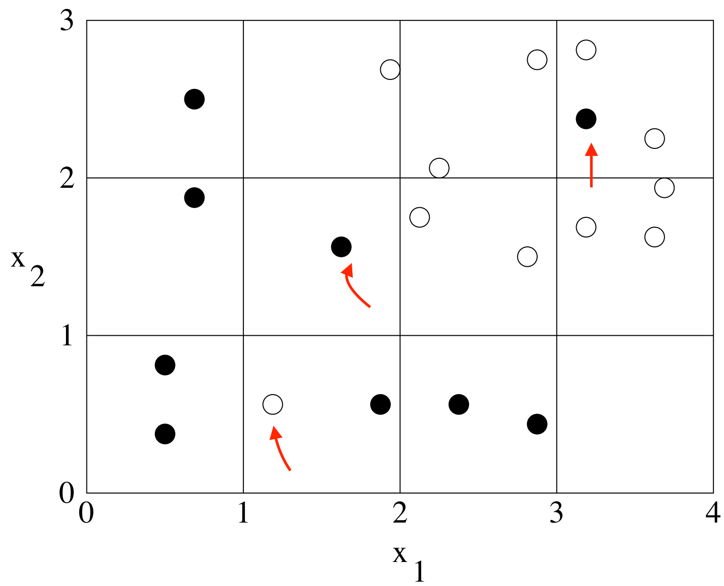
Answers to Part C - thanks to Radford Neal

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.) i,e, 20

Three of the cases will be mis-classified based on the others, so the estimated error rate is $3/20$.

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.

$x_1 = 1, \ x_2 = 1$

The answer is $2/3$.

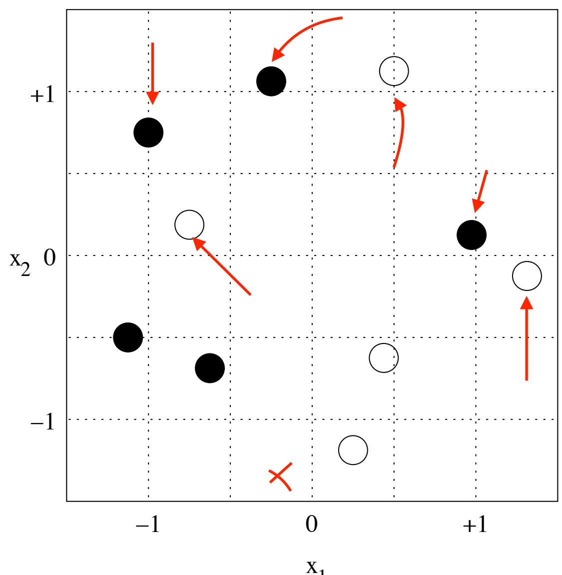
$x_1 = 2, \ x_2 = 2$

The answer is $1/3$.

$x_1 = 3, \ x_2 = 0$

The answer is 1.

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:



error is fraction of cases (n=10) that are missclassified

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?

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) 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` N N Y N N N N N N N

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

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?

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

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 $x = (-0.25, 0.25)$?

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).

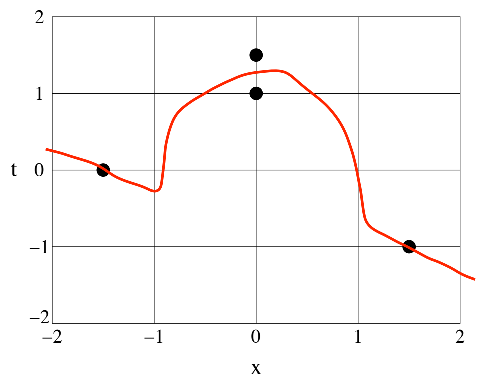
Question 3: Consider a linear basis function regression model, with one input and the following three basis functions:

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

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

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

Suppose we have four data points, as plotted below:

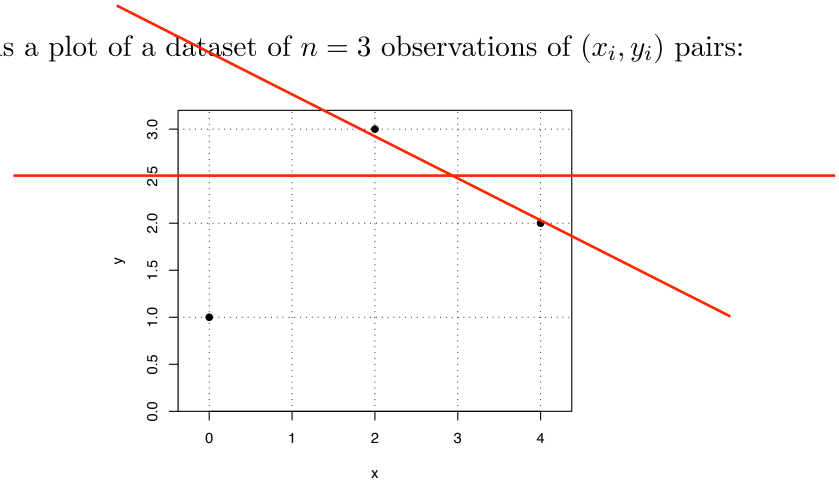


What is the maximum likelihood (least squares) estimate for the parameters β_0, β_1, and β_2? Elaborate calculations should not be necessary.

Note that $\phi_2(x)$ is zero for the data points where $x = -1.5$ and $x = +1.5$. So the value of β_2 will not affect the value of $f(x, w)$ at these points. It can therefore be used to fit the two data points at $x = 0$ (where $\phi(x) = 1$) as well as possible, regardless of what β_0 and β_1 are. This in turn means that we can use β_0 and β_1 to fit the two data points at $x = -1.5$ and $x = +1.5$. Looking at the line joining these two points, we see that the intercept is $-1/2$ and the slope is $-1/3$. We will therefore fit these points exactly if we use $\beta_0 = -1/2$ and $\beta_1 = -1/3$. Choosing $\beta_2 = 1.75$ will then lead to $f(0, w) = 1.25$, which is the best value we can have for fitting the two data points at $x = 0$.

f(0, beta) = beta_0 + beta(1 - 0) = -0.5 + beta_2 = 1.25

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?

With one point removed, the dataset will have only two points, so with λ close to zero, the regression line will pass through these two points, whereas with λ very large, the regression line will be horizontal, at the level equal to the mean of the two responses.

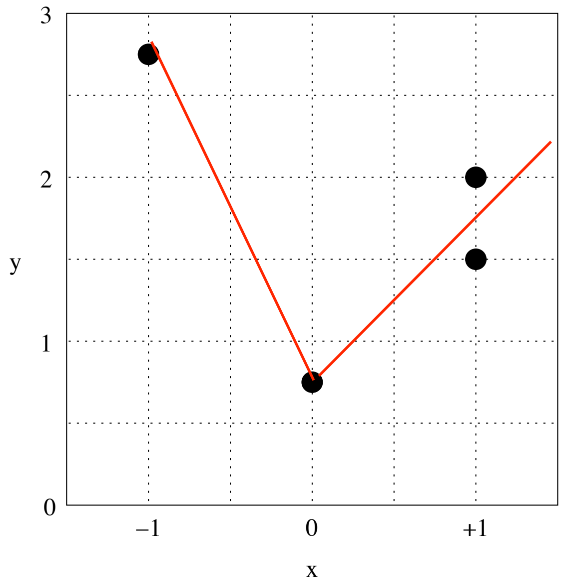
y(x,beta) = beta_0 + beta_1 x

For λ close to zero, we see that leaving out points from left to right gives squared errors of 3^2, 1.5^2, and 3^2, for a total of 20.25.

For λ very large, leaving out points from left to right gives squared errors of 1.5^2, 1.5^2, and 0, for a total of 4.5.

So based on this cross-validation assessment, we would prefer the very large value of λ.

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 $\widehat{\beta}_2$?

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) 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? beta_1 -> 0

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) 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.

infty

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

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

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 $\lambda = 1$.

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.

likelihood prior

$$\begin{aligned} P(\theta \,|\, Y_1 = 1) \;&=\; \theta \cdot 12(\theta - 1/2)^2 \Big/ \int_0^1 \theta \cdot 12(\theta - 1/2)^2 \, d\theta \\ &=\; 24\theta(\theta - 1/2)^2 \end{aligned}$$

normalizing constant

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 $P(Y_2 = 1 \,|\, Y_1 = 1)$.

$$P(Y_2 = 1 \,|\, Y_1 = 1) \;=\; \int_0^1 \theta \cdot 24\theta(\theta - 1/2)^2 \, d\theta \;=\; 4/5$$

integrate posterior given likelihood is theta again (Y_2=1)
w.r.t. parameter theta gives the predictive distribution

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 $X_1 = 0$.

$L(\theta) \;=\; P(X_1 = 0 \,|\, \theta) \;=\; 1 - \theta$

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.

uniform!

The prior density is $P(\theta) = 2$ for $\theta \in [1/2, 1]$, 0 otherwise.

The posterior density is $P(\theta \,|\, X_1 = 0) \;=\; 0$ for $\theta \notin [1/2, 1]$, and otherwise $P(\theta \,|\, X_1 = 0) \;\propto\; P(\theta)\, L(\theta) \;\propto\; 2\,(1-\theta)$. The normalizing constant can be found by evaluating $\int_{1/2}^{1} 2\,(1-\theta)\, d\theta \;=\; 1/4$, from which we find that $P(\theta \,|\, X_1 = 0) \;=\; 8\,(1-\theta)$ for $\theta \in [1/2, 1]$.

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

$P(X_2 = 1 \,|\, X_1 = 0) \;=\; \int P(X_2 = 1 \,|\, \theta)\, P(\theta \,|\, X_1 = 0)\, d\theta \;=\; = \; \int_{1/2}^{1} \theta\, 8\,(1-\theta)\, d\theta \;=\; 2/3$

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

$$
\begin{aligned}
P(X_2 = X_3 \,|\, X_1 = 0) \;&=\; \int P(X_2 = X_3 \,|\, \theta)\, P(\theta \,|\, X_1 = 0)\, d\theta \\
&\quad \text{either both 0 or both 1} \\
&=\; \int [\, P(X_2 = 0, X_3 = 0 \,|\, \theta) + P(X_2 = 1, X_3 = 1 \,|\, \theta)\,]\, P(\theta \,|\, X_1 = 0)\, d\theta \\
&=\; \int [\, P(X_2 = 0 \,|\, \theta) P(X_3 = 0 \,|\, \theta) + P(X_2 = 1 \,|\, \theta) P(X_3 = 1 \,|\, \theta)\,]\, P(\theta \,|\, X_1 = 0)\, d\theta \\
&=\; \int_{1/2}^{1} ((1-\theta)^2 + \theta^2)\, 8\,(1-\theta)\, d\theta \\
&=\; 7/12
\end{aligned}
$$

Note that X_2 and X_3 are independent given θ, but they are not independent given just X_1.

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:

implies for any phi < 0.4, likelihood for this case is 0, so overall=0

(0.1, 0), (0.3, 1), (0.4, 0), (0.6, 1), (0.7, 1), (0.8, 1)

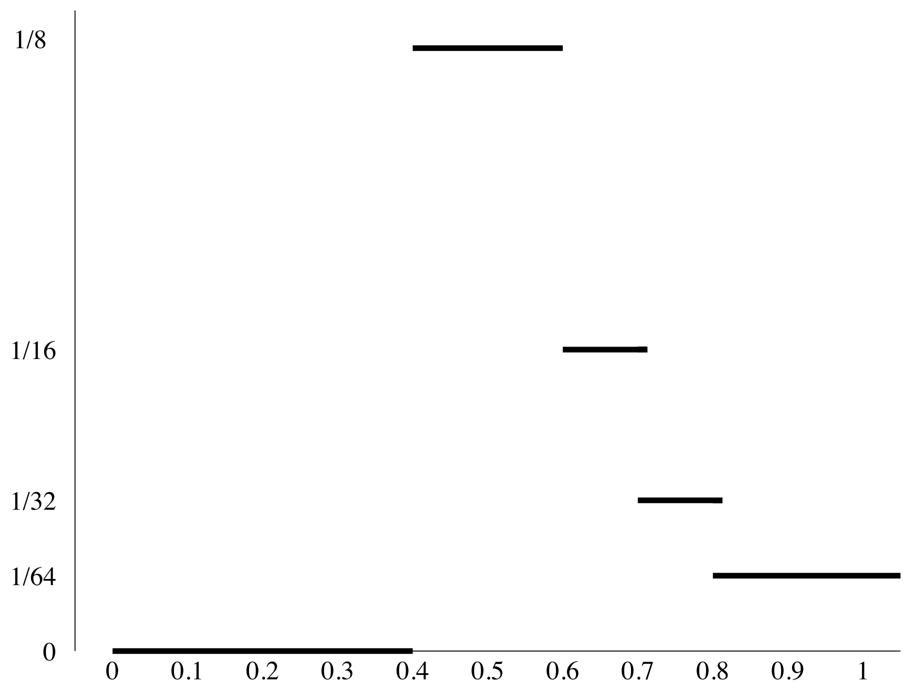
(0.4, 0.6), likelihood = 1/2 * 1/2 * 1/2 *1*1*1= 1/8

A) 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. Due to independence, 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:

p(DIphi)



p(phi)

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]

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*

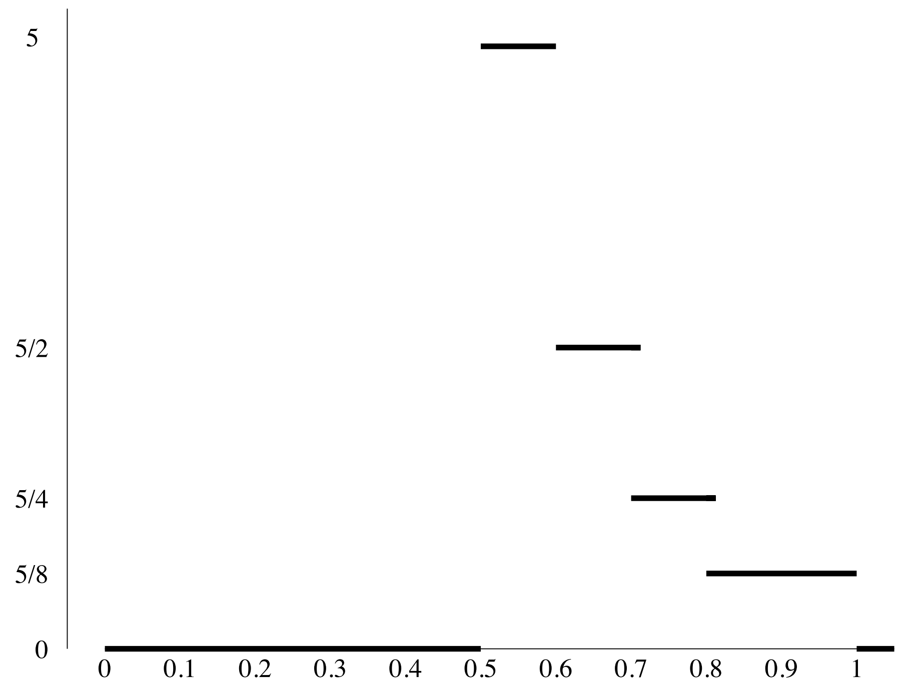
$$2 \;\times\; (0.1/8 + 0.1/16 + 0.1/32 + 0.2/64) \;=\; 2 \,\times\, 0.8\,/\,32 \;=\; 1/20$$

marginal likelihood is just integrating over likelihood function over all possible phi
i.e. the cdf over the domain for likelihood function

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.

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:



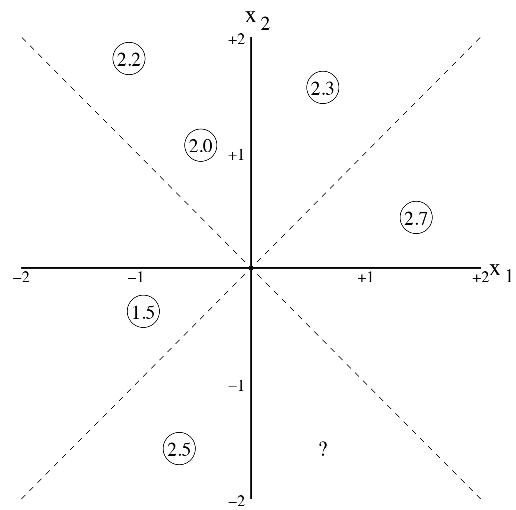
D) Find the predictive probability that $y = 1$ for each of three test cases in which x has the values 0.2, 0.6, and 0.7, based on the posterior distribution you found in part (C).

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* $y = 1$ *at that* x *is 1/2.*

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* $0.5 \times 1 \ + \ (1-0.5) \times (1/2) \ = \ 0.75$.

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* $x = 0.7$ *is* $0.75 \times 1 \ + \ (1-0.75) \times (1/2) \ = \ 0.875$.

Question 9: Below is a plot of six training cases for a regression problem with two inputs. The location of the circle for a training case gives the values of the two inputs, x_1 and x_2, for that case, and the number in the circle is the value of the response, y, for that case.



Suppose we use a linear basis function model for this data, with the following basis functions:

$$\phi_0(x) = 1$$

$$\phi_1(x) = \begin{cases} 1 & \text{if } x_1 > 0 \\ 0 & \text{if } x_1 \leq 0 \end{cases}$$

$$\phi_2(x) = \begin{cases} 1 & \text{if } x_1 + x_2 > 0 \\ 0 & \text{if } x_1 + x_2 \leq 0 \end{cases}$$

A) What will be the least squares estimates of the regression coefficients, β_0, β_1, and β_2, based on these six training cases (ignore the question mark in the lower right for now)? No elaborate matrix computations should be needed to answer this.

For the two points in the lower-left quadrant, both ϕ_1 and ϕ_2 will be zero, so for these points, y will be modeled as β_0 plus noise. The best fit possible for these points is when β_0 is their average, 2.0.

For the two points in the upper-left quadrant (both above the diagonal line), ϕ_1 will be zero and ϕ_2 will be one, so for these points, y will be modeled as $\beta_0 + \beta_2$ plus noise. The best fit possible for these points is when $\beta_0 + \beta_2$ is their average, 2.1. If we set β_0 to 2.0 in order to model the points in the lower left, we can achieve this by setting β_2 to 0.1.

For the two points in the uppper-right quadrant, both ϕ_1 and ϕ_2 will be one, so for these points, y will be modeled as $\beta_0 + \beta_1 + \beta_2$ plus noise. The best fit possible for these points is when $\beta_0 + \beta_1 + \beta_2$ is their average, 2.5. If we set β_0 to 2.0 and β_2 to 0.1 in order to model the other points, we can achieve this by setting β_1 to 0.4.

So the answer is $\hat{\beta}_0 = 2.0,\ \hat{\beta}_1 = 0.4,\ \hat{\beta}_2 = 0.1.$

B) Based on the least squares estimates for part (A), what will be the prediction for the value of the response in a test case whose x_1 and x_2 values are given by the location of the question mark in the plot above?

At this point, ϕ_1 is one and ϕ_2 is zero, so the prediction is $\hat{\beta}_0 + \hat{\beta}_1 = 2.4.$

C) For this training set, find and estimate of the average squared error of prediction using least squares estimates, by applying leave-one-out cross-validation (which is the same as six-fold cross validation here, since there are six training cases).

Reviewing the answer to (A), we see that if we leave out one point of the pair in any of the quadrants, we could fit the other point of that pair exactly, while still fitting the remaining points as well as possible. So the error in predicting each left-out point will be its difference from the other point in its pair. The answer is therefore

applying LS to 5 points, then find

$$(1^2 + 1^2 + 0.2^2 + 0.2^2 + 0.4^2 + 0.4^2) \ / \ 6 \quad = \quad 0.4$$

D) What will be the prediction for the test case at the question mark based on penalized least squares estimates, if the penalty has the form $\lambda(\beta_1^2 + \beta_2^2)$, and λ is very large?

A very large penalty for $\beta_1^2 + \beta_2^2$ will force them both to be close to zero, at which point β_0 will be set to the overall mean of the y values, which is 2.2. This will then be the prediction for the test case.

Question 10: Let $Y_1, Y_2, Y_3, \ldots$ be random quantities that are independent given a parameter θ, with each Y_t having the value 1, 2, or 3, with probabilities

$$\underbrace{P(Y_t = y \mid \theta)}_{\text{likelihood}} \;=\; \begin{cases} \theta & \text{if } y = 1 \\ 2\theta & \text{if } y = 2 \\ 1-3\theta & \text{if } y = 3 \end{cases}$$

The prior distribution for the model parameter θ is uniform on the interval $[0, 1/3]$.

For the questions below, suppose we observe that $Y_1 = 1$ and $Y_2 = 3$, but do not observe $Y_3, Y_4 \ldots$.

A) Find the marginal likelihood for this data (that is, the probability that $Y_1 = 1$ and $Y_2 = 3$, integrating over the prior distribution of θ).

$$P(Y_1 = 1,\, Y_2 = 3) \;=\; \int P(Y_1 = 1,\, Y_2 = 3 \,|\, \theta)\, P(\theta)\, d\theta \;=\; \int_0^{1/3} \theta\,(1-3\theta)\,3\, d\theta \;=\; 1/18$$

disjoint events Y_1 and Y_2

B) Find the posterior probability density function for θ (that is, the density for θ given $Y_1 = 1$ and $Y_2 = 3$), with the correct normalizing constant.

$$P(\theta \,|\, Y_1 = 1,\, Y_2 = 3) \;=\; \frac{P(Y_1 = 1,\, Y_2 = 3 \,|\, \theta)\, P(\theta)}{P(Y_1 = 1,\, Y_2 = 3)}$$

$$=\; \frac{\theta\,(1-3\theta)\,3}{1/18} \;=\; 54\,\theta\,(1-3\theta), \quad \textit{for } \theta \in (0, 1/3)$$

C) Find the predictive distribution for Y_3 given the observed data (that is, give a table of $P(Y_3 = y \,|\, Y_1 = 1,\, Y_2 = 3)$ for $y = 1, 2, 3$).

$$P(Y_3 = 1 \,|\, Y_1 = 1,\, Y_2 = 3) \;=\; \int_0^{1/3} \theta\, 54\,\theta\,(1-3\theta)\, d\theta \;=\; 1/6$$

$$P(Y_3 = 2 \,|\, Y_1 = 1,\, Y_2 = 3) \;=\; \int_0^{1/3} 2\theta\, 54\,\theta\,(1-3\theta)\, d\theta \;=\; 1/3$$

$$P(Y_3 = 1 \,|\, Y_1 = 1,\, Y_2 = 3) \;=\; \int_0^{1/3} (1-3\theta)\, 54\,\theta\,(1-3\theta)\, d\theta \;=\; 1/2$$

Question 11: 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:

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

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?

In this limit, m_n will go to zero, and S_n will go to S_0. That is, the posterior distribution will be the same as the prior distribution.

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?

In this limit, S_n will go to $\sigma^2(\Phi^T\Phi)^{-1}$ and m_n will go to $(\Phi^T\Phi)^{-1}\Phi^T y$, which is the same as the least squares (maximum likelihood) estimate.

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?

The first and second terms in the expression above for the log marginal likelihood do not depend on ω. The last term will go to zero as ω goes to infinity, since S_0^{-1} will go to zero, while (as we saw in part (B)), m_n goes to some finite limit, given by the maximum likelihood estimate. For the same reason, as ω goes to infinity, the fourth term will go to some finite limit. However, the third term, $-\log(|S_0|/|S_n|)$, will go to minus infinity as ω goes to infinity, singce $-\log(|S_0|)$ will go to minus infinity and $|S_n|$ will go to a finite limit.

The marginal likelihood will therefore go to zero as ω goes to infinity. not log marginal !

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?

Yes. The prediction for any fixed value of β will be a polynomial in x, so the expectation of this prediction with respect to the posterior distribution of β will also be a polynomial in x, since averaging any number of polynomials of some order gives another polynomial of the same order.