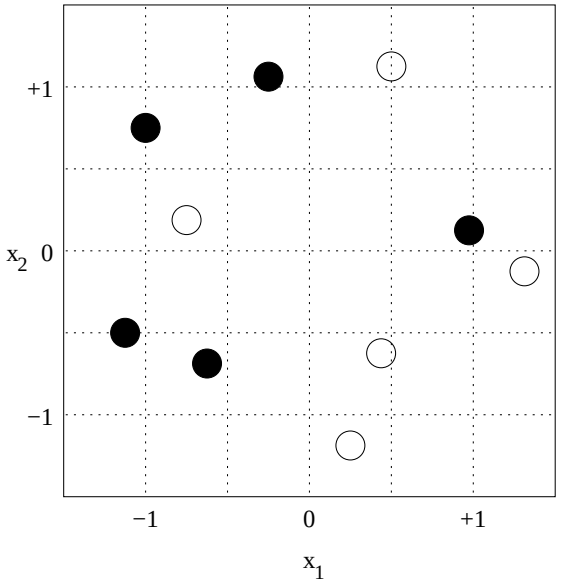
STA 414/2104 — Answers to First Test — 2012-02-09

Question 1: [30 marks total] Here is a plot of 10 training cases for a binary classification problem with two input variables, x_1 and x_2, with points in class 0 in white and points in class 1 in black:



We wish to compare three variations on the K-nearest-neighbor method for this problem, using 10-fold cross validation (ie, we leave out each training case in turn and try to predict it from the other nine). We use the fraction of cases that are misclassified as the error measure. We set $K = 1$ in all methods, so we just predict the class in a test case from the class of its nearest neighbor.

A) [8 marks] The first method looks only at x_1, so the distance between cases with input vectors x and x' is $|x_1 - x'_1|$. What is the cross-validation error for this method?

From left to right, the left out points are classified correctly (Y) or not (N) as follows:

```
Y Y N N Y Y Y Y N N
```

So the cross-validation assessment of the error rate is 4/10.

B) [8 marks] The second method looks only at x_2, so the distance between cases with input vectors x and x' is $|x_2 - x'_2|$. What is the cross-validation error for this method?

From top to bottom, the left out points are classified correctly (Y) or not (N) as follows:

```
N N Y N N N N N N N
```

So the cross-validation assessment of the error rate is 9/10.

C) [8 marks] The third method looks at both inputs, and uses Euclidean distance, so the distance between cases with input vectors x and x' is $\sqrt{(x_1 - x'_1)^2 + (x_2 - x'_2)^2}$. What is the cross-validation error for this method?

The cross-validation assessment of the error rate is 6/10.

D) [6 marks] If we use the method (from among these three) that is best according to 10-fold cross-validation, what will be the predicted class for a test case with inputs $x = (-0.25, 0.25)$?

We classify the test point based only on x_1, since that worked best in the cross-validation assessment. This leads to the test point being classified as class 1 (black).

Question 2: [36 marks total] Consider a binary classification problem in which the probability that the class, y, of an item is 1 depends on a single real-valued input, x. We use the following model for this class probability, with an unknown parameter ϕ:

$$P(y = 1 \,|\, x, \phi) \;\; = \;\; \begin{cases} 1/2 & \text{if } x \leq \phi \\ 1 & \text{if } x > \phi \end{cases}$$

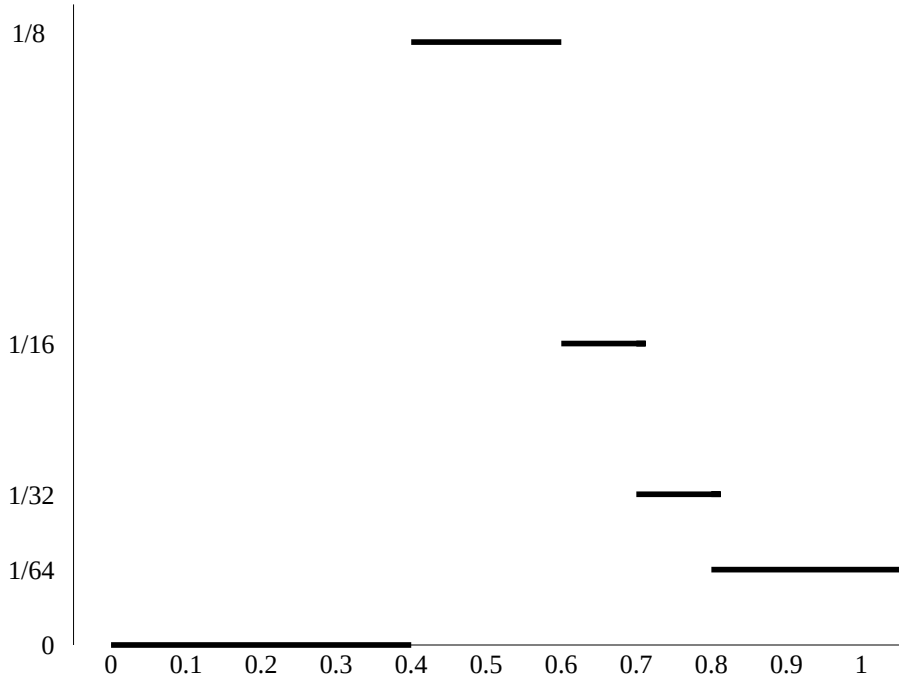
We have a training set consisting of the following six (x, y) pairs:

$$(0.1,\, 0),\;\; (0.3,\, 1),\;\; (0.4,\, 0),\;\; (0.6,\, 1),\;\; (0.7,\, 1),\;\; (0.8,\, 1)$$

A) [10 marks] Draw a graph of the likelihood function for ϕ based on the six training cases above.

The likelihood is the probability of the observed classes as a function of ϕ, with the x values taken as given. Though I forgot to say, I meant for you to assume that the classes were independent given ϕ and x. (It seems that nobody assumed otherwise.) So the probability of the data is just the product of the probabilities for the six observed classes, which are either 0, 1, or 1/2, depending for each case on y and whether or not x is greater than ϕ.

This gives the following plot of the likelihood function:



B) [8 marks] Compute the marginal likelihood for this model with this data (ie, the prior probability of the observed training data with this model and prior distribution).

As clarified in class, this question should be answered assuming the same prior as in part (C).

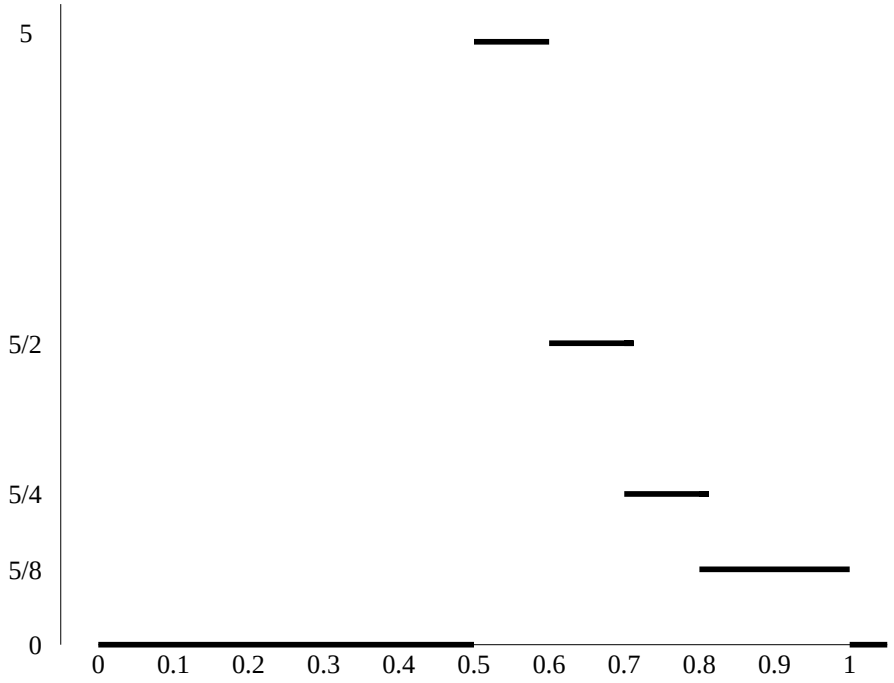
Since the prior density is zero outside the interval $[0.5, 1]$, and the prior density is 2 within this interval, the marginal likelihood is the integral over the interval $[0.5, 1]$ of 2 times the likelihood function above. This is equal to

$$2 \;\times\; (0.1/8 + 0.1/16 + 0.1/32 + 0.2/64) \;\; = \;\; 2 \;\times\; 0.8\,/\,32 \;\; = \;\; 1/20$$

C) [9 marks] Find the posterior distribution of ϕ given the six training cases above, assuming that the prior distribution of ϕ is uniform on the interval [0.5, 1]. Display this posterior distribution by drawing a graph of its probability density function.

The posterior density is zero where the prior is zero, outside the interval $[0.5, 1]$. *Within this interval, the posterior density is equal to the likelihood, times the prior density of 2, divided by the marginal likelihood of 1/20.*

This gives the following plot of the posterior density:



D) [9 marks] Find the predictive probability that $y = 1$ for three test cases in which x has the values listed below, based on the posterior distribution you found in part (C).

$x = 0.2$:

All values of ϕ *with non-zero posterior density predict that a case with* $x = 0.2$ *will have* $y = 1$ *with probability 1/2. So the predictive probability that* $y = 1$ *is 1/2.*

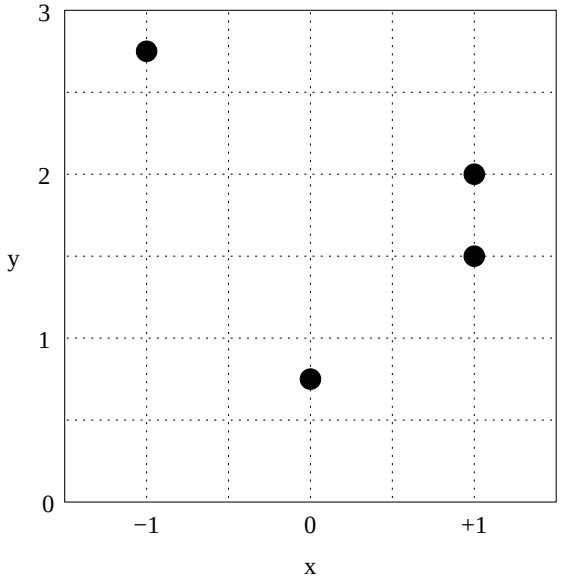
$x = 0.6$:

The posterior probability that ϕ *is less than 0.6 is* $5 \times 0.1 = 0.5$, *so the predictive probability of* $y = 1$ *when* $x = 0.6$ *is* $0.5 \times 1 + (1-0.5) \times (1/2) = 0.75$.

$x = 0.7$:

The posterior probability that ϕ *is less than 0.7 is* $5 \times 0.1 + (5/2) \times 0.1 = 0.75$, *so the predictive probability of* $y = 1$ *when* $x = 0.6$ *is* $0.75 \times 1 + (1-0.75) \times (1/2) = 0.875$.

Question 3: [34 marks total] Consider a linear basis function model for a regression problem with response y and a single scalar input, x, in which the basis functions are $\phi_0(x) = 1$, $\phi_1(x) = x$, and $\phi_2(x) = |x|$. Below is a plot of four training cases to be fit with this model:



A) [15 marks] Suppose we fit this linear basis function model by least squares. What will be the estimated coefficients for the three basis functions, $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$?

The function fit will have the form $\beta_0 + \beta_1 x + \beta_2|x|$. *This function is a straight line for* $x < 0$ *and a straight line with possibly different slope for* $x > 0$, *with the lines joining at* $x = 0$. *We can therefore choose* β_0, β_1, *and* β_2 *to pass exactly through the points at* $x = -1$ *and* $x = 0$, *and through the midpoint of the two points at* $x = +1$, *which is the best we can do to minimize squared error.*

This leads to $\widehat{\beta}_0 = 0.75$, *so that the point at* $x = 0$ *is fit exactly, to the constraint that* $\widehat{\beta}_1 + \widehat{\beta}_2 = 1$, *so that the line for* $x > 0$ *has slope 1, and to the constraint that* $\widehat{\beta}_1 - \widehat{\beta}_2 = -2$, *so that the line for* $x < 0$ *has slope* -2. *Solving these equations, we get that* $\widehat{\beta}_1 = -1/2$ *and* $\widehat{\beta}_2 = 3/2$.

B) [11 marks] Suppose we fit this linear basis function model by penalized least squares, with a penalty of $\lambda|\beta_1|$ (note that the penalty does not depend on β_0 and β_2). What will be the estimated coefficients for the three basis functions, $\widehat{\beta}_0$, $\widehat{\beta}_1$, and $\widehat{\beta}_2$ in the limit as λ goes to infinity?

An infinite penalty on β_1 *will force it to be zero, so the function will have the form* $\beta_0 + \beta_2|x|$.

Fitting this to the given data is the same as fitting to the data with the point at $x = -1$ *moved to be at* $x = +1$. *There will then be three points at* $x = +1$, *with values 2.75, 2, and 1.5. The mean of these points 6.25/3. The only other* x *point with data is* $x = 0$, *where* $y = 0.75$. *We can choose* β_0 *and* β_2 *so that the line passes exactly through* $y = 0.75$ *at* $x = 0$ *and* $y = 6.25/3$ *at* $x = +1$, *which is the best we can do to minimize squared error. This is achieved when* $\widehat{\beta}_0 = 0.75$ *and* $\widehat{\beta}_2 = 6.25/3 - 0.75 = 4/3$.

C) [8 marks] Suppose we use the form of the penalty as in part (B), but with $\lambda = 1$. Will the penalized least squares estimate for β_1 be exactly zero? Show why or why not.

The estimate for β_1 will not be exactly zero.

One way to see this is to compare the squared error plus penalty (with $\lambda = 1$) when β_1 is forced to zero and the squared error plus penalty (with $\lambda = 1$) when all coefficients are estimated without a penalty. It turns out that the latter is smaller, so the penalized least squares estimate with $\lambda = 1$ can't have $\widehat{\beta}_1 = 0$.

Here are the details of this calculation.

The best coefficients with $\widehat{\beta}_1 = 0$ were found in part (B). With these coefficients, the squared error is

$$\begin{aligned} & 0^2 \ + \ 0^2 \ + \ (2.75 - 6.25/3)^2 \ + \ (2 - 6.25/3)^2 \ + \ (1.5 - 6.25/3)^2 \\ & = \ (1/9) \ \times \ ((8.25 - 6.25)^2 \ + \ (6 - 6.25)^2 \ + \ (4.5 - 6.25)^2) \\ & = \ (1/9) \ \times \ (4 \ + \ 1/16 \ + \ 49/16) \ = \ 114/144 \end{aligned}$$

Since $\widehat{\beta}_1 = 0$, the penalty is zero.

The best coefficients with no penalty were found in part (A). With these coefficients, the squared error is

$$0^2 \ + \ 0^2 \ + \ (1/4)^2 \ + \ (1/4)^2 \ = \ 1/8$$

The penalty is $|-1/2| \ = \ 1/2$. The squared error plus penalty is therefore $5/8$, which is less than $114/144$.

Another way to answer this question is to compute the derivative with respect to β_1 of the squared error at the best estimates with $\beta_1 = 0$ that were found in part (B), which isn't too hard. The estimate for β_1 will be zero if this derivative is smaller in absolute value than λ, but it's not, when $\lambda = 1$.