STA 414/2104, Spring 2013, Answers to Practice Problem Set #3

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 a single input, x_1 (so $p = 1$), 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:

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

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 $x \le 0$ to 5 for $x \ge 1$. 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.

Question 2: Consider the factor analysis model, $x = \mu + Wz + \epsilon$, where x is an observed vector of p variables, μ is the mean vector for x, z is an unobserved vector of m common factors, W is the matrix of "factor loadings", and ϵ is a random residual. We assume that $z \sim N(0, I)$ and independently $\epsilon \sim N(0, \Sigma)$, where Σ is diagonal with diagonal entries $\sigma_1^2, \ldots, \sigma_p^2$.

Let the number of observed variables be $p = 4$ and the number of common factors be $m = 1$.

a) Give an explicit example (specifying μ, W, and Σ) showing that it is possible for the correlation of x_1 and x_2 to be negative, the correlation of x_1 and x_3 to be positive, and the correlation of x_1 and x_4 to be zero. Compute the covariance and correlation matrices of x for your example.

One possible example is $\mu = [0\ 0\ 0\ 0]^T$, $\Sigma = I$, and $W = [1\ \ {-1}\ 1\ 0]^T$. The covariance matrix of x will then be

$$E[(Wz+\epsilon)(Wz+\epsilon)^T] \;=\; E[Wzz^TW^T + \epsilon\epsilon^T] \;=\; WW^T + \Sigma \;=\; \begin{bmatrix} 2 & -1 & 1 & 0 \\ -1 & 2 & -1 & 0 \\ 1 & -1 & 2 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

The correlation matrix will be

$$\begin{bmatrix} 1 & -1/2 & 1/2 & 0 \\ -1/2 & 1 & -1/2 & 0 \\ 1/2 & -1/2 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

b) Suppose that $\mu_j = 0$ and $\sigma_j^2 = 4$ for $j = 1, 2, 3, 4$, and $W = [\,3\ 2\ 1\ 0\,]^T$. Find the covariance matrix for x, the direction of the first principal component of that covariance matrix, and the variance in that direction.

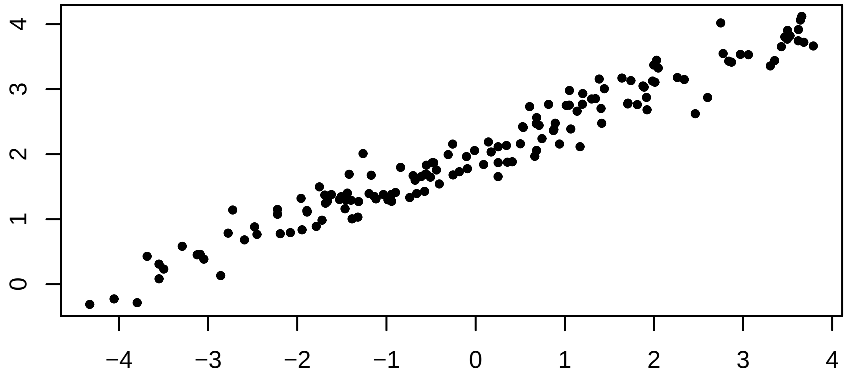
The covariance matrix is $WW^T + 4I$*, which is*

$$\begin{bmatrix} 13 & 6 & 3 & 0 \\ 6 & 8 & 2 & 0 \\ 3 & 2 & 5 & 0 \\ 0 & 0 & 0 & 4 \end{bmatrix}$$

One eigenvector of this matrix is W*, with eigenvalue* $3^2 + 2^2 + 1^2 + 0^2 + 4 = 19$*, as can be seen from* $(WW^T + 4I)W = (W^TW + 4)W$*. The other eigenvectors will be orthogonal to this eigenvector, and hence will have eigenvalue 4, since for such an eigenvector,* V*,* $(WW^T + 4I)V = W(W^TV) + 4V = 4V$*.*

So the first principal component direction is $[3\ 2\ 1\ 0]$*, and the variance in this direction is 19.*

Question 3: Below is a scatterplot of 150 observations of two variables:



a) Write down a vector pointing in the direction of the first principal component for this data. An approximate answer found by eye is sufficient. The vector need not have length one.

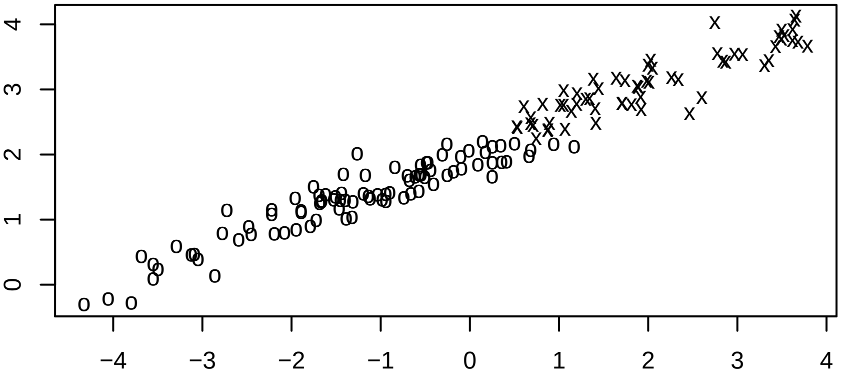
Also, draw the direction of the first principal component on the scatterplot above.

One answer is $[2\ 1]$*. I won't try to draw this on the plot.*

b) What is the approximate standard deviation in the first principal component's direction?

Somewhere around 2 or 3.

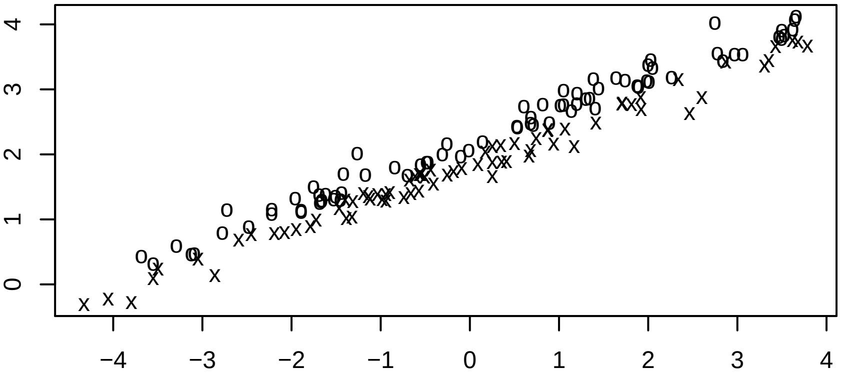
c) Suppose that each of these data points are associated with one of two classes, as shown below (with one class marked by "o" and the other by "x"):



If we reduce the data to just the projection on the first principal component, how well will we be able to classify the data points using this one number, compared to how well we would have been able to classify using the two original numbers?

We will be able to classify almost as well as with the original data.

d) Suppose instead that the two classes are as shown below:



In this case, how well will we be able to classify using just the projection on the first principal component, compared to using the two original numbers?

The projection on the first principal component will give almost no information about the class. One could do much better using the original data, since one can see in the plot that points in the circle class are usually above those in the x class. So there is a a diagonal line that separates the classes fairly well.

Question 4: 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

$$\frac{P(C_1|x_1^*, x_2^*)}{P(C_0|x_1^*, x_2^*)} \quad = \quad \frac{P(C_1)}{P(C_0)} \, \frac{P(x_1^*|C_1)}{P(x_1^*|C_0)} \, \frac{P(x_2^*|C_1)}{P(x_2^*|C_0)}$$

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

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 5: We have two i.i.d. observations of seven variables, as follows:

$$
\begin{array}{ccccccc}
5 & 7 & 8 & 2 & 3 & 5 & 2\\
3 & 3 & 6 & 6 & 1 & 1 & 0
\end{array}
$$

a) Find a 7-dimensional vector of length one that points in the direction of the first principal component of this data. Explain how you obtained it.

First, we subtract the sample means from the two observed vectors, giving the following centred data:

$$
\begin{array}{rrrrrrr}
1 & 2 & 1 & -2 & 1 & 1 & 1\\
-1 & -2 & -1 & 2 & -1 & -2 & -1
\end{array}
$$

With only two training cases, each of these vectors must point in the direction of the first principal component. Taking the first, its length is 4, so one vector of length 1 in the direction of the first principal component is

$$
\left[\begin{array}{ccccccc}\frac{1}{4} & \frac{1}{2} & \frac{1}{4} & -\frac{1}{2} & \frac{1}{4} & \frac{1}{2} & \frac{1}{4}\end{array}\right]^T
$$

The other possible answer is the negation of the above.

It's also possible to answer this question by computing

$$
XX^T = \begin{bmatrix} 16 & -16\\ -16 & 16\end{bmatrix}
$$

and then finding its eigenvectors, $[1\ \ -1]^T$ *and* $[1\ 1]^T$*, which have eigenvalues 32 and 0. PC1 is in the direction* $X^T[1\ -1]^T$*. After scaling to unit length, this gives the same answer as above.*

b) Find the projection on this principal component of the new observation shown below:

$$
4\ \ 1\ \ 9\ \ 3\ \ 2\ \ 2\ \ 1
$$

Subtracting the sample means from the training data gives $[0\ -4\ \ 2\ -1\ \ 0\ -1\ \ 0]^T$*. The dot product of this with the PC1 vector from (a) is* $-3/2$.

Question 6: Recall that in a factor analysis model an observed data point, x, is modeled using M latent factors as

$$x = \mu + Wz + \epsilon$$

where μ is a vector of means for the p components of x, W is a $p \times M$ matrix, z is a vector of M latent factors, assumed to have independent $N(0, 1)$ distributions, and ϵ is a vector of p residuals, assumed to be independent, and to come from normal distributions with mean zero. The variance of ϵ_j is σ_j^2.

Suppose that $p = 5$ and $M = 2$, and that the parameters of the model are mean $\mu = [0\ 0\ 0\ 0\ 0]^T$, residual standard deviations $\sigma_1 = 1$, $\sigma_2 = 1$, $\sigma_3 = 2$, $\sigma_4 = 2$, $\sigma_5 = 2$, and

$$W = \begin{bmatrix} 1 & 2 \\ -1 & 1 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}$$

a) Find the covariance matrix for x.

$$\text{Cov}(x) = E[(Wz + \epsilon)\,(Wz + \epsilon)^T] = WW^T + \text{diag}(\sigma_1^2, \ldots, \sigma_5^2)$$

$$= \begin{bmatrix} 6 & 1 & 1 & 1 & 2 \\ 1 & 3 & -1 & -1 & 1 \\ 1 & -1 & 5 & 1 & 0 \\ 1 & -1 & 1 & 5 & 0 \\ 2 & 1 & 0 & 0 & 5 \end{bmatrix}$$

b) Suppose that we don't observe vectors x of dimension five, but rather we observe vectors y of dimension four, where $y_1 = x_1$, $y_2 = 3x_2$, $y_3 = -x_3$, and $y_4 = 2x_4 + x_5$. Assuming that the distribution of x is given by the factor analysis model with parameters above, write down a factor analysis model (including values of its parameters) for the distribution of y.

Using the relation of y to x and the model for x above, we can write $y = W'z + \epsilon'$, where

$$W' = \begin{bmatrix} 1 & 2 \\ -3 & 3 \\ -1 & 0 \\ 2 & 1 \end{bmatrix}$$

The standard deviations of the ϵ'_i will be $\sigma'_1 = \sigma_1 = 1$, $\sigma'_2 = 3\sigma_2 = 3$, $\sigma'_3 = \sigma_3 = 2$, and $\sigma'_4 = \sqrt{4\sigma_4^2 + \sigma_5^2} = \sqrt{20}$.

Question 7: Consider a binary classification task in which a 0/1 response, y, is to be predicted from three binary covariates, x_1, x_2, x_3. We have six training cases, as follows:

y	x_1	x_2	x_3
0	1	0	1
0	0	1	0
1	1	0	1
1	1	1	0
1	0	0	1
1	1	0	0

We decide to use a naive Bayes model for this task, in which the three covariates are modeled as being independent within each class. The distribution for covariate j within class k is modeled as Bernoulli(θ_{kj}). We estimate the probabilities of the classes and θ_{kj} for $k = 0, 1$ and $j = 1, 2, 3$ from the training data, by maximum likelihood.

a) Based on the training data above, what will be the estimates for the class probabilities and for the θ_{kj} parameters?

The class probabilities will be estimated from the frequencies in the training data as $P(y = 0) = 2/6 = 1/3$ *and* $P(y = 1) = 4/6 = 2/3$.

The probabilities for the x_i *given* $y = 0$ *will be estimated from the two training cases with* $y = 0$ *as* $\theta_{01} = \theta_{02} = \theta_{03} = 1/2$.

The probabilities for the x_i *given* $y = 1$ *will be estimated from the four training cases with* $y = 0$ *as* $\theta_{11} = 3/4$, $\theta_{12} = 1/4$, *and* $\theta_{13} = 2/4 = 1/2$.

b) According to this naive Bayes model, using the training data above, what is that probability that $y = 1$ for each of the test cases below?

- $x_1 = 1,\ x_2 = 1,\ x_3 = 0$
  Answer:

$$
\begin{aligned}
&P(y = 1 \,|\, x_1 = 1, x_2 = 1, x_3 = 0) \\
&\quad = \frac{P(y=1)\,P(x_1=1,x_2=1,x_3=0|y=1)}{P(y=0)\,P(x_1=1,x_2=1,x_3=0|y=0) \;+\; P(y=1)\,P(x_1=1,x_2=1,x_3=0|y=1)} \\
&\quad = \frac{(2/3)\,(3/4)\,(1/4)\,(1/2)}{(1/3)\,(1/2)\,(1/2)\,(1/2) \;+\; (2/3)\,(3/4)\,(1/4)\,(1/2)} \\
&\quad = 3/5
\end{aligned}
$$

- $x_1 = 1,\ x_2 = 0,\ x_3 = 1$
  Answer:

$$
\begin{aligned}
&P(y = 1 \,|\, x_1 = 0, x_2 = 0, x_3 = 1) \\
&\quad = \frac{P(y=1)\,P(x_1=1,x_2=0,x_3=1|y=1)}{P(y=0)\,P(x_1=1,x_2=0,x_3=1|y=0) \;+\; P(y=1)\,P(x_1=1,x_2=0,x_3=1|y=1)} \\
&\quad = \frac{(2/3)\,(3/4)\,(3/4)\,(1/2)}{(1/3)\,(1/2)\,(1/2)\,(1/2) \;+\; (2/3)\,(3/4)\,(3/4)\,(1/2)} \\
&\quad = 9/11
\end{aligned}
$$

c) Suppose that the loss from classifying an item as being in class 1 when it is really in class 0 is twice as large as the loss from classifying an item as being in class 0 when it is really in class 1. How should you classify each of the following test cases?

- $x_1 = 1,\ x_2 = 1,\ x_3 = 0$

  Let the loss classifying as class 1 when really class 0 be 2, and the loss classifying as class 0 when really class 1 be 1.
  Expected loss if you classify as class 0 is $1 \times P(y = 1|x_1 = 1, x_2 = 1, x_3 = 0) \;=\; 3/5$.
  Expected loss if you classify as class 1 is $2 \times P(y = 0|x_1 = 1, x_2 = 1, x_3 = 0) \;=\; 4/5$.
  So you should classify as class 0.

- $x_1 = 1,\ x_2 = 0,\ x_3 = 1$

  Expected loss if you classify as class 0 is $1 \times P(y = 1|x_1 = 1, x_2 = 0, x_3 = 1) \; = \; 9/11.$
  Expected loss if you classify as class 1 is $2 \times P(y = 0|x_1 = 1, x_2 = 0, x_3 = 1) \; = \; 4/11.$
  So you should classify as class 1.