CSCI E-181 Midterm 1
Spring 2014

Name: DAVID WIHL

Start Time: 2 30 pm   3/26/14

End Time: 5 30 pm

1	/ 10
2	/ 20
3	/ 15
4	/ 15
5	/ 10
6	/ 15
7	/ 15
Total	/ 100

1. **K-Means++** {10pts}

> Imagine that you have N data and you wish to find K clusters using K-Means++. Assuming that $N > K$, can the K-Means++ algorithm choose the same datum twice to become a cluster center? Why or why not?

K-Means++ offers an improvement over K-Means by choosing an initial center from the dataset rather a random point.

It will <u>not</u> choose the same datum twice because the next center will be chosen from the remaining datapoints, in proportion to the distance to the cluster center. If the original is already the best center, it will not be chosen again.

2. Hierarchical Agglomerative Clustering {20pts}

In the two figures below, draw the dendrogram for the data on the left, where the y-axis provides their values. In the top figure, use the single-linkage criterion (min over between-group distances) and in the bottom figure use the complete-linkage criterion (max over between-group distances).
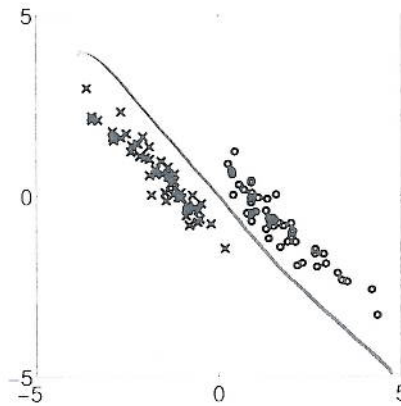
(a) {10pts} Single Linkage:
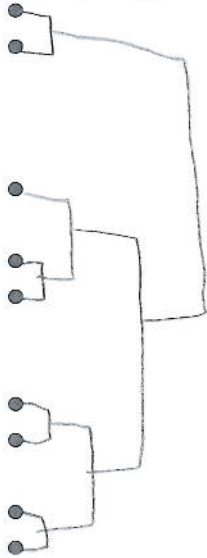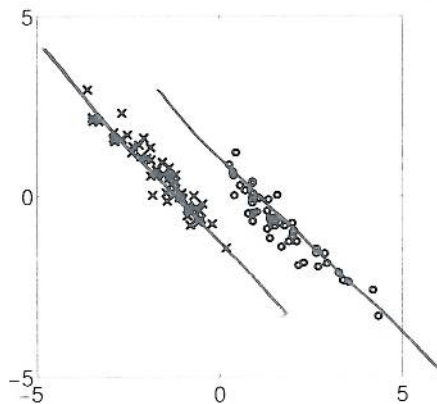
(b) {10pts} Complete Linkage:
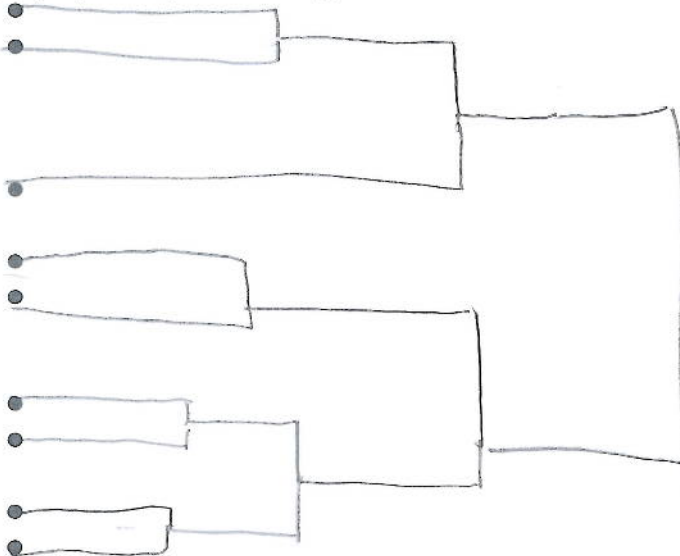
3. Subspace Projection {15pts}

In each of the following figures, draw the line or lines onto which we would project the data. There are two classes shown in each figure, one by 'x' and the other by 'o'.
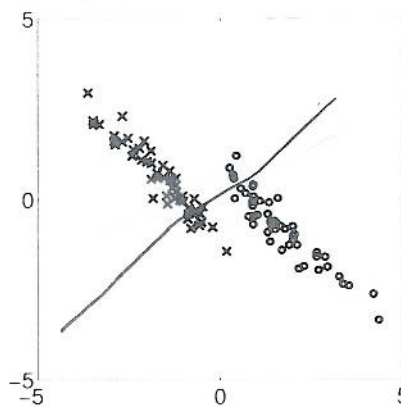
(a) {5pts}Draw the single line for a PCA projection, ignoring the class labels.

(b) {5pts}Draw the two lines for PCA, treating each class separately.

(c) {5pts} Draw the single line for projection with Fisher's Linear Discriminant.

4

4. Weighted Linear Regression {15pts}

Consider a data set in which each of the N data points is associated with a "weighting" $r_n \geq 0$. You might imagine doing something like this if some of the data were more important than others in your training set. Consider the weighted sum-of-squares loss function for basis function linear regression, i.e., where the target $t_n \in \mathbb{R}$:

$$E_D(w) = \frac{1}{2} \sum_{n=1}^{N} r_n \{ t_n - w^\top \phi(x_n) \}^2.$$

Find an expression for the solution w^* that minimizes this error function.

Let R = diagonal matrix $(r_1, r_2, \ldots, r_n)$

$$E_D(\underline{w}) = \frac{1}{2} (\underline{t} - \Phi \underline{w})^\top R (\underline{t} - \Phi \underline{w})$$

$$\frac{\partial}{\partial \underline{w}} \left(\frac{1}{2} (\underline{t} - \Phi \underline{w})^\top R (\underline{t} - \Phi \underline{w}) \right) = 0$$

$$\underline{w}^* = \left(\Phi^\top R \Phi \right)^{-1} \Phi^\top R \underline{t}$$

5. Biased Coins {10pts}

You have a box full of coins. There are two types of coins, C_1 and C_2. Coins of type C_1 come up heads with probability 0.8 and coin of type C_2 come up heads with probability 0.2. There are many more C_1 coins in the box than C_2 coins, in fact 90% of the coins are of type C_1. You grab a coin at random from inside the box and flip it 10 times, getting five heads and five tails. Compute $p(D|C_1)$, $p(D|C_2)$. How probable is it that you have a coin of type C_1, given these ten flips?

$n = $ # of flips
$k = $ # of heads
$p = $ probability of heads

$$p(D|C_1) = \binom{n}{k} p^k (1-p)^{n-k}$$

$$= \frac{10!}{5!(10-5)!} \cdot (0.8)^5 \cdot (1-0.8)^{10-5}$$

$$\cong 0.0264$$

$$p(D|C_2) = \frac{10!}{5!(10-5)!} \cdot (0.2)^5 (1-0.2)^{10-5}$$

$$\cong 0.0264$$

Given $p(D|C_1) = p(D|C_2)$, we learn nothing from the coin tosses,

$$p(C_1) = 0.9 \text{ (given)}$$

5. **Biased Coins {10pts}**

You have a box full of coins. There are two types of coins, C_1 and C_2. Coins of type C_1 come up heads with probability 0.8 and coin of type C_2 come up heads with probability 0.2. There are many more C_1 coins in the box than C_2 coins, in fact 90% of the coins are of type C_1. You grab a coin at random from inside the box and flip it 10 times. Compute $p(D|C_1)$, $p(D|C_2)$. How probable is it that you have a coin of type C_1, given these ten flips?

$$p(H|C_1) = 0.8 \cdot 0.8$$

$$p(T|C_1) = 0.9 \cdot 0.2$$

$$p(H|C_2) = 0.1 \cdot 0.2$$

$$p(T|C_2) = 0.1 \cdot 0.8$$
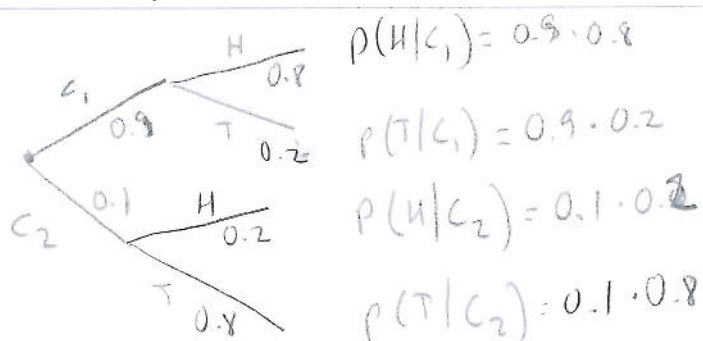
$$E[D] = p(C_1)\, p(D|C_1) + p(C_2) \cdot p(D|C_2)$$

$$p(C_1) = \frac{E[D] - p(C_2)\, p(D|C_2)}{p(D|C_1)}$$

6

6. Redundant Features in Naïve Bayes [15pts]

Suppose that we use a Naïve Bayes classifier to classify binary data with binary feature vectors $x_n \in \{0,1\}^D$. We'll classify them into two classes, C_1 and C_2. With Naïve Bayes and binary features, the class conditional distributions will be of the form of a product of Bernoulli distributions:

$$p(x \mid C_k) = \prod_{d=1}^{D} \mu_{kd}^{x_d} (1 - \mu_{kd})^{(1-x_d)},$$

where $x_d \in \{0,1\}$, and $\mu_{kd} = p(x_d = 1 \mid C_k)$. Assume also that the class priors are uniform, i.e., $p(C_1) = p(C_2) = \frac{1}{2}$.

(a) {5pts} If $D = 1$ (i.e., there is only one feature), use the equations above to write out $\ln \frac{p(C_1 \mid x)}{p(C_2 \mid x)}$ for a single binary feature x.

$$p(c_d \mid x) = \frac{p(c_k) \cdot p(x \mid c_k)}{p(x)}$$

$$\ln \frac{p(c_1 \mid x)}{p(c_2 \mid x)} = \ln \left(\frac{p(c_1) \cdot p(x \mid c_1)}{p(x)} \right) - \ln \left(\frac{p(c_2) \cdot p(x \mid c_2)}{p(x)} \right)$$

(b) {5pts} Now suppose we change our feature representation so that instead of using just a single feature, we use two redundant features (i.e., two features that always have the same value so that $x_1 = x_2$). Since they are the same, you can assume that $\mu_{k1} = \mu_{k2}$ also. With this feature representation, let's write $\hat{x} = x_1 \cdot x_2$, since there can only be two configurations of the x_1, x_2 pair, instead of four. What is $\ln \frac{p(C_1 \mid \hat{x})}{p(C_2 \mid \hat{x})}$ in terms of the value for $\ln \frac{p(C_1 \mid x)}{p(C_2 \mid x)}$ you calculated in part (a)?

$$p(c_k \mid \hat{x}) = \frac{p(\hat{x} \mid c_k) \cdot p(c_k)}{\sum_{k=1}^{} p(\hat{x} \mid c_k) \, p(c_k)}$$

(c) {5pts} Does this seem like a bug or a feature? Why?

They seems like a bug since our calculated probability is lower even though the information content has not changed.

7. Binomial Regression {15pts}

You've been hired by a startup to build a ratings system for restaurants. Users rate the restaurants on a scale of 0 to 10 (i.e., $t_n \in \{0, 1, \ldots, 10\}$) and you have a set of real-valued features for each restaurant, $x_n \in \mathbb{R}^D$. Given the range of the t_n, it seems like a binomial distribution would be a good choice for building a regression model:

$$p(k \mid \rho) = \binom{10}{k} \rho^k (1 - \rho)^{10-k},$$

where ρ parameterizes the distribution and takes values in $(0, 1)$, while k is the rating. Recall that $\binom{N}{k}$ is the binomial coefficient, i.e., $N!/(K!(N-K)!)$.

(a) {5pts} We cook up some basis functions $\phi_i(x)$ and we plan to weight them using a set of weights w to determine ρ. However, $\phi(x)^T w$ can be negative and can be greater than one. How can we map it into the right space?

$$\hat{\rho} = \frac{\phi(x)^T w}{\|\phi(x)^T w\|}$$, using an eigenvalue to normalize and change the sign

(b) {5pts} Having figured out how to get a map into the right space, write down the log likelihood of a set of N data $\{t_n, x_n\}_{n=1}^N$. You can ignore constants in the sum that don't depend on the inputs or w.

$$\log p(t \mid N) = \sum_n t_n \log(y_n) + (1 - t_n) \log(1 - y_n)$$

$$y(\phi) = \sigma(w^T \phi)$$

$$E(w) = -\log p(t \mid w)$$

(c) {5pts} Compute the gradient of the log likelihood in terms of w. Hint: the derivative of the logistic function is $\frac{d}{dz}\sigma(z) = \sigma(z)(1 - \sigma(z))$.

Error = $-\log$ likelihood $E(w) = -\log\left(p(t \mid w)\right)$

$$\nabla E(w) = \sum_{n=1}^N (y_n - t_n) \phi_n$$