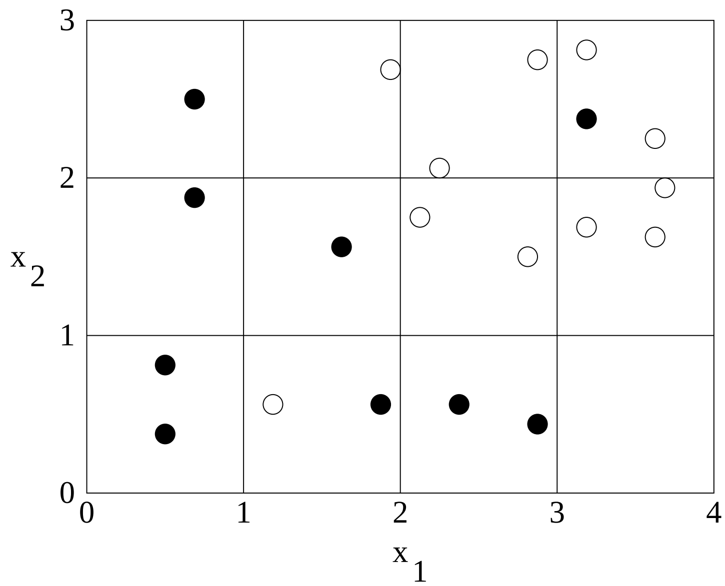
STA 414/2104, Spring 2013, Practice Problem Set #1

Note: these problems are not for credit, and not to be handed in

Question 1: Consider a classification problem in which there are two real-valued inputs, x_1 and x_2, and a binary (0/1) target (class) variable, y. There are 20 training cases, plotted below. Cases where $y = 1$ are plotted as black dots, cases where $y = 0$ as white dots, with the location of the dot giving the inputs, x_1 and x_2, for that training case.



A) Estimate the error rate of the one-nearest-neighbor (1-NN) classifier for this problem using leave-one-out cross validation. (That is, using S-fold cross validation with S equal to the number of training cases, in which each training case is predicted using all the other training cases.)

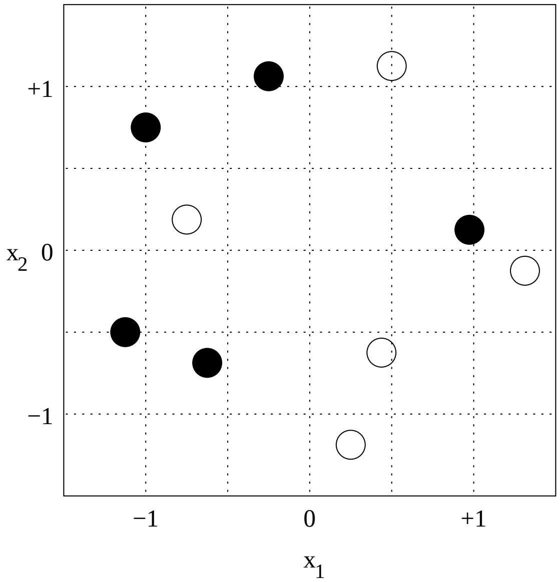
B) Suppose we use the three-nearest-neightbor (3-NN) method to estimate the probability that a test case is in class 1. For test cases with each of the following sets of input values, find the estimated probability of class 1.

$x_1 = 1, \ x_2 = 1$

$x_1 = 2, \ x_2 = 2$

$x_1 = 3, \ x_2 = 0$

Question 2: 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) 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?

B) 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?

C) 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?

D) 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 $x = (-0.25,\ 0.25)$?

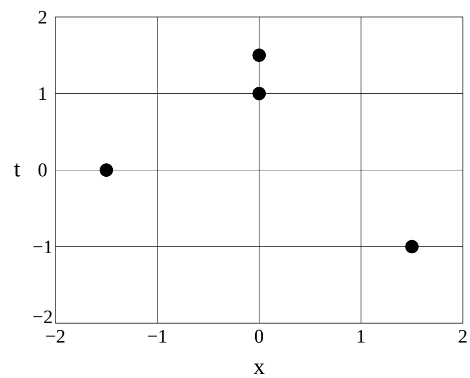
Question 3: Consider a linear basis function regression model, with one input and the following three basis functions:

$$\begin{aligned}\phi_0(x) &= 1\\ \phi_1(x) &= x\\ \phi_2(x) &= \begin{cases} 1-x^2 & \text{if } |x|<1\\ 0 & \text{if } |x|\geq 1\end{cases}\end{aligned}$$

The model for the target variable, y, is that $P(y\,|\,x,\beta) = N(y\,|\,f(x,\beta),1)$, where

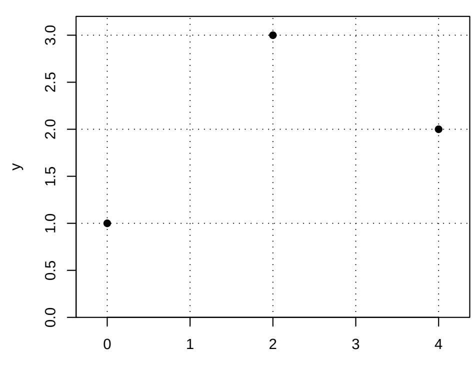
$$f(x,\beta) \;=\; \sum_{j=0}^{m-1} \beta_j \phi_j(x)$$

Suppose we have four data points, as plotted below:



What is the maximum likelihood (least squares) estimate for the parameters β_0, β_1, and β_2? Elaborate calculations should not be necessary.

Question 4: Below is a plot of a dataset of $n=3$ observations of (x_i, y_i) pairs:

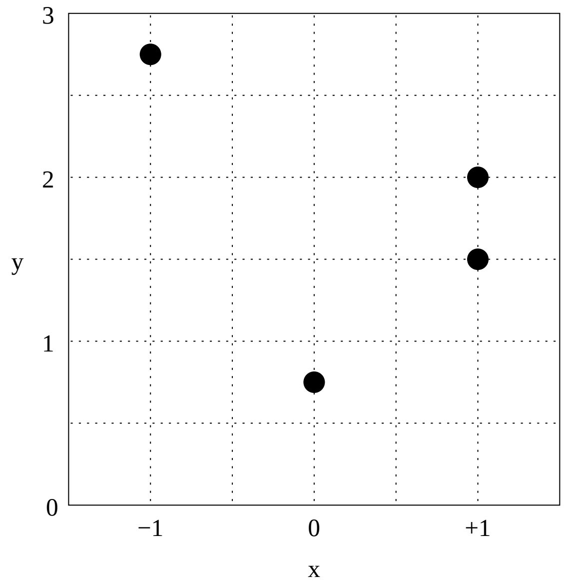


In other words, the data points are $(0,1)$, $(2,3)$, $(4,2)$.

Suppose we model this data with a linear basis function model with $m=2$ basis functions given by $\phi_0(x)=1$ and $\phi_1(x)=x$. We use a quadratic penalty of the form $\lambda\beta_1^2$, which penalizes only the regression coefficient for $\phi_1(x)$, not that for $\phi_0(x)$.

Suppose we use squared error from three-fold cross-validation (ie, with each validation set having only one case) to choose the value of λ. Suppose we consider only two values for λ — one very close to zero, and one very large. For the data above, will we choose λ near zero, or λ that is very big?

Question 5: 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) 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 $\widehat{\beta}_2$?

B) 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?

C) 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.

Question 6: Suppose that we observe a binary (0/1) variable, Y_1. We do not know the probability, θ, that Y_1 will be 1, but we have a prior distribution for θ, that has the following density function on the interval $(0, 1)$:

$$P(\theta) \;=\; 12\left(\theta - \frac{1}{2}\right)^2$$

A) Find as simple a formula as you can for the density function of the posterior distribution of θ given that we observe $Y_1 = 1$. Your formula should give the correcty normalized density.

B) Suppose that Y_2 is a future observation, that is independent of Y_1 given θ. Find the predictive probability that $Y_2 = 1$ given that $Y_1 = 1$ — ie, find $P(Y_2 = 1 \,|\, Y_1 = 1)$.

Question 7: Let X_1, X_2, X_3, ... for a sequence of binary (0/1) random variables. Given a value for θ, these random variables are independent, and $P(X_i = 1) = \theta$ for all i. Suppose that we are sure that θ is at least $1/2$, and that our prior distribution for θ for values $1/2$ and above is uniform on the interval $[1/2, 1]$. We have observed that $X_1 = 0$, but don't know the values of any other X_i.

A) Write down the likelihood function for θ, based on the observation $X_1 = 0$.

B) Find an expression for the posterior probability density function of θ given $X_1 = 0$, simplified as much as possible, with the correct normalizing constant included.

C) Find the predictive probability that $X_2 = 1$ given that $X_1 = 0$.

D) Find the probability that $X_2 = X_3$ given that $X_1 = 0$.

Question 8: 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, with the classes for different cases being independent, given a parameter ϕ and x. We use the following model for this class probability in terms of the unknown parameter ϕ:

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

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) Draw a graph of the likelihood function for ϕ based on the six training cases above.

B) 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), assuming that the prior distribution of ϕ is uniform on the interval [0.5, 1]

C) Find the posterior distribution of ϕ given the six training cases above, and the prior from part (B) Display this posterior distribution by drawing a graph of its probability density function.

D) Find the predictive probability that $y = 1$ for each of three test cases in which x has the values 0.2, 0.6, and 0.7, based on the posterior distribution you found in part (C).

Question 9: Answer the following questions about Bayesian inference for linear basis function models. Recall that if the noise variance is σ^2, and the prior distribution for β is Gaussian with mean zero and covariance matrix S_0, the posterior distribution for β is Gaussian with mean m_n and covariance matrix S_n that can be written as follows:

$$S_n \;=\; \left[S_0^{-1} \,+\, (1/\sigma^2)\Phi^T\Phi\right]^{-1}, \qquad m_n \;=\; S_n\Phi^T y/\sigma^2$$

and the log of the marginal likelihood for the model is

$$-\frac{n}{2}\log(2\pi) \;-\; \frac{n}{2}\log(\sigma^2) \;-\; \frac{1}{2}\log\left(\frac{|S_0|}{|S_n|}\right) \;-\; \frac{1}{2}||y-\Phi m_n)||^2/\sigma^2 \;-\; \frac{1}{2}m_n^T S_0^{-1} m_n$$

For the questions below, assume that $S_0 = \omega^2 I$, for some positive ω.

A) Suppose we set the noise variance, σ^2, to be bigger and bigger, while fixing other aspects of the model. What will be the limiting values of the the posterior mean and covariance matrix?

B) Suppose we set ω^2, the prior variance of the β_j, to be bigger and bigger, while fixing other aspects of the model. What will be the limiting values of the the posterior mean, m_n, and covariance matrix, S_n?

C) Suppose we set ω^2 to be bigger and bigger while fixing other aspects of the model. What will be the limiting value of the marginal likelihood?

D) Suppose there is only one input (so x is a scalar), and the basis functions are $\phi_j(x) = x^j$, for $j = 0, \ldots, m-1$. The Bayesian mean prediction for the value of y in a test case with input x is found by integrating the prediction based on β (ie, the expected value of y given x and β) with respect to the posterior distribution of β. Will this final mean prediction be a polynomial function of x?