CS 181 Midterm 1 — Solutions

Spring 2014

Name:

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

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

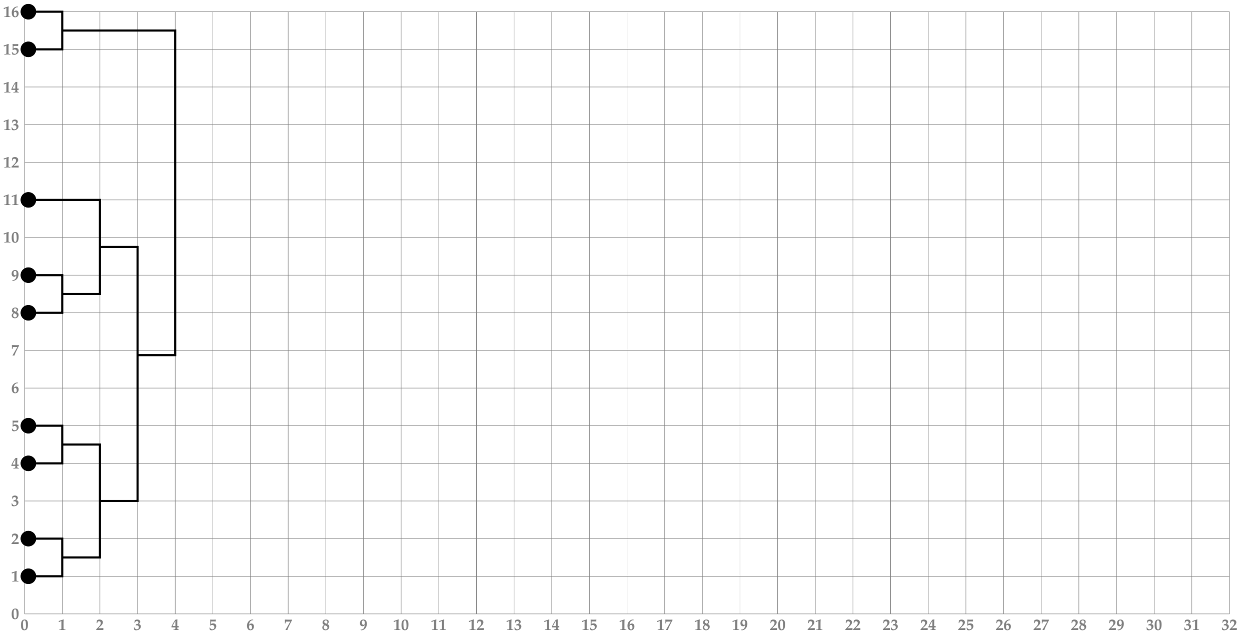
Imagine that you have N data and you wish to find K clusters using K-Means++. Assuming that $N > K$, can the K-Means++ algorithm choose the same datum twice to become a cluster center? Why or why not?

The K-Means++ algorithm will never choose the same datum twice to become a center. This is because the distribution over the data items is proportional to the squared distance to the closest cluster center. When a datum is a cluster center, this distribution will be zero for that item.

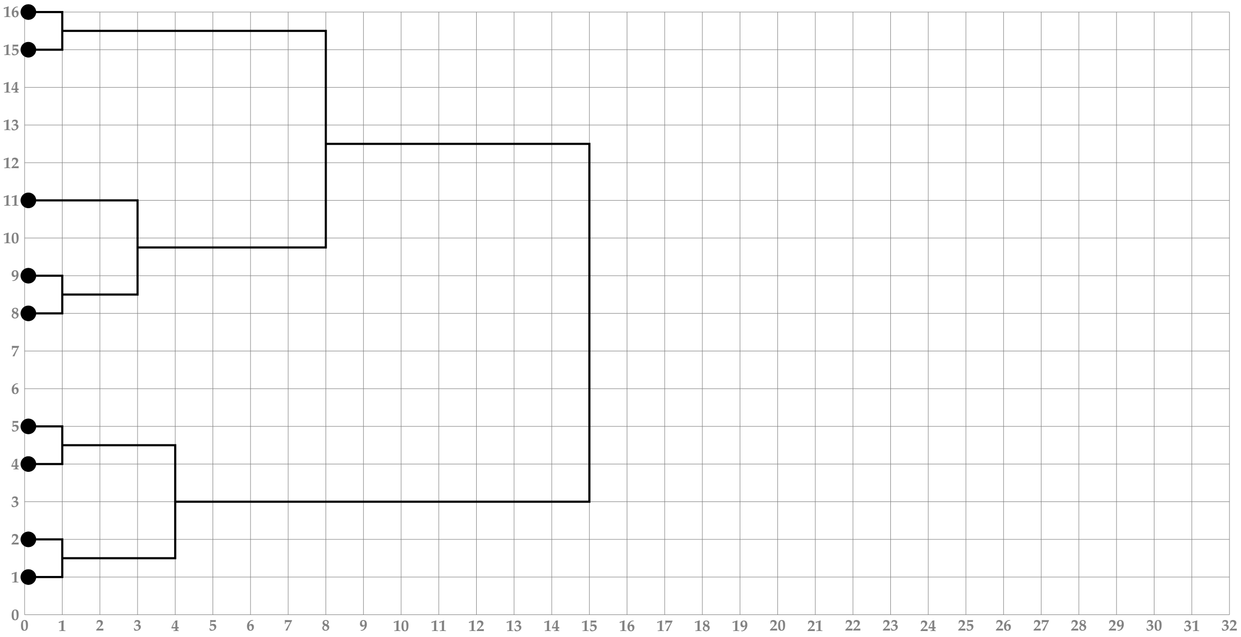
2. **Hierarchical Agglomerative Clustering** {20pts}

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:



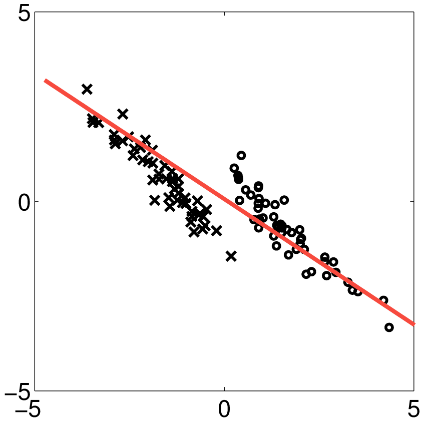
(b) {10pts} Complete Linkage:



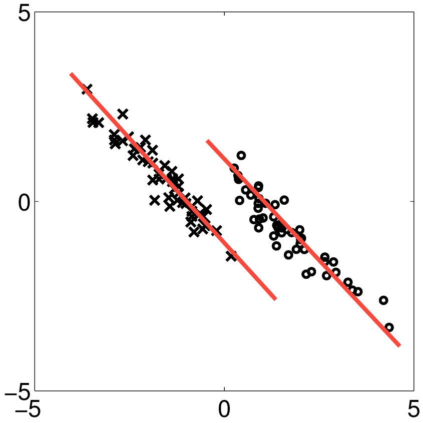
3. **Subspace Projection** {15pts}

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'.

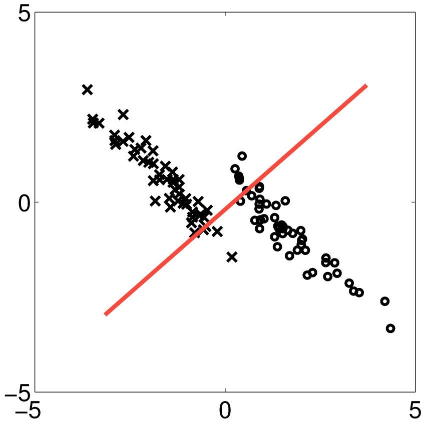
(a) {5pts}Draw the single line for a PCA projection, ignoring the class labels.



(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. **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 $t_n \in \mathbb{R}$:

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

Find an expression for the solution $\boldsymbol{w}^*$ that minimizes this error function.

Take the gradient and set to zero:

$$\frac{\partial}{\partial \boldsymbol{w}} E_D(\boldsymbol{w}) = -\sum_{n=1}^{N} r_n\{t_n - \boldsymbol{w}^\top \phi(\boldsymbol{x}_n)\}\phi(\boldsymbol{x}_n) = 0.$$

Solve for $\boldsymbol{w}$:

$$\begin{aligned}\sum_{n=1}^{N} r_n t_n \phi(\boldsymbol{x}_n) &= \left(\sum_{n=1}^{N} r_n \phi(\boldsymbol{x}_n)\phi(\boldsymbol{x}_n)^\mathsf{T}\right)\boldsymbol{w} \\ \boldsymbol{w} &= \left(\sum_{n=1}^{N} r_n \phi(\boldsymbol{x}_n)\phi(\boldsymbol{x}_n)^\mathsf{T}\right)^{-1}\left(\sum_{n=1}^{N} r_n t_n \phi(\boldsymbol{x}_n)\right)\end{aligned}$$

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

You have a box full of coins. There are two types of coins, $\mathcal{C}_1$ and $\mathcal{C}_2$. Coins of type $\mathcal{C}_1$ come up heads with probability 0.8 and coin of type $\mathcal{C}_2$ come up heads with probability 0.2. There are many more $\mathcal{C}_1$ coins in the box than $\mathcal{C}_2$ coins, in fact 90% of the coins are of type $\mathcal{C}_1$. You grab a coin at random from inside the box and flip it 10 times, getting five heads and five tails. Compute $p(D \,|\, \mathcal{C}_1)$, $p(D \,|\, \mathcal{C}_2)$. How probable is it that you have a coin of type $\mathcal{C}_1$, given these ten flips?

The prior probabilities are $p(\mathcal{C}_1) = 0.9$ and $p(\mathcal{C}_2) = 0.1$. We compute:

$$\begin{aligned} p(D \,|\, \mathcal{C}_1) &= \binom{10}{5}(0.8)^5(0.2)^{10-5} \\ p(D \,|\, \mathcal{C}_2) &= \binom{10}{5}(0.2)^5(0.8)^{10-5} \end{aligned}$$

Each coin has identical likelihoods, so it is only the prior that matters: $p(\mathcal{C}_1 \,|\, D) = 0.9$.

6. **Redundant Features in Naïve Bayes** [15pts]

> Suppose that we use a Naïve Bayes classifier to classify binary data with binary feature vectors $\boldsymbol{x}_n \in \{0,1\}^D$. We'll classify them into two classes, $\mathcal{C}_1$ and $\mathcal{C}_2$. With Naïve Bayes and binary features, the class conditional distributions will be of the form of a product of Bernoulli distributions:
>
> $$p(\boldsymbol{x} \,|\, \mathcal{C}_k) = \prod_{d=1}^{D} \mu_{kd}^{x_d} (1 - \mu_{kd})^{(1-x_d)},$$
>
> where $x_d \in \{0,1\}$, and $\mu_{kd} = p(x_d = 1 \,|\, \mathcal{C}_k)$. Assume also that the class priors are uniform, i.e., $p(\mathcal{C}_1) = p(\mathcal{C}_2) = \frac{1}{2}$.

(a) {5pts} If $D = 1$ (i.e., there is only one feature), use the equations above to write out $\ln \frac{p(\mathcal{C}_1 \,|\, x)}{p(\mathcal{C}_2 \,|\, x)}$ for a single binary feature x.

Because priors are equal:

$$\ln \frac{p(\mathcal{C}_1 \,|\, x)}{p(\mathcal{C}_2 \,|\, x)} = \ln \frac{p(x \,|\, \mathcal{C}_1)}{p(x \,|\, \mathcal{C}_2)}$$

So

$$\ln \frac{p(\mathcal{C}_1 \,|\, x)}{p(\mathcal{C}_2 \,|\, x)} = x \ln \mu_1 + (1-x)\ln(1-\mu_1) - x \ln \mu_2 - (1-x)\ln(1-\mu_2)$$

(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 $\ln \frac{p(\mathcal{C}_1 \,|\, \hat{x})}{p(\mathcal{C}_2 \,|\, \hat{x})}$ in terms of the value for $\ln \frac{p(\mathcal{C}_1 \,|\, x)}{p(\mathcal{C}_2 \,|\, x)}$ you calculated in part (a)?

$$\begin{aligned}\ln \frac{p(\mathcal{C}_1 \,|\, \hat{x})}{p(\mathcal{C}_2 \,|\, \hat{x})} &= \ln \frac{p(x_1 \,|\, \mathcal{C}_1)p(x_2 \,|\, \mathcal{C}_1)}{p(x_1 \,|\, \mathcal{C}_2)p(x_2 \,|\, \mathcal{C}_2)} \\ &= 2\left(\hat{x} \ln \mu_1 + (1-\hat{x})\ln(1-\mu_1) - \hat{x}\ln \mu_2 - (1-\hat{x})\ln(1-\mu_2)\right)\end{aligned}$$

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

This is a bug because it is now more confident than it should be. These features are tightly coupled, but naïve Bayes assumes they are independent.

7. **Binomial Regression** $\{15\text{pts}\}$

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, $\boldsymbol{x}_n \in \mathbb{R}^D$. Given the range of the t_n, it seems like a binomial distribution would be a good choice for building a regression model:

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

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

Solution:
This is a perfect use case for the logistic or sigmoid function $\sigma(z) = \frac{1}{1+e^{-z}}$.

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

Solution:
We have that the likelihood is:

$$p(\{t_n\} \,|\, \{\boldsymbol{x}_n\}, \boldsymbol{w}) = \prod_n \binom{10}{k} \sigma(\boldsymbol{\phi}(x_n)^\mathsf{T}\boldsymbol{w})^{t_n} (1 - \sigma(\boldsymbol{\phi}(x_n)^\mathsf{T}\boldsymbol{w}))^{10-t_n}$$

The log-likelihood is:

$$\ln p(\{t_n\} \,|\, \{\boldsymbol{x}_n\}, \boldsymbol{w}) = \sum_n t_n \ln \sigma(\boldsymbol{\phi}(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w}) + (10 - t_n) \ln(1 - \sigma(\boldsymbol{\phi}(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w})),$$

disregarding the $\binom{10}{k}$ since it is a constant.

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

Solution:

Taking the derivative, we have:

$$\begin{aligned}
&\frac{d}{d\boldsymbol{w}} \ln p(\{t_n\} \mid \{\boldsymbol{x}_n\}, \boldsymbol{w}) \\
&= \sum_n \frac{t_n}{\sigma(\phi(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w})} \sigma(\phi(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w})(1 - \sigma(\phi(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w}))\phi(\boldsymbol{x}_n) \\
&\quad + \frac{10 - t_n}{1 - \sigma(\phi(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w})} (-\sigma(\phi(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w})(1 - \sigma(\phi(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w})))\phi(\boldsymbol{x}_n) \\
&= \sum_n t_n(1 - \sigma(\phi(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w}))\phi(x_n) - (10 - t_n)\sigma(\phi(\boldsymbol{x}_n)^\mathsf{T}\boldsymbol{w})\phi(\boldsymbol{x}_n).
\end{aligned}$$

Further simplification is possible but unnecessary.