Stat 6910, Section 003
Statistical Learning and Data Mining II
Fall 2018

Homework 2

Jared Hansen

A-number: A01439768

e-mail: jrdhansen@gmail.com

Stat 6910 Statistical Learning and Data Mining II	Fall 2018
Homework Assignment 2	
100 Points	Due Friday 10/26/2018 (via Canvas by 5:00 pm)

1. Linear Algebra Review (10 pts)

   (a) (5 pts) Show that if U is an orthogonal matrix, then for all $\mathbf{x} \in \mathbb{R}^d$, $\|\mathbf{x}\| = \|U\mathbf{x}\|$, where $\|\cdot\|$ indicates the Euclidean norm.

   We'll make use of the following properties

   - Since U is orthogonal, by definition $U^TU = UU^T = \mathbb{I}$
   - $\|\mathbf{x}\| = (\mathbf{x}^T\mathbf{x})^{1/2}$

   From the definition of $\|\cdot\|$, $\|U\mathbf{x}\| = ((U\mathbf{x})^T(U\mathbf{x}))^{1/2} = (\mathbf{x}^TU^TU\mathbf{x})^{1/2}$. Since we know that $U^TU = \mathbb{I}$, $(\mathbf{x}^TU^TU\mathbf{x})^{1/2} = (\mathbf{x}^T\mathbb{I}\mathbf{x})^{1/2} = (\mathbf{x}^T\mathbf{x})^{1/2} = \|\mathbf{x}\|$ by definition. Therefore, $\|\mathbf{x}\| = \|U\mathbf{x}\|$ $\forall \mathbf{x} \in \mathbb{R}^d$ when U is an orthogonal matrix.

   (b) (5 pts) Show that all 2×2 orthogonal matrices have the form
   $\begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}$ *or* $\begin{bmatrix} cos\theta & sin\theta \\ sin\theta & -cos\theta \end{bmatrix}$

   Since all row vectors and column vectors $\begin{bmatrix} x_i \\ y_i \end{bmatrix}$ of A have unit length and are in $\mathbb{R}^2$, each vector can be thought of as a vector with tail at $(0, 0)$ and head on the unit circle.
   Therefore, $\exists$ some angle θ whose cosine $= x_i$ and whose sine $= y_i$. Instead of describing a row or column vector of A as $\begin{bmatrix} x_i \\ y_i \end{bmatrix}$ we may describe it as $\begin{bmatrix} cos\theta \\ sin\theta \end{bmatrix}$.
   Let column 1 of A be $\begin{bmatrix} cos\theta \\ sin\theta \end{bmatrix}$. In order for A to remain an orthogonal matrix, the remaining column vector must be orthogonal to $\begin{bmatrix} cos\theta \\ sin\theta \end{bmatrix}$. Let the angle of this other vector be α. For two angle to be orthogonal to one another, they

need to have a difference of $\pi/2$ radians. In this case, $\alpha = \theta \pm \pi/2$.
This allows column 2 of A to be:

- $\begin{bmatrix} cos(\alpha) \\ sin(\alpha) \end{bmatrix} = \begin{bmatrix} cos(\theta + \pi/2) \\ sin(\theta + \pi/2) \end{bmatrix} = \begin{bmatrix} -sin(\theta) \\ cos(\theta) \end{bmatrix}$
- $\begin{bmatrix} cos(\alpha) \\ sin(\alpha) \end{bmatrix} = \begin{bmatrix} cos(\theta - \pi/2) \\ sin(\theta - \pi/2) \end{bmatrix} = \begin{bmatrix} sin(\theta) \\ -cos(\theta) \end{bmatrix}$

Thus, the only two forms of A that are possible are $\begin{bmatrix} cos\theta & -sin\theta \\ sin\theta & cos\theta \end{bmatrix}$ *or* $\begin{bmatrix} cos\theta & sin\theta \\ sin\theta & -cos\theta \end{bmatrix}$

2. Probability (18 pts)

(a) (9 pts) Let random variables X and Y be jointly continuous with pdf $p(x, y)$. Prove the following results:

i. $\mathbb{E}[X] = \mathbb{E}_Y[\mathbb{E}_X[X|Y]]$ where $\mathbb{E}_Y$ is the expectation with respect to Y.

$E_Y[E_X[X|Y]] = \int_Y E_X[X|Y] \cdot p(y)\,dy = \int_Y \int_X \left(x \cdot \frac{p(x,y)}{p(y)}\,dx\right) p(y)\,dy$ since $p(x|y) = \frac{p(x,y)}{p(y)}$.

Since $p(y)$ is not dependent on x, we can rearrange as $\int_Y \int_X x \cdot p(x,y)\,dx \cdot \frac{p(y)}{p(y)}\,dy = \int_Y \int_X x \cdot p(x,y)\,dx\,dy$

By defn: $\int_Y p(x,y)\,dy = p(x)$

Now we apply Fubini's Theorem to rearrange, and have $\int_X \int_Y x \cdot p(x,y)\,dy\,dx$

Since x is not dependent on y we can rearrange as $\int_X x \left(\int_Y p(x,y)\,dy\right) dx$

We're left with $\int_X x \cdot p(x)\,dx = E[X]$ by definition. Thus, we've shown that $E[X] = E_Y[E_X[X|Y]]$ and proven the desired result.

ii. $\mathbb{E}[\mathbf{1}[X \in C]] = Pr(X \in C)$ where $\mathbf{1}[X \in C]$ is the indicator function of an arbitrary set C. That is, $\mathbf{1}[X \in C] = 1$ if $X \in C$ and 0 otherwise.

For a discrete random variable X, $E(X) = \sum_{i=1}^{n} x_i \cdot p(x_i)$

By definition $P(\mathbf{1}[X \in C] = 1) = P(X \in C)$ and by the complement rule $P(\mathbf{1}[X \in C] = 0) = 1 - P(X \in C)$.

Applying the expected value definition to $\mathbf{1}[X \in C]$: $E[\mathbf{1}[X \in C]] = (1)(P(X \in C)) + (0)(1 - P(X \in C))$

Giving $E[\mathbf{1}[X \in C]] = P(X \in C)$ and proving the desired result.

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

By definition, if X and Y are independent then $p(x,y) = p(x)p(y)$.

$E(XY) = \int_X \int_Y xy \cdot p(x,y)\,dy\,dx = \int_X \int_Y x \cdot y \cdot p(x) \cdot p(y)\,dy\,dx$. Since X and Y aren't dependent on each other we can rearrange and get

$E(XY) = \left(\int_X x \cdot p(x)\,dx\right)\left(\int_Y y \cdot p(y)\,dy\right) = E(X) \cdot E(Y)$. Thus we've proven the desired result. Similar proof holds for discrete random variables by using definitions involving sums instead of integrals.

(b) (9 pts) For the following equations, describe the relationship between them. Write one of four answers to replace the question mark: "=", "≤", "≥", "depends". Choose the most specific relation that always holds and briefly explain why. Assume all probabilities are non-zero.

i. $Pr(X = x, Y = y) \boxed{\leq} Pr(X = x)$

ii. $Pr(X = x|Y = y) \boxed{depends} Pr(X = x)$

iii. $Pr(X = x|Y = y) \boxed{\geq} Pr(Y = y|X = x)Pr(X = x)$

(i.) If we think of $P(X=x)$ and $P(Y=y)$ as constraints, we know that imposing the additional constraint of $P(Y=y)$ to $P(X=x)$ can result only in a lower probability if $y \notin x$. However if $y \subset x$ then $P(X=x, Y=y) = P(X=x)$ since all events $\in Y$ are also $\in X$.

(ii.) If X and Y are independent, then $P(X|Y) = P(X)$ by definition.

If X and Y are not independent: we know $P(X|Y) = \frac{P(X,Y)}{P(Y)}$ → rewrite original expression as $\frac{P(X,Y)}{P(Y)}$? $P(X)$.

We know from 2b(i.) that $P(X,Y) \leq P(X)$, so we can rewrite again as $\frac{\text{"} \leq P(X) \text{"}}{P(Y)}$? $P(X)$.

Therefore, the wiggle room in the "$\leq P(X)$" makes the answer "depends".

Examples: $\left[\frac{0.3}{0.7} = 0.429\right] > 0.3$, $\left[\frac{0.1}{0.7} = 0.143\right] < 0.3$, $\left[\frac{0.2}{0.4} = 0.5\right] = 0.5$

(iii.) 1: X and Y are independent: if X and Y are independent then we have $P(X)$? $P(Y) \cdot P(X)$.

Since $0 \leq P(Y) \leq 1$ and $0 \leq P(X) \leq 1$ we know $P(X) = P(X)P(Y)$ if $P(Y)=1$, and $P(X) > P(X)P(Y)$ otherwise.

2: X and Y not independent: Since $P(X|Y) = P(X,Y)/P(Y)$ and $P(Y|X) = P(Y,X)/P(X) \Rightarrow P(Y,X) = P(Y|X)P(X)$.

We can rewrite the original relation as $\frac{P(X,Y)}{P(Y)}$? $P(Y,X)$. Since $P(X,Y) = P(Y,X)$ this implies that the sign must be $\geq$. If $P(Y)=1$ then the two expressions are equal, but if $P(Y) < 1$ then the quantity $\frac{P(X,Y)}{P(Y)}$ will be larger than $P(X,Y)$. (This is a property of probabilities $\in (0,1)$.)

3. Positive (semi-) definite matrices

(a.) If $\underset{\sim}{u}_i$ is the i^{th} column of U, then $\underset{\sim}{u}_i$ is an eigenvector of A with corresponding eigenvalue λ_i.

$A = U \Lambda U^T$ (given info) $\longrightarrow A\,\underset{\sim}{u}_i = U \Lambda U^T \underset{\sim}{u}_i \longrightarrow$

$$\begin{bmatrix} \underset{\sim}{u}_1 \rightarrow \\ \vdots \\ \underset{\sim}{u}_i \rightarrow \\ \vdots \\ \underset{\sim}{u}_d \rightarrow \end{bmatrix} \begin{bmatrix} \\ \\ \underset{\sim}{u}_i \\ \\ \\ \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ \underset{\sim}{u}_i^T \underset{\sim}{u}_i \\ \vdots \\ 0 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i^{th} \text{ row}$$

Now we have $A\,\underset{\sim}{u}_i = U \Lambda \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i^{th}$ row $\longrightarrow A\,\underset{\sim}{u}_i = U \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & & \vdots \\ \vdots & & \lambda_i & 0 \\ 0 & \cdots & 0 & \lambda_d \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \leftarrow i^{th}$ row

Giving $A\,\underset{\sim}{u}_i = U \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_i \\ 0 \\ \vdots \\ 0 \end{bmatrix} \longrightarrow A\,\underset{\sim}{u}_i = \begin{bmatrix} \uparrow & \uparrow & & \uparrow & & \uparrow \\ \underset{\sim}{u}_1 & \underset{\sim}{u}_2 & \cdots & \underset{\sim}{u}_i & \cdots & \underset{\sim}{u}_d \\ \downarrow & \downarrow & & \downarrow & & \downarrow \end{bmatrix} \begin{bmatrix} 0 \\ \vdots \\ 0 \\ \lambda_i \\ 0 \\ \vdots \\ 0 \end{bmatrix}$

$$\rightarrow A\,\underset{\sim}{u}_i = \begin{bmatrix} (u_{1,1})(0) + (u_{1,2})(0) + \ldots + (u_{1,i})(\lambda_i) + (u_{1,i+1})(0) + \ldots + (u_{1,d})(0) \\ (u_{2,1})(0) + (u_{2,2})(0) + \ldots + (u_{2,i})(\lambda_i) + (u_{2,i+1})(0) + \ldots + (u_{2,d})(0) \\ \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \qquad\qquad \vdots \\ (u_{d,1})(0) + (u_{d,2})(0) + \ldots + (u_{d,i})(\lambda_i) + (u_{d,i+1})(0) + \ldots + (u_{d,d})(0) \end{bmatrix}$$

$$\rightarrow A\,\underset{\sim}{u}_i = \begin{bmatrix} (u_{1,i})(\lambda_i) \\ (u_{2,i})(\lambda_i) \\ \vdots \\ (u_{d,i})(\lambda_i) \end{bmatrix} \longrightarrow A\,\underset{\sim}{u}_i = \lambda_i \begin{bmatrix} (u_{1,i}) \\ (u_{2,i}) \\ \vdots \\ (u_{d,i}) \end{bmatrix} \longrightarrow A\,\underset{\sim}{u}_i = \lambda_i\,\underset{\sim}{u}_i$$

Since $A\,\underset{\sim}{u}_i = \lambda_i\,\underset{\sim}{u}_i$ by employing the spectral decomposition (and algebraic manipulation) this implies that $\underset{\sim}{u}_i$ is an eigenvector of A with corresponding eigenvalue λ_i, proving the desired result.

3b. Show that A is PSD iff $\lambda_i \geq 0$ for each i.

Using $A = U \Lambda U^T$ rewrite $\underset{\sim}{x}^T A \underset{\sim}{x}$ as $\underset{\sim}{x}^T U \Lambda U^T \underset{\sim}{x}$. Now using hint $U \Lambda U^T = \sum_{i=1}^{d} \lambda_i \underset{\sim}{u}_i \underset{\sim}{u}_i^T$ we can rewrite again as $\sum_{i=1}^{d} \lambda_i \underset{\sim}{x}^T \underset{\sim}{u}_i \underset{\sim}{u}_i^T \underset{\sim}{x}$. We know that $(\underset{\sim}{x}^T \underset{\sim}{u}_i)$ and $(\underset{\sim}{u}_i^T \underset{\sim}{x})$ are scalars.

So we can rewrite again as $\sum_{i=1}^{d} \lambda_i (\underset{\sim}{x}^T \underset{\sim}{u}_i)^2$. Since the term $(\underset{\sim}{x}^T \underset{\sim}{u}_i)$ is squared, it must be positive.

Therefore, $\sum_{i=1}^{d} \lambda_i (\underset{\sim}{x}^T \underset{\sim}{u}_i)^2$ is only ≥ 0 if $\lambda_i \geq 0 \; \forall i$.

Thus, we've shown A is PSD "if" $\lambda_i \geq 0 \; \forall i$.

Now let's show that A is PSD "and only if" $\lambda_i \geq 0 \; \forall i$.

Let $\underset{\sim}{x} = \underset{\sim}{u}_1$, giving $U^T \underset{\sim}{u}_1 = \|\underset{\sim}{u}_1\|^2$ (since $\underset{\sim}{u}_i^T \underset{\sim}{u}_1 = 0 \; \forall i \neq 1$).

Let $\lambda_1 < 0$. By the spectral decomposition, $\underset{\sim}{x}^T A \underset{\sim}{x} = \sum_{i=1}^{d} \lambda_i (\underset{\sim}{x}^T \underset{\sim}{u}_i)^2$. Since we let $\underset{\sim}{x} = \underset{\sim}{u}_1 \Rightarrow \underset{\sim}{x}^T = \underset{\sim}{u}_1^T$ giving $\sum_{i=1}^{d} \lambda_i (\underset{\sim}{u}_1^T \underset{\sim}{u}_i)^2 = \lambda_1 (\underset{\sim}{u}_1^T \underset{\sim}{u}_1)^2 + \lambda_2 (\underset{\sim}{u}_1^T \underset{\sim}{u}_2)^2 + \ldots + \lambda_d (\underset{\sim}{u}_1^T \underset{\sim}{u}_d)^2 = \lambda_1 (\|\underset{\sim}{u}_1\|)^2 + \lambda_2(0) + \ldots + \lambda_d(0)$

$= \lambda_1 (\|\underset{\sim}{u}_1\|)^2$. Assuming $\underset{\sim}{u}_1 \neq \underset{\sim}{0}$ this implies that $(\|\underset{\sim}{u}_1\|)^2 > 0$.

Since we let $\lambda_1 < 0 \Rightarrow \lambda_1$ is negative, and since $(\|\underset{\sim}{u}_1\|)^2$ is positive, the quantity $\lambda_1 (\|\underset{\sim}{u}_1\|)^2$ must be negative since (neg)(pos) = neg. $\lambda_1 (\|\underset{\sim}{u}_1\|)^2 < 0 \Rightarrow A$ is not PSD.

Therefore, A is PSD only if $\lambda_i \geq 0$, and we've shown that A is PSD "and only if" $\lambda_i \geq 0 \; \forall i$

Since we've proven that A is PSD if $\lambda_i \geq 0 \; \forall i$ and A is PSD only if $\lambda_i \geq 0 \; \forall i$ we've proven the desired result.

3c. Show that A is PD iff $\lambda_i > 0$ for each i

Using $A = U\Lambda U^T$ rewrite $\underset{\sim}{x}^T A \underset{\sim}{x}$ as $\underset{\sim}{x}^T U\Lambda U^T \underset{\sim}{x}$. Now using hint $U\Lambda U^T = \sum_{i=1}^{d} \lambda_i \underset{\sim}{u}_i \underset{\sim}{u}_i^T$ we can rewrite again as $\sum_{i=1}^{d} \lambda_i \underset{\sim}{x}^T \underset{\sim}{u}_i \underset{\sim}{u}_i^T \underset{\sim}{x}$. We know that $(\underset{\sim}{x}^T \underset{\sim}{u}_i)$ and $(\underset{\sim}{u}_i^T \underset{\sim}{x})$ are scalars.

So we can rewrite again as $\sum_{i=1}^{d} \lambda_i (\underset{\sim}{x}^T \underset{\sim}{u}_i)^2$. Since the term $(\underset{\sim}{x}^T \underset{\sim}{u}_i)$ is squared, it must be positive.

Therefore, $\sum_{i=1}^{d} \lambda_i (\underset{\sim}{x}^T \underset{\sim}{u}_i)^2$ is only > 0 if $\lambda_i > 0 \ \forall i$.

Thus we've shown A is PD "<u>if</u>" $\lambda_i > 0 \ \forall i$.

Now Let's show that A is PD "<u>and only if</u>" $\lambda_i > 0 \ \forall i$.

Let $\underset{\sim}{x} = \underset{\sim}{u}_1$, giving $U^T \underset{\sim}{u}_1 = \|\underset{\sim}{u}_1\|^2$ (since $\underset{\sim}{u}_i^T \underset{\sim}{u}_1 = 0 \ \forall i \neq 1$).

Let $\lambda_1 \leq 0$. By the spectral decomposition, $\underset{\sim}{x}^T A \underset{\sim}{x} = \sum_{i=1}^{d} \lambda_i (\underset{\sim}{x}^T \underset{\sim}{u}_i)^2$.

Since we let $\underset{\sim}{x} = \underset{\sim}{u}_1 \Rightarrow \underset{\sim}{x}^T = \underset{\sim}{u}_1^T$

giving $\sum_{i=1}^{d} \lambda_i (\underset{\sim}{u}_1^T \underset{\sim}{u}_i)^2 = \lambda_1 (\underset{\sim}{u}_1^T \underset{\sim}{u}_1)^2 + \lambda_2 (\underset{\sim}{u}_1^T \underset{\sim}{u}_2)^2 + \ldots + \lambda_d (\underset{\sim}{u}_1^T \underset{\sim}{u}_d) = \lambda_1 (\|\underset{\sim}{u}_1\|)^2 + \lambda_2 (0) + \ldots + \lambda_d (0)$

$= \lambda_1 (\|\underset{\sim}{u}_1\|)^2$. Assuming $\underset{\sim}{u}_1 \neq \underset{\sim}{0}$ this implies that $(\|\underset{\sim}{u}_1\|)^2 > 0$.

Since we let $\lambda_1 \leq 0$ and $(\|\underset{\sim}{u}_1\|)^2$ is positive, the quantity $\lambda_1 (\|\underset{\sim}{u}_1\|)^2 \leq 0$.

This implies that A is not PD.

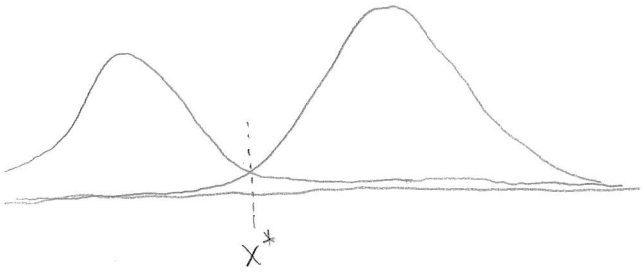
Therefore, A is PD only if $\lambda_i > 0$, and we've shown that a is PD "<u>and only if</u>" $\lambda_i \geq 0 \ \forall i$.

Since we're proven that A is PD if $\lambda_i > 0 \ \forall i$ and A is PD only if $\lambda_i > 0 \ \forall i$ we've proven the desired result.

4. The Bayes Classifier (37 pts).

Let X be a random variable representing a 1-dimensional feature space and let Y be a discrete random variable taking values in in {0,1} (i.e., Y is the corresponding class label). If Y = 0, then the posterior distribution of X for class 0 is Gaussian with mean μ_0 and variance σ^2_0. If Y = 1, then the posterior distribution of X for class 1 is Gaussian with mean μ_1 and variance σ^2_1. Let w0 = Pr(Y = 0) and w1 = Pr(Y = 1) = 1−w0.

(a) (5 pts) Derive the Bayes classifier for this problem as a function of wi, μi, and σi where i ∈{0,1}.



In this case, the Bayes Classifier will be the x value, x^*, where distribution$_0$ and distribution$_1$ intersect. Therefore, to solve for x^* we simply set $\pi_0 p_0(x) = \pi_1 p_1(x)$ where $\pi_0 = w_0$, $\pi_1 = w_1$, p_0 is the pdf of distn$_0$, and p_1 is the pdf of distn$_1$, and then solve for x.

$$w_0 p_0(x) = w_1 p_1(x) \longrightarrow w_0\left(\frac{1}{\sqrt{2\pi\sigma_0^2}}\right)\left(e^{\frac{-(x-\mu_0)^2}{2\sigma_0^2}}\right) = w_1\left(\frac{1}{\sqrt{2\pi\sigma_1^2}}\right)\left(e^{\frac{-(x-\mu_1)^2}{2\sigma_1^2}}\right)$$

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

$$\longrightarrow \ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right] = \frac{\sigma_0^2}{\sigma_0^2}\cdot\frac{(x-\mu_1)^2}{2\sigma_1^2} - \frac{(x-\mu_0)^2}{2\sigma_0^2}\cdot\frac{\sigma_1^2}{\sigma_1^2}$$

$$\longrightarrow 2\ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right] = \frac{\sigma_0^2(x-\mu_1)^2 - \sigma_1^2(x-\mu_0)^2}{\sigma_1^2\sigma_0^2} \longrightarrow \sigma_1^2\sigma_0^2\, 2\ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right] = \sigma_0^2(x-\mu_1)^2 - \sigma_1^2(x-\mu_0)^2$$

$$\longrightarrow \sigma_1^2\sigma_0^2\, 2\ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right] = \sigma_0^2x^2 - 2\mu_1\sigma_0^2x + \mu_1^2\sigma_0^2 - (\sigma_1^2x^2 - 2\mu_0\sigma_1^2x + \mu_0^2\sigma_1^2)$$

$$\longrightarrow \sigma_1^2\sigma_0^2\, 2\ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right] = x^2(\sigma_0^2-\sigma_1^2) + x(2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2) + (\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2)$$

$$\longrightarrow 0 = x^2(\sigma_0^2-\sigma_1^2) + x(2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2) + \left[(\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2) - \sigma_1^2\sigma_0^2\, 2\ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right]\right]$$

$$0 = \underbrace{x^2(\sigma_0^2 - \sigma_1^2)}_{a} + \underbrace{x(2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2)}_{b} + \underbrace{\left[(\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2) - (\sigma_1^2\sigma_0^2 2\ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right]\right]}_{c}$$

This is of the form $0 = ax^2 + bx + c$, so we can apply the quadratic formula to solve for x.

However, we need to solve for two cases: ① when $\sigma_0 = \sigma_1$ and ② when $\sigma_0 \neq \sigma_1$.

① When $\sigma_0 = \sigma_1$, the a constant $= 0$, so the ax^2 term $= 0$.

Thus, we're left with $x(2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2) = \sigma_1^2\sigma_0^2 2\ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right] - (\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2)$

Thus, $x = \dfrac{\sigma_1^2\sigma_0^2\, 2\ln\left[\frac{w_1\sigma_0}{w_0\sigma_1}\right] - (\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2)}{(2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2)}$

② When $\sigma_0 \neq \sigma_1$, we use the quadratic formula and get that

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

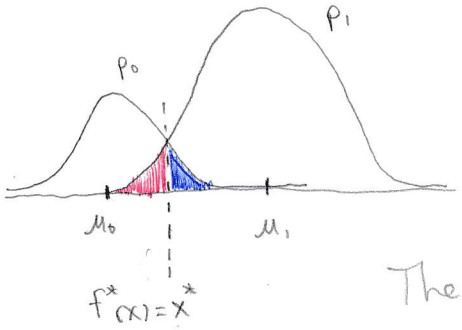
Thus, the Bayes Classifier is:

$$f^*(x) = \begin{cases} \left(\sigma_1^2\sigma_0^2 2\ln\left(\frac{w_1\sigma_0}{w_0\sigma_1}\right) - (\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2)\right) \Big/ (2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2), & \text{for } \sigma_0 = \sigma_1 \\ \dfrac{-(2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2) \pm \sqrt{(2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2)^2 - 4(\sigma_0^2 - \sigma_1^2)\left((\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2) - (\sigma_1^2\sigma_0^2 2\ln\left(\frac{w_1\sigma_0}{w_0\sigma_1}\right)\right)}}{2(\sigma_0^2 - \sigma_1^2)}, & \text{for } \sigma_0 \neq \sigma_1 \end{cases}$$

↑ Also, since this case returns two values of x select the one s.t. $\mu_0 < x < \mu_1$

4. The Bayes Classifier (continued)

(b) (5 pts) Derive the Bayes error rate for this classification problem as a function of wi, μi, and σi where i ∈{0,1}. You may write your solution in terms of the Q function where if Z is a standard normal random variable, then Q(z) = Pr(Z > z).



We can think of the Bayes error rate as the shaded red area (misclassified p_1 values) plus the shaded blue area (misclassified p_0 values).

The formula for this area is $w_0 \cdot p_0(X > x^*) + w_1 \cdot p_1(X < x^*)$

$$= w_0 \cdot p_0\left(Z > \frac{x^* - \mu_0}{\sigma_0}\right) + w_1 \cdot \left(1 - p_1\left(Z > \frac{x^* - \mu_1}{\sigma_1}\right)\right)$$

In terms of $Q(z) = P(Z > z)$ the Bayes error rate is

$$= w_0 Q\left(\frac{x^* - \mu_0}{\sigma_0}\right) + w_1\left(1 - Q\left(\frac{x^* - \mu_1}{\sigma_1}\right)\right)$$

4. The Bayes Classifier (continued)

(c) (5 pts) Describe how to perform cross-validation for a classification problem.

I will explain how k-fold cross-validation is performed. First, the data is divided up into k equal pieces. Then, we fit a model on all but one of those pieces. Call this left-out piece k. Using our newly fitted model (trained on data pieces 1,2,....,k-1), we then predict onto k. Finally, the true classification values (labels) of k are compared with the predicted labels. Then an error rate for k is calculated as: $error_K$ = (number of incorrectly labeled observations in k)/(total number of observations in k).

Next, we "shift" our data, and again fit a model on all but one of those pieces. Say the second time we fit on pieces 1,2,....,k-2, and k, leaving out piece k-1. Then we calculate the error rate for k-1.

We iterate through this process until we've calculated an error rate for all k pieces of the total data. Finally, we obtain the overall cross-validated error rate of our classifier by averaging the error rate for all k pieces.

(d) Set $\mu0 = 0$, $\mu1 = 1.5$, $\sigma0 = \sigma1 = \sigma = 1$, $w0 = 0.3$ and $w1 = 0.7$. For the sample sizes $N \in \{100,200,500,1000\}$, simulate the above classification problem. Apply the Bayes classifier, logistic regression, and the k-nearest neighbor (nn) classifier to the simulated data. Run this simulation for 100 trials and report the following for each sample size:

i. (3 pts) The Bayes error rate.

From part 4(a) we derived the formula for the decision boundary (Bayes classifier), including for when $\sigma0 = \sigma1$:

$$f^*(x) = \left(\sigma_1^2\sigma_0^2\, 2\ln\left(\frac{w_1\sigma_0}{w_0\sigma_1}\right) - (\mu_1^2\sigma_0^2 - \mu_0^2\sigma_1^2)\right) \Big/ (2\mu_0\sigma_1^2 - 2\mu_1\sigma_0^2)$$

To solve for x* we simply substitute values for $\mu0$, $\mu1$, $\sigma0$, $\sigma1$, $w0$, and $w1$, giving:

$$f^*(x) = x^* = \left((1)(1)\, 2\ln\left(\frac{(0.7)(1)}{(0.3)(1)}\right) - ((1.5)^2(1^2) - (0^2)(1)^2)\right) \Big/ (2(0)(1^2) - 2(1.5)(1^2)) \longrightarrow x^* = 0.185135$$

From part 4(b) we derived the formula for the Bayes error rate:

$$w_0 Q\left(\frac{x^* - \mu_0}{\sigma_0}\right) + w_1\left(1 - Q\left(\frac{x^* - \mu_1}{\sigma_1}\right)\right) = \text{Bayes error rate}$$

Now that we have all necessary components, we can simply substitute in x*, $\mu0$, $\mu1$, $\sigma0$, $\sigma1$, $w0$, and $w1$, giving:

$$w_0 Q\left(\frac{x^* - \mu_0}{\sigma_0}\right) + w_1\left(1 - Q\left(\frac{x^* - \mu_1}{\sigma_1}\right)\right) = (0.3)\left(1 - \Phi\left(\frac{0.185135 - 1.5}{1}\right)\right) + (0.7)\left(\Phi\left(\frac{0.185135 - 0}{1}\right)\right)$$

$$= (0.3)(1 - 0.5734388) + (0.7)(0.0942776)$$

Bayes error rate = 0.1939628

ii. (3 pts) The average value of *k* as selected using cross-validation.

- For N = 100: $avg_k = 5.37$
- For N = 200: $avg_k = 7.81$
- For N = 500: $avg_k = 7.07$
- For N = 1000: $avg_k = 9.22$

In reporting these figures, it should be noted that I don't have particularly high confidence in them. I was able to write a function to extract the cross-validated error rate for the *k*-nn classifier just fine, but had a lot of difficulty extracting the *k* values. I think that I came up with a workaround, but it was very hackish and I have no idea if it works as I think it does.

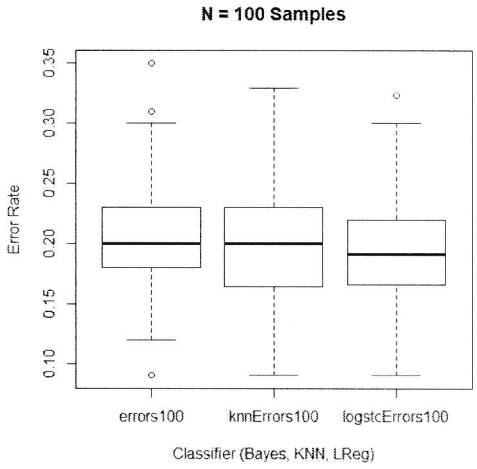
iii. The classification error of each classifier. Calculate the error of the logisctic regression and *k*-nn classifiers for each trial using cross-validation and describe how you performed cross-vladiation (e.g. 5-fold, 10-fold cross-validation, etc.; 1pt). You may use built-in logistic regression and *k*-nn classifier functions (i.e. you do not need to code these up from scratch).
Report the mean and standard deviation of the error in

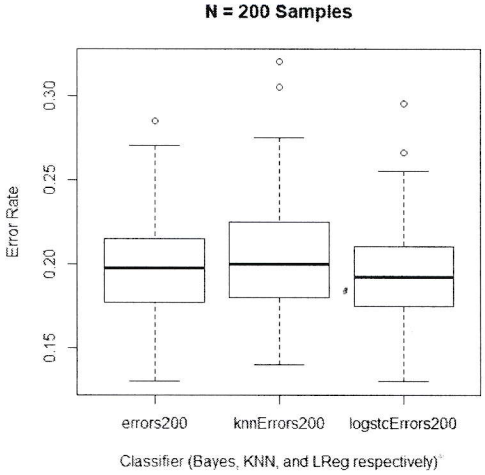
A. (6pts) Table form

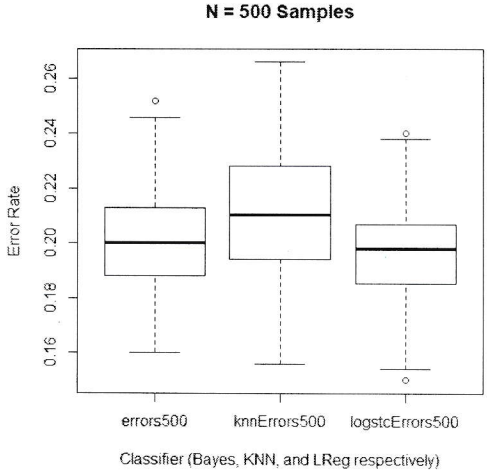
Classifier	Bayes Classifier				*K*-NN Classifier				Logistic Regression			
N	100	200	500	1000	100	200	500	1000	100	200	500	1000
Mean	0.203700	0.19785	0.20028	0.19819	0.199978	0.202415	0.210615	0.210067	0.195831	0.193717	0.195693	0.194938
Std Dev	0.042891	0.026258	0.018757	0.012757	0.048482	0.034904	0.023038	0.015589	0.043726	0.027471	0.019079	0.012228

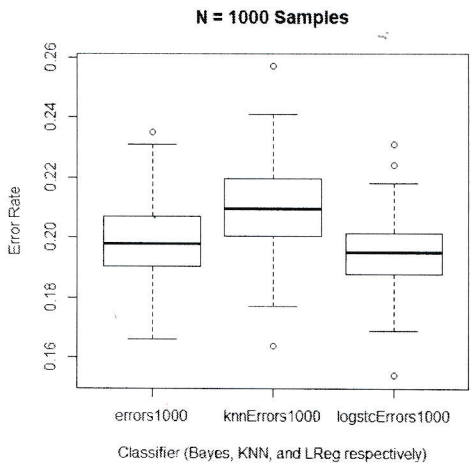
B. (9 pts) Graphical form. Make a plot with sample size on the x-axis and the error on the y-axis. Plot the mean and standard deviation using error bars. Plot the results for all 3 classifiers on the same plot. You may need to use log scales for better visualization. You do not need to turn in your code.

NOTE: I wasn't sure if the instruction to put results for all 3 classifiers on the same plot meant for all values of N, or to take the average for each classifier across the N. I opted to plot each of the 3 classifiers on the same plot for each N value. As such, each plot has a slightly different y-axis since there are differing amounts of variation for differing values of N.









5. How long did this assignment take you? (5 pts)

DAY	Total Time
Tue, 10/23: 9:00am-12:00pm = **3** hrs, 6:30pm-9:00pm = **2.5** hrs	5.50 hrs
Wed, 10/24: 8:15am-9:45am=**1.5** hrs, 1:00pm-2:00pm=**1** hr, 5:00pm-10:00pm = **5** hrs	7.50 hrs
Thu, 10/25: 9:00am-11:45am=**2.75** hrs, 4:00-4:15=**0.25** hrs, 6:00pm-12:15am=**6.25** hrs	9.25 hrs
Fri, 10/26: 9:00am-11:15am = **2.25** hrs, 12:30pm-1:15pm = **0.75** hrs, 2:30pm-3:30pm=**1** hr	4.00 hrs

TOTAL TIME SPENT ON HOMEWORK 2: 26.25 HOURS

- From Tuesday morning until Friday at 3:30 pm I spent very almost every waking hour (outside of lectures, meetings, and meals) working on this assignment.
- In all fairness, I ought to qualify this statement by acknowledging that I'm not an exceptional student in the context of this class.
- Also, thanks for your help on the homework. It helped **a ton**, and it would have taken much longer if I wouldn't have been able to come talk with you about it.

6. Type up homework solutions. (5 pts)

I typed what things I could but didn't want to take the homework assignment from 26 hours to 30+, especially since I'm a fairly novice LaTex-er.