*ANSWERS FOR SECOND TEST*

STA 414/2104 — Second Test — 2007-03-13

No books, notes, or calculators are allowed. Time is 50 minutes.

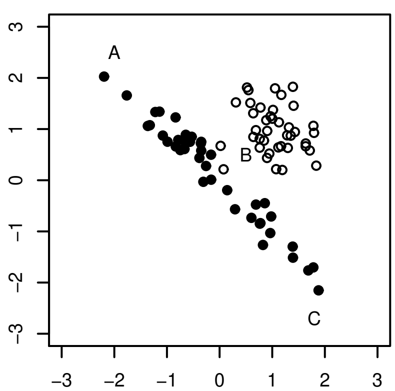
| | |
|---|---|
| 1/36 | |
| 2/30 | |
| 3/34 | |
| T/100 | |

Recall that linear and quadratic discriminants can be derived from Gaussian models for the distribution of the inputs within each class. Using such models, we can find the probability of class $k$ given values for the inputs, $x$, in a case by Bayes' Rule:

$$P(C_k|x) \quad = \quad \frac{P(x|C_k)P(C_k)}{\sum_j P(x|C_j)P(C_j)}$$

Linear discriminants, in which a case is classified according to the value of $w^T x + w_0$, for some vector $w$ and scalar $w_0$, arise when we assume that $P(x|C_k)$ is Gaussian, with the same covariance matrix for all classes, $k$. Quadratic discriminants arise when we allowed different covariance matrices for different classes. In both cases, we find the means and covariance matrices by maximum likelihood.

**Question 1:** [ 36 marks ] Each of the two scatterplots below shows training cases for a classification problem with two inputs and two classes, with the class of the case indicated by the colour of dot (black is class 1). Three test cases with inputs $x_A$, $x_B$, and $x_C$, are also indicated, by the letters A, B, and C. For each scatterplot, write down the approximate probability of class 1 for each of the three test cases, if the probabilities are found assuming that the covariance matrix is the same for both classes (as with linear discriminants) and if the probabilities are found assuming that the covariance matrices for the two classes may be different (as with quadratic discriminants). You should give rough approximations for these probabilities — either "near 0", "near 1", or "near 1/2".



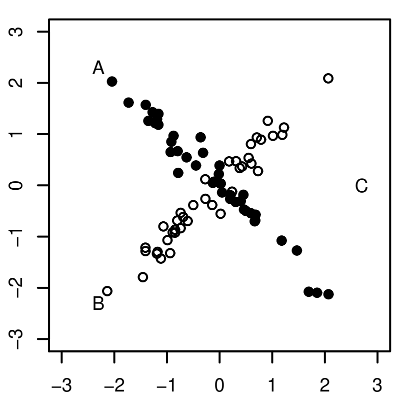

Linear discriminant:

$$P(C_1|x_A) = 1/2 \quad P(C_1|x_B) = 1/2 \quad P(C_1|x_C) = 1$$

*With the covariance for the two classes constrained to be the same, the classification boundary will be a straight line going approximately through points* A *and* B.

Quadratic discriminant:

$$P(C_1|x_A) = 1 \quad P(C_1|x_B) = 1/2 \quad P(C_1|x_C) = 1$$




Linear discriminant:

$$P(C_1|x_A) = 1/2 \quad P(C_1|x_B) = 1/2 \quad P(C_1|x_C) = 1/2$$

*Class means are nearly the same, covariances constrained to be the same, so class distributions are nearly identical, and all class probabilities are near 1/2.*

Quadratic discriminant:

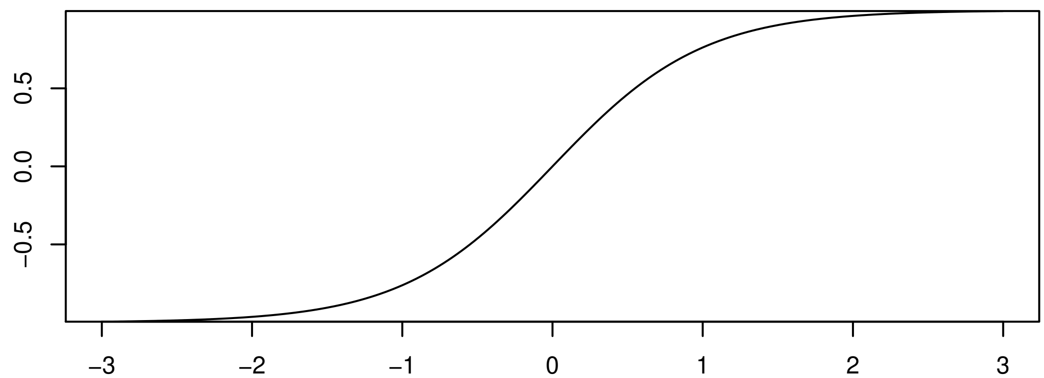
$$P(C_1|x_A) = 1 \quad P(C_1|x_B) = 0 \quad P(C_1|x_C) = 1/2$$

Recall that a multilayer perceptron network with one layer of $M$ hidden units computes an output, $y$, from the input vector, $x$, for a case, with the function computed depending on the settings of the network parameters, $w$, as follows:

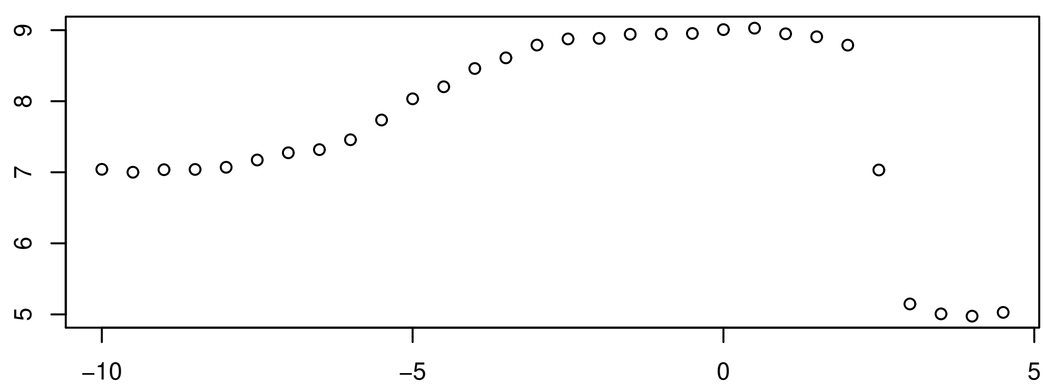
$$\begin{aligned} y &= f(s) \\ s &= w_0^{(2)} + \sum_{k=1}^{M} w_k^{(2)} z_k \\ z_k &= h(a_k), \quad \text{for } k = 1, \ldots, M \\ a_k &= w_{k0}^{(1)} + \sum_{j=1}^{D} w_{kj}^{(1)} x_j, \quad \text{for } k = 1, \ldots, M \end{aligned}$$

**Question 2:** [ 30 marks ]

For this question, suppose that $f(s) = s$ and that $h(a) = \tanh(a)$, where tanh is the hyperbolic tangent function, $\tanh(a) = (e^a - e^{-a}) / (e^a + e^{-a})$. Here is a plot of the tanh function:




Suppose that we use such a multilayer perceptron network with $M = 2$ hidden units to fit the 30 training cases plotted below:




The plot shows the one-dimensional input, $x$, on the horizontal axis and the target value, $t$, on the vertical axis. Suppose we set the network parameters, $w$, to minimize $E(w) = (1/2) \sum_{i=1}^{N} (t^{(i)} - y^{(i)})^2$, where $t^{(i)}$ and $y^{(i)}$ are the values of the target variable and the network output in training case $i$. List on the next page the approximate values that will be found for the network parameters. Rough approximations found by eye are OK. There is more than one correct solution.

*One possible solution (numbers are approximate). Other solutions can be obtained by permuting the hidden units or negating all weights involving a hidden unit.*

$w_0^{(2)} = 6$ *Adjusted for right value at $x = -10$, given other weights*

$w_1^{(2)} = 1$ *First bump up is of height 2, same as tanh*

$w_2^{(2)} = -2$ *First bump down is of height 4, twice that of tanh*

$w_{10}^{(1)} = 2.5$ *This and next adjusted so first bump is centred at $x = -5$*

$w_{11}^{(1)} = 0.5$

$w_{20}^{(1)} = -5$ *This and next adjusted so second bump is centred at $x = 2.5$*

$w_{21}^{(1)} = 2$

**Question 3:** [ 34 marks ] For this question, suppose that there are $D = 2$ inputs, that there is $M = 1$ hidden unit, that $f(s) = s$, and that $h(a) = a^{1/3}$. Suppose also that we have two training cases, for which the values of the input and target variables are as follows:

| $x_1$ | $x_2$ | $t$ |
|---|---|---|
| 1 | 0 | 3 |
| 2 | 3 | 2 |

Consider a value for $w$ in which the parameters are as follows:

$$w_0^{(2)} = -1, \qquad w_1^{(2)} = 2, \qquad w_{10}^{(1)} = 0, \qquad w_{11}^{(1)} = 1, \qquad w_{12}^{(1)} = 2$$

Compute the gradient of $E(w) = (1/2)\sum_{i=1}^{N}(t^{(i)} - y^{(i)})^2$ at this value of $w$ — ie, find the partial derivatives of $E(w)$ with respect to each component of $w$. Actual numerical values are required.

*Here is what we get for each training case by forward propagation to find the values of hidden units and outputs:*

| $x_1$ | $x_2$ | $t$ | $a$ | $z$ | $s$ | $y$ |
|---|---|---|---|---|---|---|
| 1 | 0 | 3 | 1 | 1 | 1 | 1 |
| 2 | 3 | 2 | 8 | 2 | 3 | 3 |

*Here is what we get by backpropagation to find the derivatives of* $E_i(w) = (1/2)(t^{(i)} - y^{(i)})^2$*, for* $i = 1, 2$*:*

| $\partial E_i/\partial s$ | $\partial E_i/\partial z$ | $\partial E_i/\partial a$ |
|---|---|---|
| $-2$ | $-4$ | $-4/3$ |
| 1 | 2 | $1/6$ |

*From these forward and backward values, we can get the derivaties of* $E(w)$ *with respect to each weight by summing the contributions from the two training cases, with the following results:*

$$\begin{aligned}
\partial E/\partial w_0^{(2)} &= (-2) + (1) = -1 \\
\partial E/\partial w_1^{(2)} &= (-2)(1) + (1)(2) = 0 \\
\partial E/\partial w_{10}^{(1)} &= (-4/3) + (1/6) = -7/6 \\
\partial E/\partial w_{11}^{(1)} &= (-4/3)(1) + (1/6)(2) = -1 \\
\partial E/\partial w_{12}^{(1)} &= (-4/3)(0) + (1/6)(3) = 1/2
\end{aligned}$$