SLDM II - Homework 2

Matt Isaac

October 25, 2018

1. Linear Algebra Review

a. Show that if U is an orthogonal matrix, then for all $\mathbf{x} \in \mathbb{R}$, $||\mathbf{x}|| = ||\mathbf{Ux}||$, where $||.||$ indicates the Euclidean norm.

By definition of the Euclidean norm, we begin with

$$||x|| = \sqrt{x^T x}$$

Then, since U is an orthogonal matrix, and $U^T U = U^{-1} U = I$,

$$||x|| = \sqrt{x^T U^T U x}.$$

If we remember the fact that $A^T B^T = (BA)^T$, we can see that

$$||x|| = \sqrt{(Ux)^T (Ux)},$$

which implies that

$$||x|| = ||Ux||$$

b. Show that all 2×2 orthogonal matrices have the form:

$$\begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

or

$$\begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}$$

Let U be a 2×2 orthogonal matrix, such that such that $U = \begin{bmatrix} \mathbf{u_1 u_2} \end{bmatrix}$, where $\mathbf{u_i}$ is the ith column of U. Because U is orthogonal, $\mathbf{u_1}^T \mathbf{u_1} = 1$ and $\mathbf{u_1}^T \mathbf{u_2} = 0$. Since $\mathbf{u_1}^T \mathbf{u_1} = ||\mathbf{u_1}|| = 1$, we know that $\mathbf{u}_i$ lies on the unit circle. Thus, $\mathbf{u_1} = \begin{bmatrix} \cos\theta \\ \sin\theta \end{bmatrix}$, and $\mathbf{u_2} = \begin{bmatrix} \cos\tilde{\theta} \\ \sin\tilde{\theta} \end{bmatrix}$, where $\tilde{\theta} = \theta + \frac{\pi}{2}$. So, if we find the possible values of the elements of $\mathbf{u_2}$, we see that $\mathbf{u_2} = \begin{bmatrix} -\sin\theta \\ \cos\theta \end{bmatrix}$ or $\mathbf{u_2} = \begin{bmatrix} \sin\theta \\ -\cos\theta \end{bmatrix}$. Thus,

$$U = \begin{bmatrix} \mathbf{u_1 u_2} \end{bmatrix} = \begin{bmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{bmatrix}$$

or

$$\begin{bmatrix} \cos\theta & \sin\theta \\ \sin\theta & -\cos\theta \end{bmatrix}.$$

2. Probability

a.

i. $\mathbf{E[X] = E_Y[E_X[X|Y]]}$

$$E_Y[E_X[X|Y]] = E\left[\int\limits_x xPr(X=x|Y=y)dx\right]$$

$$\implies E_Y[E_X[X|Y]] = \int\limits_y \int\limits_x yxPr(X=x|Y=y)P(Y=y)dxdy$$

$$\implies E_Y[E_X[X|Y]] = \int\limits_y \int\limits_x yxPr(X=x, Y=y)dxdy$$

$$\implies E_Y[E_X[X|Y]] = \int\limits_x xPr(X=x)dx$$

$$\implies E_Y[E_X[X|Y]] = E[X]$$

ii. $\mathbf{E[1[X \in C]] = Pr(X \in C)}$, where $\mathbf{1}[X \in C]$ is the indicator function of an arbitrary set C.
By definition,

$$E(X) = \sum_x xp(x),$$

$$Pr(\mathbf{1}[X \in C] = 1) = Pr(X \in C)$$

and

$$Pr(\mathbf{1}[X \in C] = 0) = 1 - Pr(X \in C).$$

Then in this case, because our indicator function results in a discrete random variable (i.e. $\mathbf{1}[X \in C] \in \{0,1\}$),

$$E(1[X \in C]) = (0)(1 - Pr(X \in C)) + (1)(Pr(X \in C)) = Pr(X \in C)$$

.

iii. If X and Y are independent, then $E[XY] = E[X]E[Y]$

By definition, if X and Y are independent, $p(x,y) = p(x)p(y)$. Also, recall that for a continuous random variable, $E(X) = \int\limits_x xp(x)dx$. Expanding this to $E[XY]$, we see that

$$E[XY] = \int\limits_x \int\limits_y xy \cdot p(xy)dxdy$$

Since X and Y are independent,

$$E[XY] = \int\limits_x \int\limits_y xy \cdot p(x)p(y)dxdy$$

which can be rearranged as

$$E[XY] = \int\limits_x xp(x)dx \int\limits_y yp(y)dy = E[X]E[Y].$$

Thus, if X and Y are independent,

$$E[XY] = E[X]E[Y].$$

b. For the following equations, describe the relationship between them. Write one of four answers to replace the question mark: "=", "≤", "≥" or "depends".

i. $\mathbf{Pr(X = x, Y = y)}$? $\mathbf{Pr(X = x)}$.
The event $(X = x, Y = y)$ is either more restrictive, or just as restrictive as the event $(X = x)$. If it is more restrictive, then $Pr(X = x, Y = y) < Pr(X = x)$. If $(X = x, Y = y)$ is just as restrictive as $(X = x)$, that is if $x \subseteq y$, then $Pr(X = x, Y = y) = Pr(X = x)$. Combining these two cases, $Pr(X = x, Y = y) \leq Pr(X = x)$.

ii. $\mathbf{Pr(X = x|Y = y)}$? $\mathbf{Pr(X = x)}$.
First, recall that $Pr(X = x|Y = y) = \frac{Pr(X=x,Y=y)}{Pr(Y=y)}$.
Also, recall from (i) that $Pr(X = x, Y = y) \leq Pr(X = x)$. By the same reasoning, $Pr(X = x, Y = y) \leq Pr(Y = y)$. It follows, then that $\frac{Pr(X=x,Y=y)}{Pr(Y=y)}$ will result in a value larger than $Pr(X = x|Y = y)$, since $0 < P(Y = y) \leq 1$. However, we don't know whether $\frac{Pr(X=x,Y=y)}{Pr(Y=y)}$ will be greater than or less than $P(X = x)$. So, we conclude that it "depends".

iii. $\mathbf{Pr(X = x|Y = y)}$? $\mathbf{Pr(Y = y|X = x)Pr(X = x)}$.
First, notice that
$Pr(X = x|Y = y) = \frac{Pr(X=x,Y=y)}{Pr(Y=y)}$ and $Pr(Y = y|X = x)Pr(X = x) = \frac{Pr(X=x,Y=y)}{Pr(X=x)}Pr(X = x) = Pr(X = x, Y = y)$. Thus, the question is really $\frac{Pr(X=x,Y=y)}{Pr(Y=y)}$? $Pr(X = x, Y = y)$. Once this is realized it is easy to see that, if $Pr(Y = y) = 1$, $\frac{Pr(X=x,Y=y)}{Pr(Y=y)} = Pr(X = x, Y = y)$. Otherwise, if $Pr(Y = y) < 1$, $\frac{Pr(X=x,Y=y)}{Pr(Y=y)} > Pr(X = x, Y = y)$. Combining both cases, we come to the conclusion that $\frac{Pr(X=x,Y=y)}{Pr(Y=y)} \geq Pr(X = x, Y = y)$.

3. Positive (semi-)definite matrices.

Let A be a real, symmetric $d \times d$ matrix. We say A is positive semi-definite (PSD) if for all $x \in \mathbb{R}^d, x^T Ax \geq 0$.
A is positive definite (PD) if for all $x \neq 0$, $X^T Ax > 0$.
The spectral theorem says that every real symmetric matrix A can be expressed via the spectral decomposition

$$A = U\Lambda U^T$$

where U is a $d \times d$ orthogonal matrix and $\Lambda = diag(\lambda_1, ..., \lambda_d)$.
Using the spectral decomposition, show the following:

a. If $\mathbf{u_i}$ is the ith column of U then $\mathbf{u_i}$ is an eigenvector of A with corresponding eigenvalue λ_i.

We begin with the spectral decomposition:

$$A = U\Lambda U^T$$

Then, because U is orthogonal, and by definition $U^{-1} = U^T \implies U^T U = U^{-1}U = I$,

$$AU = U\Lambda U^T U \implies AU = U\Lambda.$$

Showing the value of the entries, we obtain the following matrices:

$$\begin{bmatrix} a_{11} & \cdots & a_{1d} \\ \vdots & \ddots & \vdots \\ a_{d1} & \dots & a_{dd} \end{bmatrix} \begin{bmatrix} u_{11} & \dots & u_{1d} \\ \vdots & \ddots & \vdots \\ u_{d1} & \dots & u_{dd} \end{bmatrix} = \begin{bmatrix} u_{11} & \dots & u_{1d} \\ \vdots & \ddots & \vdots \\ u_{d1} & \dots & u_{dd} \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \lambda_d \end{bmatrix}$$

Multiplying the matrices together, we get

$$\begin{bmatrix} a_{11}u_{11} + ... + a_{1d}u_{d1} & \cdots & a_{11}u_{1d} + ... + a_{1d}u_{dd} \\ \vdots & \ddots & \vdots \\ a_{d1}u_{11} + ... + a_{dd}u_{d1} & \cdots & a_{d1}u_{1d} + ... + a_{dd}u_{dd} \end{bmatrix} = \begin{bmatrix} u_{11}\lambda_1 & \dots & u_{1d}\lambda_d \\ \vdots & \ddots & \vdots \\ u_{d1}\lambda_1 & \dots & u_{dd}\lambda_d \end{bmatrix}$$

Upon inspection, we can see that column 1 of the matrix on the lefthand side is the vector resulting from the calculation $A\mathbf{u_1}$ (i.e. matrix A times $\mathbf{u_1}$, the first column of matrix U). This equals the first column of the righthand side, which is $\lambda_1\mathbf{u_1}$ (i.e. λ_1 times the first column of matrix U). This pattern continues until the last index, d, and we see that columnd d of the lefthand side matrix is the vector resulting from $A\mathbf{u_d}$, which equals the last column of the matrix on the righthand side, $\lambda_d\mathbf{u_d}$. Generalizing this pattern, we can see that for the ith column of U, $A\mathbf{u_i} = \lambda_\mathbf{i}\mathbf{u_i}$. This is the definition of the relationship between an eigen vector $\mathbf{u_i}$ and the eigen value λ_i.

b. A is PSD iff $\lambda_i \geq 0$ for each i.

By spectral decomposition, we see that

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T U\Lambda U^T\mathbf{x}$$

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i \mathbf{x}^T\mathbf{u}_i\mathbf{u}_i^T\mathbf{x}$$

Note that $\mathbf{x}^T u_i$ and $u_i^T\mathbf{x}$ will both end up being scalars, so actually, $\mathbf{x}^T u_i = u_i^T\mathbf{x}$. Thus, we can write:

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i(\mathbf{x}^T\mathbf{u}_i)^2.$$

Since $(\mathbf{x}^T\mathbf{u}_i)^2$ will always be greater than or equal to 0, $\mathbf{x}^T A\mathbf{x}$ will be ≥ 0 if the λ_is are ≥ 0

That is, $\mathbf{x}^T A\mathbf{x} > 0$ if $\lambda_i > 0$ for each i.

For the 'other direction' of the proof, we will begin with the assumption that A is positive-semi definite. We will also assume, by way of contradiction, and withouth loss of generality, that $\lambda_1 < 0$. Again, as shown previously,

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i(\mathbf{x}^T\mathbf{u}_i)^2$$

Since the definition of a positive definite matrix holds for any $\mathbf{x}$, we will choose to let $\mathbf{x} = u_1$. We then have

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i(\mathbf{u}_1^T\mathbf{u}_i)^2$$

Recall that U is an orthogonal matrix, thus $U^{-1} = U^T \implies UU^T = I$. Also recall that $\mathbf{u_i}$ is the ith column of U. This means that $\mathbf{u}_i^T\mathbf{u}_j = 1$ if $i = j$ and 0 otherwise.
Returning to our equation, we see that when we expand out the summation on the righthand side, using the fact just stated (that $\mathbf{u}_i^T\mathbf{u}_j = 1$ if $i = j$ and 0 otherwise), we see that

$$\mathbf{x}^T A\mathbf{x} = \lambda_1 + 0 + \cdots + 0$$

We have reached a contradiction at this point, because we began by assuming that A was PSD, and that $\lambda_1 < 0$. However, we found an $\mathbf{x}$ that led to $\mathbf{x}^T A\mathbf{x} < 0$, which is a contradiction.

Thus, $\lambda_i \geq 0$ for each i if $\mathbf{x}^T A\mathbf{x} \geq 0$.

So overall, we conclude that $\mathbf{x}^T A\mathbf{x} \geq 0 \iff \lambda_i \geq 0$ for each i.

c. A is PD iff $\lambda_i > 0$ for each i.

By spectral decomposition, we see that

$$\mathbf{x}^T A\mathbf{x} = \mathbf{x}^T U\Lambda U^T\mathbf{x}$$

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i \mathbf{x}^T \mathbf{u}_i \mathbf{u}_i^T \mathbf{x}$$

Note that $\mathbf{x}^T u_i$ and $u_i^T\mathbf{x}$ will both end up being scalars, so actually, $\mathbf{x}^T u_i = u_i^T\mathbf{x}$. Thus, we can write:

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i (\mathbf{x}^T \mathbf{u}_i)^2.$$

Since $(\mathbf{x}^T\mathbf{u}_i)^2$ will always be greater than 0, the sign of $\mathbf{x}^T A\mathbf{x}$ will depend on the sign of λ_i being positive.

That is, $\mathbf{x}^T A\mathbf{x} > 0$ if $\lambda_i > 0$ for each i.

For the 'other direction' of the proof, we will begin with the assumption that A is positive definite. We will also assume, by way of contradiction, and withouth loss of generality, that $\lambda_1 < 0$. Again, as shown previously,

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i (\mathbf{x}^T \mathbf{u}_i)^2$$

Since the definition of a positive definite matrix holds for any $\mathbf{x}$, we will choose to let $\mathbf{x} = u_1$. We then have

$$\mathbf{x}^T A\mathbf{x} = \sum_{i=1}^{d} \lambda_i (\mathbf{u}_1^T \mathbf{u}_i)^2$$

Recall that U is an orthogonal matrix, thus $U^{-1} = U^T \implies UU^T = I$. Also recall that $\mathbf{u_i}$ is the ith column of U. This means that $\mathbf{u}_i^T\mathbf{u}_j = 1$ if $i = j$ and 0 otherwise.
Returning to our equation, we see that when we expand out the summation on the righthand side, using the fact just stated (that $\mathbf{u}_i^T\mathbf{u}_j = 1$ if $i = j$ and 0 otherwise), we see that

$$\mathbf{x}^T A\mathbf{x} = \lambda_1 + 0 + \cdots + 0$$

We have reached a contradiction at this point, because we began by assuming that A was PD, and that $\lambda_1 < 0$. However, we found an $\mathbf{x}$ that led to $\mathbf{x}^T A\mathbf{x} < 0$, which is a contradiction.

Thus, $\lambda_i > 0$ for each i if $\mathbf{x}^T A\mathbf{x} > 0$.

So overall, we conclude that $\mathbf{x}^T A\mathbf{x} > 0 \iff \lambda_i > 0$ for each i.

4. The Bayes Classifier

Let X be a random variable representing a 1-dimensional feature space and let Y be a discrete random variable taking values in $\{0, 1\}$. If $Y = 0$, then the posterior distribution of X for class 0 is Gaussian with mean μ_0 and variance σ_0^2. If $Y = 1$, then the posterior distribution of X for class 1 is Gaussian with mean μ_1 and variance σ_1^2. Let $w_0 = Pr(Y = 0)$ and $w_1 = Pr(Y = 1) = 1 - w_0$.

a. Derive the Bayes classifier for this problem as a function of w_i, μ_i, and σ_i where $i \in \{0, 1\}$.

$$\pi_0 p_0(x) = \pi_1 p_1(x)$$

$$w_0\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}e^{\frac{-(x-\mu_0)^2}{2\sigma_0^2}}\right)=w_1\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}}\right)$$

$$\frac{w_0}{w_1}\left(\frac{e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{e^{\frac{(x-\mu_0)^2}{2\sigma_0^2}}}\right)=\left(\frac{\sqrt{2\pi}\sigma_0}{\sqrt{2\pi}\sigma_1}\right)$$

$$\log\left(\frac{e^{\frac{(x-\mu_1)^2}{2\sigma_1^2}}}{e^{\frac{(x-\mu_0)^2}{2\sigma_0^2}}}\right)=\log\left(\frac{\sigma_0 w_1}{\sigma_1 w_0}\right)$$

$$\left(\frac{(x-\mu_1)^2}{2\sigma_1^2}-\frac{(x-\mu_0)^2}{2\sigma_0^2}\right)=\log\left(\frac{\sigma_0 w_1}{\sigma_1 w_0}\right)$$

$$\sigma_0^2(x-\mu_1)^2-\sigma_1^2(x-\mu_0)^2=\sigma_0^2\sigma_1^2 2\log\left(\frac{\sigma_0 w_1}{\sigma_1 w_0}\right)$$

$$\sigma_0^2(x^2-2\mu_1 x+\mu_1^2)-\sigma_1^2(x^2-2\mu_0 x+\mu_0^2)=\sigma_0^2\sigma_1^2 2\log\left(\frac{\sigma_0 w_1}{\sigma_1 w_0}\right)$$

$$(\sigma_0^2-\sigma_1^2)x^2+(-2\mu_1\sigma_0^2+2\mu_0\sigma_1^2)x=\sigma_0^2\sigma_1^2 2\log\left(\frac{\sigma_0 w_1}{\sigma_1 w_0}\right)-\mu_1^2\sigma_0^2+\mu_0^2\sigma_1^2$$

We now need to consider two cases:
If $\sigma_0^2=\sigma_1^2$, then we get

$$x=\frac{\sigma_0^2\sigma_1^2 2\log\left(\frac{\sigma_0 w_1}{\sigma_1 w_0}\right)-\mu_1^2\sigma_0^2+\mu_0^2\sigma_1^2}{2\mu_0\sigma_1^2-2\mu_1\sigma_0^2}.$$

Otherwise, we have

$$x=\frac{-(2\mu_0\sigma_1^2-2\mu_1\sigma_0^2)\pm\sqrt{(2\mu_0\sigma_1^2-2\mu_1\sigma_0^2)^2-4(\sigma_0^2-\sigma_1^2)\left(\mu_1^2\sigma_0^2-\mu_0^2\sigma_1^2-\sigma_0^2\sigma_1^2 2\log\left(\frac{\sigma_0 w_1}{\sigma_1 w_0}\right)\right)}}{2(\sigma_0^2-\sigma_1^2)}$$

b. Derive the Bayes error rate for this classification problem as a function of $w_i, \mu_i, and \sigma_i$ where $i \in \{0,1\}$. You may write solution in terms of Q where if Z is a standard normal random variable, then $Q(z)=Pr(Z>z)$.

Assume $\mu_0<\mu_1$, and let b be the cutoff value such that observations $>b$ are classified as $Y=1$ and observations $<b$ are classified as $Y=0$. Mathematically, we would right the Bayes classifier as

$$f(x)=\begin{cases}0 & x\le b\\ 1 & x>b\end{cases}.$$

The Bayes error rate can then be thought of as the probability that we get an observation greater than our cutoff that has $Y=0$ plus the probability that we get an observation less than our cutoff that has $Y=1$. In other words, the Bayes error Rate for Bayes classifier f is

$$R(f)=w_0p_0(X>b)+w_1p_1(X<b)$$

$$\implies R(f)=w_0p_0\left(Z>\frac{b-\mu_0}{\sigma_0}\right)+w_1\left(1-p_1\left(Z>\frac{b-\mu_1}{\sigma_1}\right)\right)$$

$$\implies R(f)=w_0Q_0\left(\frac{b-\mu_0}{\sigma_0}\right)+w_1\left(1-Q_1\left(\frac{b-\mu_1}{\sigma_0}\right)\right)$$

So, the Bayes classification error rate is:

$$R(f)=w_0Q_0\left(\frac{b-\mu_0}{\sigma_0}\right)+w_1\left(1-Q_1\left(\frac{b-\mu_1}{\sigma_0}\right)\right)$$

c. Describe how to perform cross-validation for a classification problem.

Since it was not specified in the prompt, I will describe k-fold cross-validation.
The first step is to randomly divide your training data into k_i groups. These groups should all be about the same size. Next, the k_1 group is set apart from the rest of the $k-1$ groups. The classifier in question is fit to (i.e. trained on) the $k-1$ groups, and then used to predict the classes of the observation in the k_1 group. The misclassification rate, E_1 is then calculated. This process is repeated for the rest of the $k-1$ groups; each taking a turn 'sitting out' during the training, and then acting as the validation data. At the end of the process, k misclassification rates should be obtained. The cross-validated misclassification rate is simply the average of the k misclassification rates: $\frac{1}{k}\sum_{i=1}^{k} E_i$.
It can be noted that Leave One Out Cross Validation (LOOCV) is performed when k equals the number of observations ($k = n$).

d. Set the following values: $\mu_0 = 0, \mu_1 = 1.5, \sigma_0 = \sigma_1 = 1, w_0 = 0.3, w_1 = 0.7$. Simulate the classification problem 100 times for $N \in \{100, 200, 500, 1000\}$. Apply the Bayes classifier logistic regression, and the k-nearest neighbor classifer to the data.

i. Bayes error rate: Theoretically, this can be calcualted to be 0.19396. However, for each level of N, in our simulation, we obtained:

N	Bayes Error Rate
100	0.20
200	0.19
500	0.19
1000	0.19

ii. Average value of k:

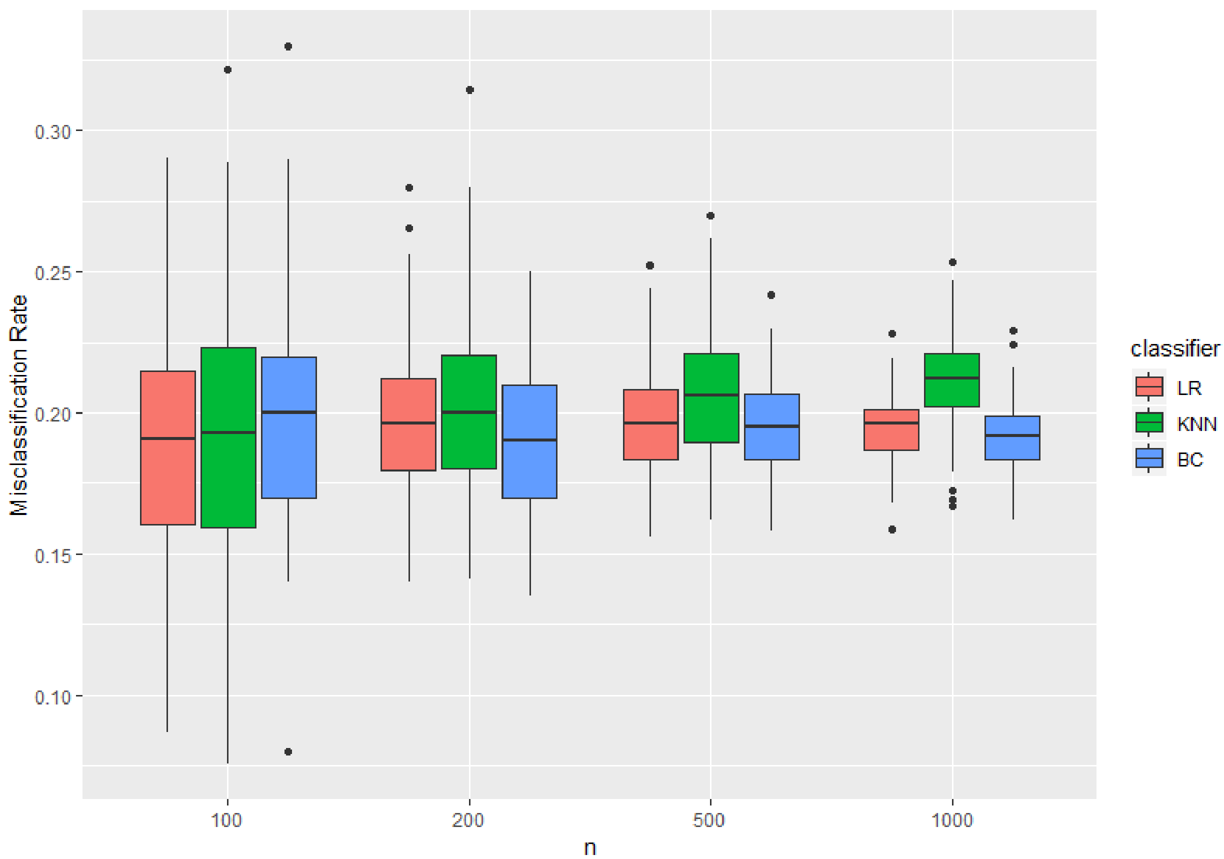
N	k
100	7.22
200	7.68
500	7.62
1000	8.16

iii. Classification error of each classifier in both table and graphical form. Describe how you performed cross validation.

10-fold cross validation was used.
Table of (mean, standard deviation) for each classifier and level of N:

N	k Nearest Neighbors	Logistic Regression	Bayes Classifier
100	(0.19, 0.047)	(0.19, 0.040)	(0.20, 0.040)
200	(0.20, 0.033)	(0.20, 0.027)	(0.19, 0.028)
500	(0.21, 0.022)	(0.20, 0.019)	(0.19, 0.017)
1000	(0.21, 0.016)	(0.19, 0.013)	(0.19, 0.013)

Plot shown on next page (Figure 1).



0.30
0.25
0.20
0.15
0.10
Misclassification Rate
100
200
500
1000
n
classifier
LR
KNN
BC

Figure 1:

5. How long did this assignment take you?

Somewhere around 20 hours.

6. Type up homework solutions:

Check.