Name: Michael Chen **Student ID:** 23567341

CS 189: Introduction to Machine Learning

Homework 2

Due: February 18, 2016 at 11:59pm

Instructions

- Homework 2 is completely a written assignment; no coding involved.
- We prefer that you typeset your answers using the LaTeX template on bCourses. If there is not enough space for your answer, you may continue your answer on the next page. Make sure to start each question on a new page.
- Neatly handwritten and scanned solutions will also be accepted. Make sure your answers are readable!
- Submit a PDF with your answers to the Homework 2 assignment on Gradescope. You should be able to see CS 189/289A on Gradescope when you log in with your bCourses email address. Please make a Piazza post if you have any problems accessing Gradescope.
- While submitting to Gradescope, you will have to select the pages containing your answer for each question.
- The assignment covers concepts in probability, linear algebra, matrix calculus, and decision theory.
- **Start early. This is a long assignment. Some of the material may not have been covered in lecture; you are responsible for finding resources to understand it.**

Problem 1: Expected Value.

A target is made of 3 concentric circles of radii $1/\sqrt{3}$, 1 and $\sqrt{3}$ feet. Shots within the inner circle are given 4 points, shots within the next ring are given 3 points, and shots within the third ring are given 2 points. Shots outside the target are given 0 points.

Let X be the distance of the hit from the center (in feet), and let the probability density function of X be

$$f(x) = \begin{cases} \frac{2}{\pi(1+x^2)} & x > 0 \\ 0 & \text{otherwise} \end{cases}$$

What is the expected value of the score of a single shot?

Solution:

$$E[F(x)] = 4\Pr(0 < x < \tfrac{1}{\sqrt{3}}) + 3\Pr(\tfrac{1}{\sqrt{3}} < x < 1) + 2\Pr(1 < x < \sqrt{3})$$

$$= 4\int_0^{1/\sqrt{3}} \frac{2}{\pi(1+x^2)}dx + 3\int_{1/\sqrt{3}}^{1} \frac{2}{\pi(1+x^2)}dx + 2\int_1^{\sqrt{3}} \frac{2}{\pi(1+x^2)}dx$$

$$= 4\left(\frac{2}{\pi}\cdot\frac{\pi}{6}\right) + 3\left(\frac{2}{\pi}\cdot\frac{\pi}{12}\right) + 2\left(\frac{2}{\pi}\cdot\frac{\pi}{12}\right)$$

$$= \frac{4}{3} + \frac{1}{2} + \frac{1}{3}$$

$$= \frac{13}{6}$$

Problem 2: MLE.

Assume that the random variable X has the exponential distribution

$$f(x;\theta) = \theta e^{-\theta x} \qquad x \geq 0, \theta > 0$$

where θ is the parameter of the distribution. Use the method of maximum likelihood to estimate θ if 5 observations of X are $x_1 = 0.9$, $x_2 = 1.7$, $x_3 = 0.4$, $x_4 = 0.3$, and $x_5 = 2.6$, generated i.i.d. (i.e., independent and identically distributed).

Solution:

$$\mathcal{L}(\theta; x_1, \ldots, x_5) = f(0.9;\theta) f(1.7;\theta) f(0.4;\theta) f(0.3;\theta) f(2.6;\theta)$$

$$= \theta^5 e^{-\theta(0.9+1.7+0.4+0.3+2.6)}$$

$$= \theta^5 e^{-5.9\theta}$$

$$\ln(\mathcal{L}) = 5\ln\theta - 5.9\theta$$

$$\frac{\partial \ln(\mathcal{L})}{\partial\theta} = 0 = \frac{\partial}{\partial\theta}(5\ln\theta - 5.9\theta)$$

$$0 = \frac{5}{\theta} - 5.9 \implies \theta = \frac{5}{5.9} \approx 0.847$$

Definition. Let $A \in \mathbb{R}^{n \times n}$ be a symmetric matrix. We say that A is **positive definite** if $\forall x \in \mathbb{R}^n \mid x \neq \vec{0},\ x^\top A x > 0$. Similarly, we say that A is **positive semidefinite** if $\forall x \in \mathbb{R}^n,\ x^\top A x \geq 0$.

Problem 3: Positive Definiteness.

Let $x = \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix}^\top \in \mathbb{R}^n$, and let $A \in \mathbb{R}^{n \times n}$ be the square matrix

$$A = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}$$

(a) Give an explicit formula for $x^\top A x$. Write your answer as a sum involving the elements of A and x.

(b) Show that if A is positive definite, then the entries on the diagonal of A are positive (that is, $a_{ii} > 0$ for all $1 \leq i \leq n$).

Solution:

a) $$[x_1 \dots x_n] \begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \\ a_{1n} & & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} = [x_1 \dots x_n] \begin{bmatrix} a_{11}x_1 + \cdots + a_{1n}x_n \\ \vdots \\ a_{n1}x_1 + \cdots + a_{nn}x_n \end{bmatrix}$$

$$= (a_{11}x_1 + \cdots + a_{1n}x_n)x_1 + \cdots + (a_{n1}x_1 + \cdots + a_{nn}x_n)x_n$$

$$= \sum_{i=1}^{n} x_i \sum_{j=1}^{n} a_{ij} x_j$$

b) consider the set of vectors in $\mathbb{R}^n$ denoted by e_i, where $i \in [1, \dots, n]$, that form an orthonormal basis for $\mathbb{R}^n$ where

$$e_1 = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix},\ e_2 = \begin{bmatrix} 0 \\ 1 \\ \vdots \\ 0 \\ 0 \end{bmatrix}, \dots, e_i = \begin{bmatrix} 0 \\ \vdots \\ 1 \\ \vdots \\ 0 \end{bmatrix}, \dots, e_n = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad \forall i \in [1, \dots, n]$$

Then $e_i^\top A e_i = 0 + \cdots 1 \cdot a_{ii} \cdot 1 + \cdots + 0 = a_{ii}$

since $e_i^\top A e_i > 0$ for a positive definite matrix, that means $\forall i \in [1, \dots, n]$, $a_{ii} > 0$ where $\begin{bmatrix} a_{11} & & 0 \\ & a_{ii} & \\ 0 & & a_{nn} \end{bmatrix}$

Problem 4: Short Proofs.

A is symmetric in all parts.

(a) Let A be a positive semidefinite matrix. Show that $A + \gamma I$ is positive definite for any $\gamma > 0$.

(b) Let A be a positive definite matrix. Prove that all eigenvalues of A are greater than zero.

(c) Let A be a positive definite matrix. Prove that A is invertible. (Hint: Use the previous part.)

(d) Let A be a positive definite matrix. Prove that there exist n linearly independent vectors $x_1, x_2, ..., x_n$ such that $A_{ij} = x_i^\top x_j$. (Hint: Use the spectral theorem and what you proved in (b) to find a matrix B such that $A = B^\top B$.)

Solution:

a) $x^T(A + \gamma I)x$

$= x^T A x + x^T \gamma I x$

We note that $x^T A x \geq 0$ b/c A is semi-definite

Also $x^T \gamma I x = \gamma x^T I x$

$= \gamma x^T x = \gamma \sum_{i=1}^{n} x_i^2 > 0$ if $\gamma > 0$; $\vec{x} \neq \vec{0}$

b) An arbitrary eigenvalue λ of A must, by definition, satisfy the following

$$A\vec{x} = \lambda\vec{x}$$

If we left multiply by $\vec{x}^T$

$$x^T A x = x^T \lambda x$$

Given the definition of positive definite

$$0 < x^T A x = x^T \lambda x = \underbrace{\lambda}_{\text{a scalar}} x^T x$$

$0 < \lambda$, so the eigenvalues are greater than 0

c) If A were not invertible, it would have a non-trivial null-space, and thus some $\vec{x}$ such that

$$A\vec{x} = 0\,\vec{x}$$

but that would mean 0 is an eigenvalue of A, but since A is positive definite, its eigenvalues must be greater than 0. Thus 0 is not an eigenvalue of A and A is invertible

d) $A = UDU^T$

$= U(\sqrt{D}\sqrt{D})U^T$

$= U\sqrt{D}U^T U\sqrt{D}U^T$

$= (U\sqrt{D}U^T)^T(U\sqrt{D}U^T)$

so we have found $B = U\sqrt{D}U^T$ such that $A = B^T B$

From the Spectral Theorem, we know $U := [x_1, ..., x_n]$ is orthogonal,

Thus we know B is diagonalizable, and so is linearly independent and has n linearly independent vectors.

Problem 5: Derivatives and Norm Inequalities.

Derive the expression for following questions. Do not write the answers directly.

(a) Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$. Derive $\frac{\partial(\mathbf{x}^T\mathbf{a})}{\partial\mathbf{x}}$.

(b) Let $\mathbf{A} \in \mathbb{R}^{n\times n}, \mathbf{x} \in \mathbb{R}^n$. Derive $\frac{\partial(\mathbf{x}^T\mathbf{A}\mathbf{x})}{\partial\mathbf{x}}$.

(c) Let $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{n\times n}$. Derive $\frac{\partial \text{Trace}(\mathbf{XA})}{\partial\mathbf{X}}$.

(d) Let $\mathbf{x} \in \mathbb{R}^n$. Prove that $\|\mathbf{x}\|_2 \le \|\mathbf{x}\|_1 \le \sqrt{n}\|\mathbf{x}\|_2$. (Note that $\|\mathbf{x}\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$ and $\|\mathbf{x}\|_1 = \sum_{i=1}^n |x_i|$.) (Hint: The Cauchy-Schwarz inequality may come in handy.)

Solution:

a) $$\frac{\partial(x^Ta)}{\partial x} \underset{\uparrow}{=} \frac{\partial(a^Tx)}{\partial x} = a^T\frac{\partial(x)}{\partial x} = a^T$$

This works b/c the dot product is the same either way

b) $$x^TAx = [x_1\ x_2 \cdots x_n]\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & & \\ \vdots & & \ddots & \\ a_{n1} & & & a_{nn}\end{bmatrix}\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n\end{bmatrix} = [x_1\ x_2 \cdots x_n]\begin{bmatrix} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n \\ \vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n\end{bmatrix}$$

$$= \sum_{j=1}^n \sum_{i=1} a_{ij}x_ix_j$$

we note that

$$\frac{\partial(x^TAx)}{\partial x_K} = \underbrace{\sum_{j=1}^n a_{Kj}x_j}_{\text{sum rows } K} + \underbrace{\sum_{i=1}^n a_{iK}x_i}_{\text{sum cols } K}$$

so $$\frac{\partial(x^TAx)}{\partial x} = \begin{pmatrix} \sum_j a_{1j}x_j \\ \vdots \\ \sum_j a_{nj}x_j \end{pmatrix} + \begin{pmatrix} \sum_i a_{i1}x_i \\ \vdots \\ \sum_i a_{in}x_i\end{pmatrix}$$

$$= Ax + A^Tx$$
$$= (A + A^T)x$$

c) $$XA = \text{Tr}\left(\begin{bmatrix} x_{11} & \cdots & x_{1n} \\ \vdots & \ddots & \\ x_{n1} & & x_{nn}\end{bmatrix}\begin{bmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \\ a_{n1} & & a_{nn}\end{bmatrix}\right) = (a_{11}x_{11} + \cdots + a_{n1}x_{1n}) + \cdots + (a_{1n}x_{n1} + \cdots + a_{nn}x_{nn})$$

$$= \sum_{j=1}^n\sum_{i=1}^n a_{ji}x_{ij} \Rightarrow \frac{\partial \text{Tr}(XA)}{\partial x_{ij}} = a_{ji} \Rightarrow \frac{\partial \text{Tr}(XA)}{\partial X} = A^T$$

d) we show the left inequality first

$$\|x\|_2 \overset{?}{\le} \|x\|_1$$

$$\sqrt{\sum_{i=1}^n x_i^2} \overset{?}{\le} \sum_{i=1}^n |x_i|$$

$$\sum_{i=1}^n x_i^2 \overset{\checkmark}{\le} \sum_{i=1}^n\sum_{j=1}^n |x_ix_j|$$

which is true b/c the terms in the left series are a <u>subset</u> of the terms in the right series, and we are only summing positive terms.

We now show the right inequality

$$\|\vec{x}\|_1 \overset{?}{\le} \sqrt{n}\|\vec{x}\|_2$$

$$\langle \vec{1} \cdot \vec{x} \rangle \overset{\checkmark}{\le} \|\vec{1}\|_2 \|\vec{x}\|_2$$

which is true by the Cauchy-Schwarz inequality

Problem 6: Weighted Linear Regression.

Let $\mathbf{X}$ be a $n \times d$ data matrix, $\mathbf{Y}$ be the corresponding $n \times 1$ target/label matrix and $\mathbf{\Lambda}$ be the diagonal $n \times n$ matrix containing a weight for each example. More explicitly, we have

$$\mathbf{X} = \begin{bmatrix} (\mathbf{x}^{(1)})^T \\ (\mathbf{x}^{(2)})^T \\ \dots \\ (\mathbf{x}^{(n)})^T \end{bmatrix} \qquad \mathbf{Y} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \mathbf{y}^{(2)} \\ \dots \\ \mathbf{y}^{(n)} \end{bmatrix} \qquad \mathbf{\Lambda} = \text{diag}(\lambda^{(1)}, \lambda^{(2)}, \dots, \lambda^{(n)})$$

where $\mathbf{x}^{(i)} \in \mathbb{R}^d$, $\mathbf{y}^{(i)} \in \mathbb{R}$, and $\lambda^{(i)} > 0 \quad \forall \ i \in \{1 \dots n\}$. $\mathbf{X}$, $\mathbf{Y}$ and $\mathbf{\Lambda}$ are fixed and known.

In this question, we will try to fit a weighted linear regression model for this data. We want to find the value of weight vector $\mathbf{w}$ which best satisfies the following equation $\mathbf{y}^{(i)} = \mathbf{w}^T\mathbf{x}^{(i)} + \epsilon^{(i)}$, where ϵ is noise. This is achieved by minimizing the weighted noise for all the examples. Thus, our risk (cost) function is defined as follows:

$$\begin{aligned} R[\mathbf{w}] &= \sum_{i=1}^{n} \lambda^{(i)}(\epsilon^{(i)})^2 \\ &= \sum_{i=1}^{n} \lambda^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} - \mathbf{y}^{(i)})^2 \end{aligned}$$

(a) Write this risk function $R[\mathbf{w}]$ in matrix notation (i.e., in terms of $\mathbf{X}$, $\mathbf{Y}$, $\mathbf{\Lambda}$ and $\mathbf{w}$).

(b) Find the weight vector $\mathbf{w}$ that minimizes the risk function obtained in the previous part. You can assume that $\mathbf{X}^T\mathbf{\Lambda}\mathbf{X}$ is full rank. (Hint: You may use the expression you derived in Question 5(b).)

(c) The L_2 regularized risk function, for $\gamma > 0$, is

$$R[\mathbf{w}] = \sum_{i=1}^{n} \lambda^{(i)}(\mathbf{w}^T\mathbf{x}^{(i)} - \mathbf{y}^{(i)})^2 + \gamma\|\mathbf{w}\|_2^2$$

Rewrite this new risk function in matrix notation as in (a) and solve for $\mathbf{w}$ as in (b).

(d) How does γ affect the regression model? How does this fit in with what you already know about L_2 regularization? (Hint: Observe the different expressions for $\mathbf{w}$ obtained in (b) and (c).)

Solution:

Problem 6:

a) $R[\omega] = \sum_{i=1}^{n} \lambda_i (\omega^T x_i - y_i)^2$

$= \lambda_1(\omega^T x_1 - y_1)^2 + \lambda_2(\omega^T x_2 - y_2)^2 + \dots + \lambda_n(\omega^T x_n - y_n)^2$

we note that: $\begin{bmatrix} - x_1 - \\ \vdots \\ - x_n - \end{bmatrix} \begin{bmatrix} | \\ \omega \\ | \end{bmatrix} - \begin{bmatrix} | \\ y \\ | \end{bmatrix} = \begin{pmatrix} \omega^T x_1 - y_1 \\ \vdots \\ \omega^T x_n - y_n \end{pmatrix}$

$= X\omega - Y$

So: $R[\omega] = (X\omega - Y)^T \Lambda (X\omega - Y)$

b) $\frac{\partial}{\partial \omega} (X\omega - Y)^T \Lambda (X\omega - Y) = 0$

$2X^T\Lambda(X\omega - Y) = 0$

$X^T\Lambda X\omega - X^T\Lambda Y = 0$

$X^T\Lambda X\omega = X^T\Lambda Y \implies \omega = (X^T\Lambda X)^{-1} X^T\Lambda Y$

c) $R[\omega] = \sum_{i=1}^{n} \lambda_i (\omega^T x_i - y_i)^2 + \gamma \|\omega\|^2$

$= (X\omega - Y)^T \Lambda (X\omega - Y) + 2\gamma \omega^T\omega$

$\frac{\partial}{\partial \omega} R[\omega] = X^T\Lambda X\omega - X^T\Lambda Y + 2\gamma\omega = 0$

$X^T\Lambda X\omega + 2\gamma\omega = X^T\Lambda Y$

$(X^T\Lambda X + 2\gamma)\omega = X^T\Lambda Y$

$\omega = (X^T\Lambda X + 2\gamma)^{-1} X^T\Lambda Y$

d) Generally speaking, larger weights $\vec{\omega}$ indicate overfitting. With a larger γ value, a larger magnitude $\vec{\omega}$ (e.g. $\|\omega\|^2$) would lead to a higher cost function. So a larger γ keeps the weight smaller, thus preventing overfitting (but maybe underfitting?)

Problem 7: Classification.

Suppose we have a classification problem with classes labeled $1, \ldots, c$ and an additional doubt category labeled as $c+1$. Let the loss function be the following:

$$\ell(f(x) = i, y = j) = \begin{cases} 0 & \text{if } i = j \quad i, j \in \{1, \ldots, c\} \\ \lambda_r & \text{if } i = c+1 \\ \lambda_s & \text{otherwise} \end{cases}$$

where λ_r is the loss incurred for choosing doubt and λ_s is the loss incurred for making a misclassification. Note that $\lambda_r \geq 0$ and $\lambda_s \geq 0$.

Hint: The risk of classifying a new datapoint as class $i \in \{1, 2, \ldots, c+1\}$ is

$$R(\alpha_i|x) = \sum_{j=1}^{c} \ell(f(x) = i, y = j) P(\omega_j|x)$$

(a) Show that the minimum risk is obtained if we follow this policy: (1) choose class i if $P(\omega_i|x) \geq P(\omega_j|x)$ for all j and $P(\omega_i|x) \geq 1 - \lambda_r/\lambda_s$, and (2) choose doubt otherwise.

(b) What happens if $\lambda_r = 0$? What happens if $\lambda_r > \lambda_s$? Is this consistent with your intuition?

Solution:

a) $R(\alpha_i \mid x) = \sum_{j \neq i}^{c} \lambda_s P(\omega_j \mid x)$

$= \lambda_s \sum_{j \neq i}^{c} P(\omega_j \mid x) = \lambda_s (1 - P(\omega_i \mid x))$

$R(c+1 \mid x) = \lambda_r$

$R(\alpha_i \mid x) = R(c+1 \mid x)$

① $\lambda_s (1 - P(\omega_i \mid x)) = \lambda_r \implies P(\omega_i \mid x) = 1 - \frac{\lambda_r}{\lambda_s}$

Note that if we plug $P(\omega_i \mid x)$ back into ①, we notice that larger $P(\omega_i \mid x)$ leads to a lower risk/cost, so we should choose a class $i \in [1, \ldots, c]$ if

$P(\omega_i \mid x) \geq 1 - \frac{\lambda_r}{\lambda_s}$

But then which of the class i's do we choose, obviously the one with the highest probability

b) - if $\lambda_r = 0$ then we always select a class i b/c $P(\omega_i \mid x) \leq 1$

- if $\lambda_r > \lambda_s$ then we always select $c+1$ b/c $P(\omega_i \mid x) > 0 > 1 - \frac{\lambda_r}{\lambda_s}$

These results make sense intuitively b/c if the $\lambda_r = 0$, then there is no penalty for picking "doubtfull" each time; it gives the same cost as picking correctly. If $\lambda_r > \lambda_s$ then the penalty for guessing wrong is less than picking "doubtful", so we should always guess then

Problem 8: Gaussians.

Let $P(x \mid \omega_i) \sim \mathcal{N}(\mu_i, \sigma^2)$ for a two-category, one-dimensional classification problem with $P(\omega_1) = P(\omega_2) = 1/2$. Here, the classes are ω_1 and ω_2. For this problem, we have $\mu_2 \geq \mu_1$.

(a) Find the optimal Bayes decision boundary (i.e., find x such that $P(\omega_1 \mid x) = P(\omega_2 \mid x)$). What is the corresponding decision rule?

(b) Show that the Bayes error associated with this decision rule is

$$P_e = \frac{1}{\sqrt{2\pi}} \int_a^{\infty} e^{-z^2/2} dz$$

where $a = \dfrac{\mu_2 - \mu_1}{2\sigma}$. The Bayes error is the probability of misclassification:

$$P_e = P((\text{misclassified as } \omega_1) \mid \omega_2)P(\omega_2) + P((\text{misclassified as } \omega_2) \mid \omega_1)P(\omega_1).$$

Solution:

a) $P(\omega_1 \mid x) = P(\omega_2 \mid x)$

$P(x \mid \omega_1) P(\omega_1) = P(x \mid \omega_2) P(\omega_2)$

$$\cancel{\frac{1}{\sigma\sqrt{2\pi}}} \exp\left[-\frac{(x-\mu_1)^2}{2\sigma^2}\right] = \cancel{\frac{1}{\sigma\sqrt{2\pi}}} \exp\left[-\frac{(x-\mu_2)^2}{2\sigma^2}\right]$$

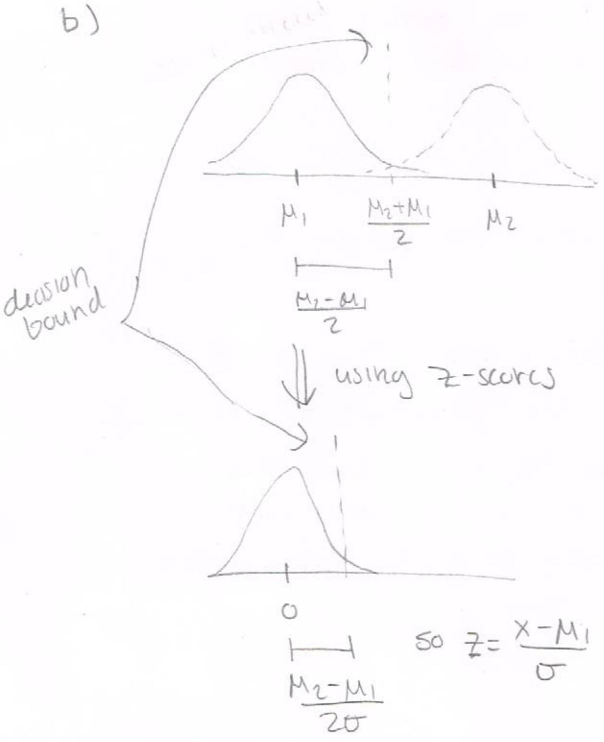
$$(x-\mu_1)^2 = (x-\mu_2)^2$$

$$x^2 - 2\mu_1 x + \mu_1^2 = x^2 - 2\mu_2 x + \mu_2^2$$

$$2\mu_2 x - 2\mu_1 x = \mu_2^2 - \mu_1^2$$

$$x = \frac{\mu_2^2 - \mu_1^2}{2(\mu_2 - \mu_1)} = \frac{\mu_2 + \mu_1}{2}$$

b)



$$P_e = P(\omega_2) \int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} P(x \mid \omega_2)\, dx + P(\omega_1) \int_{\frac{\mu_1+\mu_2}{2}}^{\infty} P(x \mid \omega_1)\, dx$$

$$= \frac{1}{2} \int_{-\infty}^{\frac{\mu_1+\mu_2}{2}} N(\mu_2, \sigma^2)\, dx + \frac{1}{2} \int_{\frac{\mu_1+\mu_2}{2}}^{\infty} N(\mu_1, \sigma^2)\, dx$$

Here we note that since the two probabilities have the same σ & the decision bound is the same distance from both centers, we can consolidate the two factors

$$= \int_{\frac{\mu_1+\mu_2}{2}}^{\infty} \frac{1}{\sqrt{2\pi}\,\sigma} e^{-(x-\mu_1)^2/2} dx$$

which is the same as the following when we use z-scores.

$$= \int_{\frac{\mu_2-\mu_1}{2\sigma}}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz$$