\mathbf{w}\|}$
$\mathbf{x}_n$
$d_n = y_n\left(\frac{\mathbf{x}_n^T\mathbf{w}}{\|\mathbf{w}\|} - b\right)$
b
x_1

We to choose the weights w and threshold b to maximise the smallest distance, d_n, to the dividing hyperplane. That is we want to maximise the margin. Denoting the margin by m then we require all the data points satisfy the constraint

$$d_n = y_n \left(\frac{\boldsymbol{x}^\mathsf{T} \boldsymbol{w}}{\|\boldsymbol{w}\|} - b \right) \geq m$$

Dividing through by m and defining $\boldsymbol{w}' = \boldsymbol{w}/(m\|\boldsymbol{w}\|)$ and $b' = b/m$ we can write the constraints as

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

where $\|\boldsymbol{w}'\| = 1/m$. Thus the problem of finding the maximum margin hyperplane reduces to finding $\boldsymbol{w}'$ and b' which minimises $\|\boldsymbol{w}'\|$ (or equivalently $\|\boldsymbol{w}'\|^2$), thus maximising the margin m. We can write this as a quadratic programming problem

choose $\boldsymbol{w}'$ and b' to minimise $\|\boldsymbol{w}'\|^2$
subject to $y_n \left(\boldsymbol{x}^\mathsf{T} \boldsymbol{w}' - b' \right) \geq 1$

(10 marks)

(c) Consider the problem of minimising

$$f(x, y) = 2\,x^2 + y^2 - 4\,x\,y$$

subject to the constraint $x - 2\,y = 2$. Write down a Lagrangian describing the problem.

The rest of this question tests students understanding of how to solving problems with constraints

$$\mathcal{L} = 2\,x^2 + y^2 - 4\,x\,y + \lambda(x - 2\,y - 2).$$

(3 marks)

(d) Write down the optimisation conditions for the Lagrangian and solve them to find the optimal values of x and y

---

$$\frac{\partial \mathcal{L}}{\partial x} = 4\,x - 4\,y + \lambda = 0 \tag{1}$$

$$\frac{\partial \mathcal{L}}{\partial y} = 2\,y - 4\,x - 2\,\lambda = 0 \tag{2}$$

$$\frac{\partial \mathcal{L}}{\partial \lambda} = x - 2\,y - 2 = 0. \tag{3}$$

From which we find

$$\begin{aligned}
(1) + (2) :& \qquad -2\,y - \lambda = 0 && \qquad y = \frac{-\lambda}{2} \\
(1) + 2(2) :& \qquad -4\,x - 3\,\lambda = 0 && \qquad x = \frac{-3\,\lambda}{4} \\
(3) :& \qquad \frac{-3\,\lambda}{4} + 2\,\frac{\lambda}{2} = \frac{\lambda}{4} = 2 && \qquad \lambda = 8
\end{aligned}$$

or

$$\lambda = 8 \qquad x = -6 \qquad y = -4$$

Thus, the optimal values of x **and** y **are** $(x^*, y^*) = (-6, -4)$

(8 marks)

(e) Show that the gradient of $f(x, y)$ at the optimal point is orthogonal to the direction of the constraint

---

This tests their deep understanding of what Lagrangians do.

$$\boldsymbol{\nabla} f\big|_{x=-6,\,y=-4} = \begin{pmatrix} \frac{\partial f}{\partial x} \\ \frac{\partial f}{\partial y} \end{pmatrix}_{x=-6,\,y=-4} = \begin{pmatrix} 4\,x - 4\,y \\ 2\,y - 4\,x \end{pmatrix}_{x=-6,\,y=-4} = \begin{pmatrix} -8 \\ 16 \end{pmatrix} = -8 \begin{pmatrix} 1 \\ -2 \end{pmatrix}$$

Consider vectors from $\boldsymbol{x}^*$ to an arbitrary point $\boldsymbol{x} = (x, y)^{\mathsf{T}}$. The dot product of such as vector with ∇f is

$$(\boldsymbol{\nabla} f)^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^*) = -8(x - 2\,y - (x^* - 2\,y^*))$$

but $x^* - 2\,y^* = -6 + 8 = 2$ so that

$$(\boldsymbol{\nabla} f)^{\mathsf{T}}(\boldsymbol{x} - \boldsymbol{x}^*) = -8(x - 2\,y - 2)$$

but on the constraint $x - 2\,y - 2 = 0$ so that the gradient is perpendicular to the constraint (as has to be the case).

(6 marks)

Question 4

(a) Describe the limitation of the perceptron in separating different classes of data

Start with simple test of background knowledge.

The perceptron is only capable of performing a linear separation of the data. Thus, if the data sets are not linearly separable then a perceptron will fail to correctly classify the data even if it is separable.

(3 marks)

(b) Describe how the MLP, RBF and SVM each overcome the limitation of the perceptron.

Test broad range of knowledge

MLP **The MLP combines many perceptrons to divide up the feature space into more complex regions. Each perceptron in the first hidden layer finds a separating hyperplane. These are combined by perceptrons further towards the output layer to build up more complicated separating regions.**

RBF **Radial basis functions divides up the feature space at the input layer according to their distance from a set of centres. These are then combined by a perceptron in the output layer.**

SVM **In support vector machines the feature space is projected into a high dimensional “transformed feature space” using the kernel trick. A maximal margin hyperplane is found in this much high dimensional space where it is much easier to separate the data. If the problem is still not separable then slack variables can be used to allow some points to be misclassified.**

(6 marks)

(c) Given a learning machine $F(\boldsymbol{x};\boldsymbol{w})$ learning a function $f(\boldsymbol{x})$, the expected generalisation performance is equal to

$$E_g(\boldsymbol{w}) = \left\langle \Big(f(\boldsymbol{x}) - F(\boldsymbol{x};\boldsymbol{w}) \Big)^2 \right\rangle_{\boldsymbol{x}}$$

where $\langle \cdots \rangle_x$ denotes the average with respect to all possible inputs. The expected generalisation performance over all data sets is

$$E_g = \langle E_g(\boldsymbol{w}) \rangle_{\mathcal{D}}$$

where $\langle \cdots \rangle_{\mathcal{D}}$ denotes the average over all training sets. Expand around the expected output of the learning machine averaged over all data sets $\langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}}$ to obtain an expression for the generalisation performance in terms of the bias and variance. Explain what these terms are and describe the Bias-Variance Dilemma.

Test knowledge of standard derivation and its meaning

Expanding around $\langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}}$

$$\begin{aligned}
E_g &= \left\langle \Big(f(x) - F(\boldsymbol{x};\boldsymbol{w}) \Big)^2 \right\rangle_{\mathcal{D},\boldsymbol{x}} \\
&= \left\langle \Big(f(x) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} + \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} - F(\boldsymbol{x};\boldsymbol{w}) \Big)^2 \right\rangle_{\mathcal{D},\boldsymbol{x}} \\
&= \left\langle \Big(f(x) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big)^2 \right\rangle_{\boldsymbol{x}} + \left\langle \Big(F(\boldsymbol{x};\boldsymbol{w}) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big)^2 \right\rangle_{\mathcal{D},\boldsymbol{x}} \\
&\qquad + 2 \left\langle \Big(F(\boldsymbol{x};\boldsymbol{w}) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big) \Big(f(x) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big) \right\rangle_{\mathcal{D},\boldsymbol{x}} \\
&= B + V + C
\end{aligned}$$

The cross term

$$\begin{aligned}
C &= 2 \left\langle \Big(F(\boldsymbol{x};\boldsymbol{w}) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big) \Big(f(x) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big) \right\rangle_{\mathcal{D},\boldsymbol{x}} \\
&= 2 \left\langle \Big(\langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big) \Big(f(x) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big) \right\rangle_{\mathcal{D}} = 0
\end{aligned}$$

while the bias is given by

$$B = \left\langle \Big(f(x) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big)^2 \right\rangle_{\boldsymbol{x}}.$$

This is a measure of the error of the average machine. The variance is given by

$$V = \left\langle \Big(F(\boldsymbol{x};\boldsymbol{w}) - \langle F(\boldsymbol{x};\boldsymbol{w}) \rangle_{\mathcal{D}} \Big)^2 \right\rangle_{\mathcal{D},\boldsymbol{x}}$$

which measures the fluctuation in the output of the learning machines.

Thus the expected generalisation performance can be seen as the sum of two terms. The bias which is a consequence of the having too simple a machine and the variance which is a consequence of the parameters of the machine being too sensitive to the data during learning.

The Bias-Variance dilemma describes the difficulty of choosing the complexity of the machine. If the machine is too simple then it will have a high bias, but usually a low variance. If it is too complicated then it might have a low bias, but is likely to have a high variance.

(15 marks)

(d) Explain the different approaches to solve the Bias-Variance dilemma taken by MLPs and SVMs

Test ability to integrate knowledge from different parts of the course.

In the case of MLPs the Bias-Variance dilemma can be addressed in different ways. Often the complexity of the machine (e.g. the number of hidden nodes) is carefully chosen. An alternative is early stopping where the learning is stopped before the machine over-fits the data. Another approach is to use a regulariser such as weight decay with an adjustable parameter to find the best compromise between bias and variance. In all these mechanisms the generalisation performance is usually measured on a validation set to determine the choice of which machine to use.

In support vector machines a complicated machine is used which is capable of separating (almost) all the training data. Thus the bias is small. However, by choosing the maximal-margin separating plane the SVM effectively selects the simplest machine that does the job. This is an example of structural risk minimisation.

(9 marks)

END OF PAPER