STA 414/2104, Spring 2012, Practice Problem Set #2, Answers (partial, more to come)

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

Question 1: Recall that a multilayer perceptron network with m hidden units using the tanh activation function computes a function defined as follows:

$$f(x,w) \;=\; w_0^{(2)} \;+\; \sum_{j=1}^{m} w_j^{(2)} \phi_j(x,w), \qquad \phi_j(x,w) \;=\; \tanh\Big(w_{0j}^{(1)} \;+\; \sum_{k=1}^{p} w_{kj}^{(1)} x_k\Big)$$

where w is the set of parameters (weights) for the network, and x is the vector of p inputs to the network.

Suppose we train such a network with $m = 1$ hidden units on the following set of $n = 4$ training cases, with $p = 1$ input, x_1, and one real-valued response, y:

x_1	y
-1	1
0	1
1	5
2	5

We use a Gaussian model for the response, in which y given x has a Gaussian distribution with mean $y(x,w)$ and variance one.

a) Suppose that we initialize the weights to $w_{01}^{(1)} = 0$, $w_{11}^{(1)} = 0$, $w_0^{(2)} = 0$, and $w_1^{(2)} = 0.1$. Define $E(w)$ to be the minus the log likelihood, dropping terms that don't depend on w, so that $E(w)$ is 1/2 times the sum of the squares of the residuals in the four training cases.

Find the gradient of $E(w)$, as would be needed to do gradient descent learning, evaluated at the initial value of w specified above. In other words, find the partial derivatives of E with respect to all the components of w, at the initial value of w.

With these initial weights, the hidden unit has the value 0, and the output of the network will also be 0, for all training cases.

We can split $E(w)$ into a sum over training cases, as $E(w) = E_1(w) + E_2(w) + E_3(w) + E_4(w)$, with $E_i(w) = (y_i - f(x_i,w))^2/2$. With the initial weights, the derivatives of each E_i with respect to the network output is $-(y_i - 0) = -y_i$. Working backwards, we see that the derivative of E_i with respect to the hidden unit value is $w_1^{(2)}(-y_i) = -0.1y_i$. Since the hidden unit input is zero for all training cases, where the derivative of tanh is one, this is also the derivative of E_i with respect to the hidden unit input.

We can use these results to find the derivatives of E_i with respect the the weights:

$$\begin{aligned}
\partial E_i/\partial w_0^{(2)} &= -y_i \\
\partial E_i/\partial w_1^{(2)} &= -y_i \times 0 \;=\; 0 \\
\partial E_i/\partial w_0^{(1)} &= -0.1 y_i \\
\partial E_i/\partial w_1^{(1)} &= -0.1 y_i x_i
\end{aligned}$$

Adding these up for all training cases, we get

$$\begin{aligned}
\partial E/\partial w_0^{(2)} &= -(1+1+5+5) &= -12 \\
\partial E/\partial w_1^{(2)} &= 0 \\
\partial E/\partial w_0^{(1)} &= -0.1(1+1+5+5) &= -1.2 \\
\partial E/\partial w_1^{(1)} &= -0.1(1(-1)+1(0)+5(1)+5(2)) &= -1.4
\end{aligned}$$

b) If gradient descent learning to minimize minus the log likelihood is done from the initial weights specified in part (a) above, what weights will the learning converge to (assuming that the learning rate used is small enough to ensure stability)? You may not be able to say exactly what the values of all the weights will be, but say as much as you can.

The network can only fit a shifted and scaled tanh function to the data. Such a function can fit this data exactly in the limit as $w_1^{(1)}$ *goes to infinity, or minus infinity, as that can turn the tanh function into a step function, which goes from 1 for* $x \leq 0$ *to 5 for* $x \geq 1$*. With any finite value for* $w_1^{(1)}$*, the best fit will be when the step occurs half-way between 0 and 1, at* $x = 1/2$*. There are two such solutions:*

$$\begin{aligned}
w_1^{(1)} &= \textit{large positive value} \\
w_0^{(1)} &= -w_1^{(1)}/2 \\
w_1^{(2)} &= 2 \\
w_0^{(2)} &= 3
\end{aligned}$$

and

$$\begin{aligned}
w_1^{(1)} &= \textit{large negative value} \\
w_0^{(1)} &= -w_1^{(1)}/2 \\
w_1^{(2)} &= -2 \\
w_0^{(2)} &= 3
\end{aligned}$$

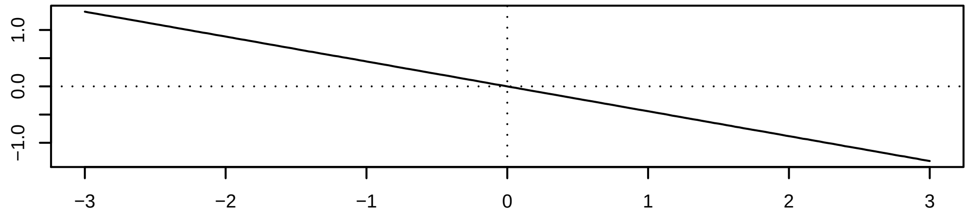
We can see from part (a) that gradient descent from the initial weights given will push the weights towards the first of these solutions, though it's possible that the value of $w_0^{(1)}$ *won't be exactly as shown above, if* $w_1^{(1)}$ *grows fast enough that the exact location of the step doesn't matter.*

c) Suppose that gradient descent learning is done from the initial weights in part (a), but with a penalty of $\lambda[w_{11}^{(1)}]^2$ added to minus the log likelihood. If λ is a small positive number, what will the learning converge to (assuming that the learning rate used is small enough to ensure stability)? You may not be able to say exactly what the values of all the weights will be, but say as much as you can.

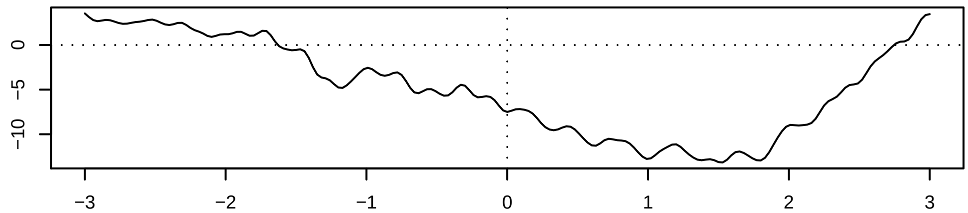
The answer is the same as for part (b), except that $w_1^{(1)}$ *will not go to infinity, but just some large value, and* $w_0^{(1)}$ *will therefore be guaranteed to converge to* $-w_1^{(1)}/2$*.*

Question 2: Below are five functions randomly drawn from five different Gaussian processes. For all five Gaussian processes, the mean function is zero. The covariance functions are one of those listed below.

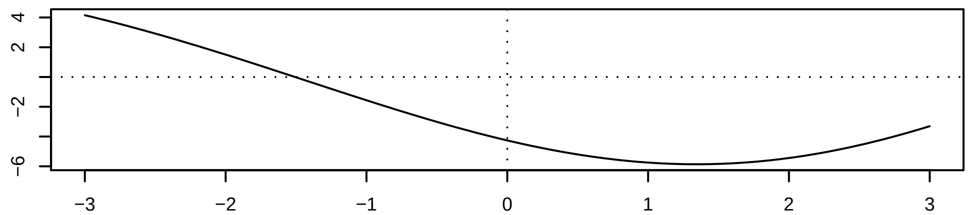
a)



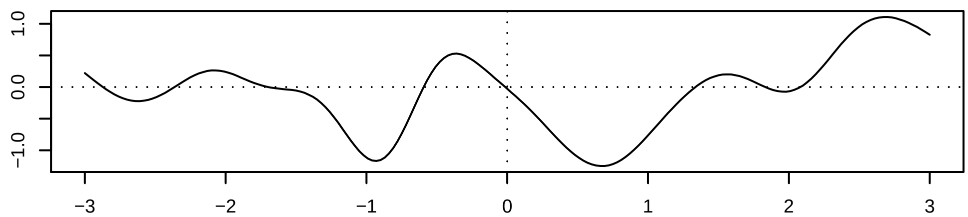
b)



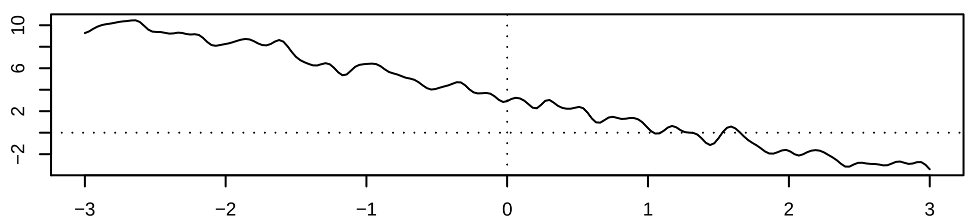
c)



d)



e)



For each of the five covariance functions below, indicate which of the five functions above is most likely to have been drawn from the Gaussian process with that covariance function.

1) $\text{Cov}(y_{i_1}, y_{i_2}) \;=\; 0.5^2 \, \exp(-((x_{i_1} - x_{i_2})\,/\,0.5)^2\,)$

Answer: (d)

2) $\text{Cov}(y_{i_1}, y_{i_2}) \;=\; x_{i_1}\, x_{i_2}$

Answer: (a)

3) $\text{Cov}(y_{i_1}, y_{i_2}) \;=\; 5^2 \;+\; 5^2\, x_{i_1}\, x_{i_2} \;+\; 0.5^2 \, \exp(-((x_{i_1} - x_{i_2})\,/\,0.1)^2\,)$

Answer: (e)

4) $\mathrm{Cov}(y_{i_1}, y_{i_2}) \;=\; 0.7^2\ \exp(\,-\,((x_{i_1} - x_{i_2})\,/\,0.1)^2\,) \;+\; 8^2\ \exp(\,-\,((x_{i_1} - x_{i_2})\,/\,2)^2\,)$

Answer: (b)

5) $\mathrm{Cov}(y_{i_1}, y_{i_2}) \;=\; 8^2\ \exp(\,-\,((x_{i_1} - x_{i_2})\,/\,5)^2\,)$

Answer: (c)

Question 3: Suppose we model the relationship of a real-valued response variable, y, to a single real input, x, using a Gaussian process model in which the mean is zero and the covariances of the observed responses are given by

$$\mathrm{Cov}(y_i, y_{i'}) \;=\; 0.5^2\delta_{i,i'} \;+\; K(x_i, x_{i'})$$

with the noise-free covariance function, K, defined by

$$K(x, x') \;=\; \begin{cases} 1 - |x - x'| & \text{if } |x - x'| < 1 \\ 0 & \text{otherwise} \end{cases}$$

Suppose we have four training cases, as follows:

x	y
0.5	2.0
2.8	3.3
1.6	3.0
3.9	2.7

Recall that the conditional mean of the response in a test case with input x_*, given the responses in the training cases, is $k^T C^{-1} y$, where y is the vector of training responses, C is the covariance matrix of training responses, and k is the vector of convariances of training responses with the response in the test case.

Find the predictive mean for the response in a test case in which the input is $x_* = 1.2$.

The covariance matrix of the training responses is

$$C \;=\; \begin{bmatrix} 1+0.5^2 & 0 & 0 & 0 \\ 0 & 1+0.5^2 & 0 & 0 \\ 0 & 0 & 1+0.5^2 & 0 \\ 0 & 0 & 0 & 1+0.5^2 \end{bmatrix}$$

The inverse of this is

$$C^{-1} \;=\; \begin{bmatrix} 0.8 & 0 & 0 & 0 \\ 0 & 0.8 & 0 & 0 \\ 0 & 0 & 0.8 & 0 \\ 0 & 0 & 0 & 0.8 \end{bmatrix}$$

The vector of covariances of the test response with the training responses is

$$k \;=\; \begin{bmatrix} 1-0.7 \\ 0 \\ 1-0.4 \\ 0 \end{bmatrix} \;=\; \begin{bmatrix} 0.3 \\ 0 \\ 0.6 \\ 0 \end{bmatrix}$$

So $k^T C^{-1} \;=\; [0.24\ 0\ 0.48\ 0]$, and the predictive mean for the test resp[onse is

$$k^T C^{-1} y \;=\; 0.24 \times 2.0 \;+\; 0.48 \times 3.0 \;=\; 1.92$$

Question 4: Recall that for a Gaussian process model the predictive distribution for the response y^* in a test case with inputs x^* has mean and variance given by

$$\begin{aligned} E[y^* \,|\, x^*, \text{training data}] &= k^T C^{-1} y \\ \text{Var}[y^* \,|\, x^*, \text{training data}] &= v - k^T C^{-1} k \end{aligned}$$

where y is the vector of observed responses in training cases, C is the matrix of covariances for the responses in training cases, k is the vector of covariances of the response in the test case with the responses in training cases, and v is the prior variance of the response in the test case.

a) Suppose we have just one training case, with $x_1 = 3$ and $y_1 = 4$. Suppose also that the noise-free covariance function is $K(x, x') = 2^{-|x-x'|}$, and the variance of the noise is 1/2. Find the mean and variance of the predictive distribution for the response in a test case for which the value of the input is 5.

The mean of the preditive distribution is

$$K(3,5)[K(3,3) + 1/2]^{-1}(4) \;=\; (1/4)[1 + 1/2]^{-1}(4) \;=\; 4/6$$

The variance of the predictive distribution is

$$[K(5,5)+1/2]-K(3,5)[K(3,3)+1/2]^{-1}K(3,5) \;=\; [1+1/2]-(1/4)[1+1/2]^{-1}(1/4) \;=\; 35/24$$

b) Repeat the calculations for (a), but using $K(x, x') = 2^{+|x-x'|}$. What can you conclude from the result of this calculation?

The mean of the preditive distribution is

$$K(3,5)[K(3,3) + 1/2]^{-1}(4) \;=\; (4)[1 + 1/2]^{-1}(4) \;=\; 32/3$$

The variance of the predictive distribution is

$$[K(5,5)+1/2]-K(3,5)[K(3,3)+1/2]^{-1}K(3,5) \;=\; [1+1/2]-(4)[1+1/2]^{-1}(4) \;=\; -55/6$$

But variances cannot be negative! We can conclude that $K(x, x') = 2^{+|x-x'|}$ is not a valid covariance function — it is not positive semi-definite.

Question 5: Consider a binary classification problem in which two inputs are available for predicting the class — input x_1, which is binary, and input x_2, which is real-valued. Suppose we use a naive Bayes model in which x_1 and x_2 are assumed to be independent within each class. Let $P(x_1 = 1 \,|\, C_0) = \theta_0$ and $P(x_1 = 1 \,|\, C_1) = \theta_1$, and assume that $x_2|C_0 \sim N(\mu_0, \sigma^2)$ and $x_2|C_1 \sim N(\mu_1, \sigma^2)$, where θ_0, θ_1, μ_0, μ_1, and σ are parameters to be estimated from the training data.

Supposing that these parameters have been estimated, as $\hat{\theta}_0$, $\hat{\theta}_1$, $\hat{\mu}_0$, $\hat{\mu}_1$, and $\hat{\sigma}$, and that some estimate for the "prior" probability of class 1, $P(C_1)$ is available, work out an expression for the probability of class 1 for a test case with inputs (x_1^*, x_2^*).

The odds in favour of class C_1 will be

$$\begin{aligned}
\frac{P(C_1|x_1^*, x_2^*)}{P(C_0|x_1^*, x_2^*)} &= \frac{P(C_1)}{P(C_0)} \frac{P(x_1^*|C_1)}{P(x_1^*|C_0)} \frac{P(x_2^*|C_1)}{P(x_2^*|C_0)} \\
&= \frac{P(C_1)}{P(C_0)} \frac{\theta_1^{x_1^*}\,(1-\theta_1)^{1-x_1^*}}{\theta_0^{x_1^*}\,(1-\theta_0)^{1-x_1^*}} \frac{(2\pi)^{-1/2}\sigma^{-1}\exp(-(x_2^*-\mu_1)/2\sigma^2)}{(2\pi)^{-1/2}\sigma^{-1}\exp(-(x_2^*-\mu_0)/2\sigma^2)} \\
&= \frac{P(C_1)}{P(C_0)} \left(\frac{\theta_1}{\theta_0}\right)^{x_1^*} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{1-x_1^*} \frac{\exp(-((x_2^*)^2 - 2\mu_1 x_2^* + \mu_1^2)/2\sigma^2)}{\exp(-((x_2^*)^2 - 2\mu_0 x_2^* + \mu_0^2)/2\sigma^2)} \\
&= \frac{P(C_1)}{P(C_0)} \left(\frac{\theta_1}{\theta_0}\right)^{x_1^*} \left(\frac{1-\theta_1}{1-\theta_0}\right)^{1-x_1^*} \frac{\exp(\mu_1 x_2^*/\sigma^2 - \mu_1^2/2\sigma^2)}{\exp(\mu_0 x_2^*/\sigma^2 - \mu_0^2/2\sigma^2)}
\end{aligned}$$

The log odds, which we'll call $a(x^*)$, will therefore be

$$\begin{aligned}
a(x^*) &= \log\left(\frac{P(C_1|x_1^*, x_2^*)}{P(C_0|x_1^*, x_2^*)}\right) \\
&= \log\left(\frac{P(C_1)}{P(C_0)}\right) + \log\left(\frac{1-\theta_1}{1-\theta_0}\right) + (\mu_0^2 - \mu_1^2)/2\sigma^2 + x_1^*\left[\log\left(\frac{\theta_1/(1-\theta_1)}{\theta_0/(1-\theta_0)}\right) + (\mu_1 - \mu_0)\,/\,\sigma^2\right]
\end{aligned}$$

The probability of class 1 can then be written as $1/(1+\exp(-a(x^*)))$.

Question 6: Recall that the maximum margin separating hyperplane, defined by $w^T x + b = 0$, can be found by solving the following optimiziation problem:

$$\text{minimize } ||w||^2, \quad \text{subject to } y_i(w^T x_i + b) \ \geq\ 1 \text{ for } i = 1, \ldots, n$$

Use this to find the maximum margin hyperplane for the following $n = 3$ data points, (x, y), in which x is one-dimensional:

$$(-1, -1),\ (2, +1),\ (3, +1)$$

(Note that a separating "hyperplane" when x is one-dimensional is a single point.)

You should produce a two-dimensional plot of the linear inequality constraints on w and b, and from this find the minimum of $||w||^2$. You should then plot the function $wx + b$, verify that the point where $wx + b = 0$ separates the classes, and find its margin.

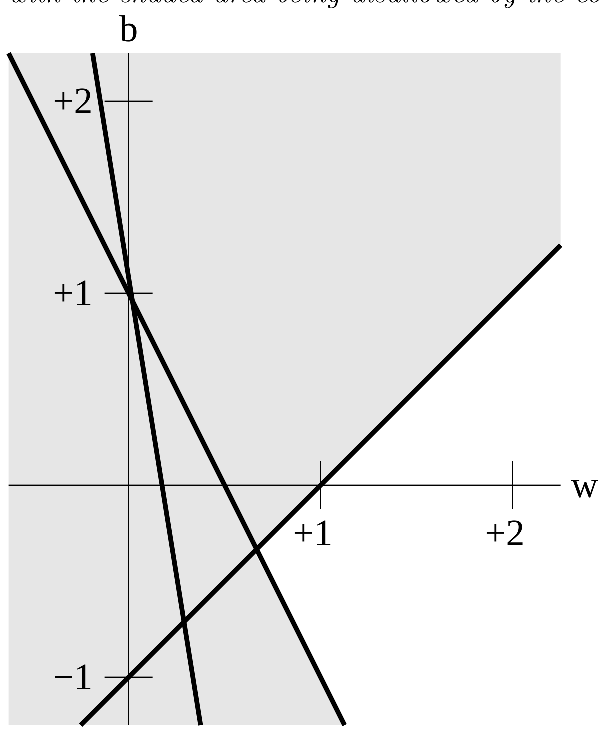
The inequality constraints are as follows:

$$\begin{aligned} (-1)(w(-1) + b) &= w - b \ \geq 1 \\ (+1)(w(2) + b) &= 2w + b \ \geq 1 \\ (+1)(w(3) + b) &= 3w + b \ \geq 1 \end{aligned}$$

which are equivalent to

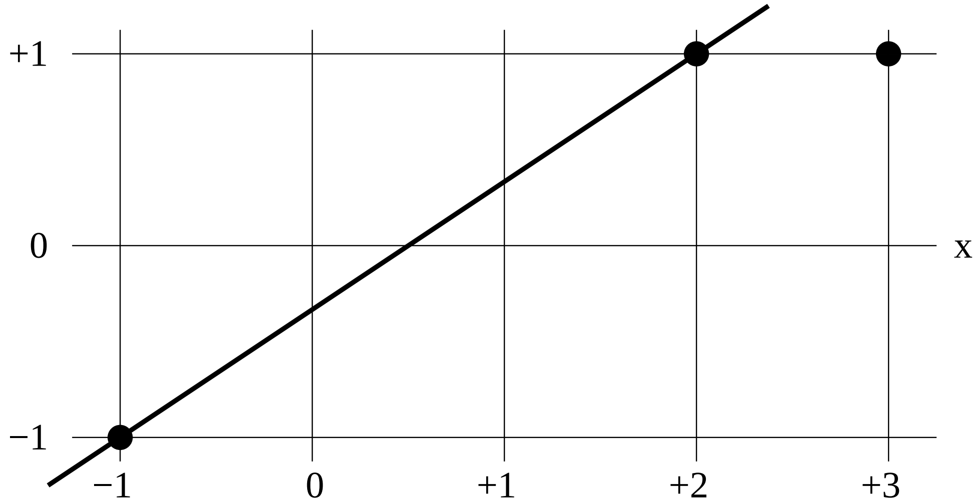
$$\begin{aligned} b &\ \leq -1 + w \\ b &\ \geq 1 - 2w \\ b &\ \geq 1 - 3w \end{aligned}$$

These are plotted below, with the shaded area being disallowed by the constraints:



The minimum allowed value for w^2 is at $w = 2/3$ and $b = -1/3$.

Here is a plot of $wx + b$ with $w = 2/3$ and $b = -1/3$, showing the training cases.



The separating point is at $x = 1/2$, with margin of $\pm 3/2$ around that point.