Outline

1. MAP Classifier for Normal Distributions
2. Performance of the Bayes Classifier
3. Error bounds
By far the most popular conditional class distribution model is the Gaussian distribution:

\[ p(x|A) = \mathcal{N}({\mu}_A, {\sigma}_A^2) = \frac{1}{\sqrt{2\pi{\sigma}_A}} \exp \left[ -\frac{1}{2} \left( \frac{x - {\mu}_A}{\sigma_A} \right)^2 \right] \] (1)

and \[ p(x|B) = \mathcal{N}({\mu}_B, {\sigma}_B^2). \]
For the two-class case where both distributions are Gaussian, the following MAP classifier can be defined as:

\[
\frac{\mathcal{N}(\mu_A, \sigma_A^2)}{\mathcal{N}(\mu_B, \sigma_B^2)} > \frac{P(B)}{P(A)} \quad (2)
\]

\[
\frac{\exp\left(-\frac{1}{2} \left(\frac{x - \mu_A}{\sigma_A}\right)^2\right)}{\exp\left(-\frac{1}{2} \left(\frac{x - \mu_B}{\sigma_B}\right)^2\right)} > \frac{\sigma_A P(B)}{\sigma_B P(A)} \quad (3)
\]
In log-likelihood form:

\[
\exp \left[ -\frac{1}{2} \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \right] \frac{A}{B} > \frac{\sigma_A P(B)}{\sigma_B P(A)}
\]

\[
\left[ -\frac{1}{2} \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \right] - \left[ -\frac{1}{2} \left( \frac{x - \mu_B}{\sigma_B} \right)^2 \right] \frac{A}{B} > \ln [\sigma_A P(B)] - \ln [\sigma_B P(A)]
\]

\[
\left[ \left( \frac{x - \mu_B}{\sigma_B} \right)^2 \right] - \left[ \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \right] \frac{A}{B} > 2 [\ln [\sigma_A P(B)] - \ln [\sigma_B P(A)]]
\]
MAP Classifier for Normal Distributions

Giving us the final form:

\[
\begin{align*}
&\left( \frac{x - \mu_B}{\sigma_B} \right)^2 - \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \\
&\quad \begin{array}{c}
A \\
B
\end{array}
\quad \begin{array}{c}
> 2 \left[ \ln \left( \sigma_A P(B) \right) - \ln \left( \sigma_B P(A) \right) \right] \\
< \\
\end{array}
\end{align*}
\] (7)

Does this look familiar?
The decision boundary (threshold) for the MAP classifier where \( P(x|A) \) and \( P(x|B) \) are Gaussian distributions can be found by solving the following expression for \( x \):

\[
\left( \frac{x - \mu_B}{\sigma_B} \right)^2 - \left( \frac{x - \mu_A}{\sigma_A} \right)^2 = 2 \left[ \ln \left( \frac{\sigma_A P(B)}{\sigma_B P(A)} \right) \right] \tag{8}
\]

\[
\chi^2 \left[ \frac{1}{\sigma_B^2} - \frac{1}{\sigma_A^2} \right] - 2x \left[ \frac{\mu_B}{\sigma_B^2} - \frac{\mu_A}{\sigma_A^2} \right] + \frac{\mu_B^2}{\sigma_B^2} - \frac{\mu_A^2}{\sigma_A^2} = 2 \ln \left( \frac{\sigma_A P(B)}{\sigma_B P(A)} \right) \tag{9}
\]
For case where $\sigma_A = \sigma_B$, $P(A) = P(B) = \frac{1}{2}$:

$$x^2 \left[ \frac{1}{\sigma_B^2} - \frac{1}{\sigma_A^2} \right] - 2x \left[ \frac{\mu_B}{\sigma_B^2} - \frac{\mu_A}{\sigma_A^2} \right] + \frac{\mu_B^2}{\sigma_B^2} - \frac{\mu_A^2}{\sigma_A^2} = 2 \ln \left[ \frac{\sigma_A P(B)}{\sigma_B P(A)} \right]$$

(10)

$$x^2(\sigma_A^2 - \sigma_B^2) - 2x(\mu_B \sigma_A^2 - \mu_A \sigma_B^2) + (\mu_B^2 \sigma_A^2 - \mu_A^2 \sigma_B^2) = 2 \ln [1]$$

(11)

Since $\ln(1) = 0$ and $\sigma_A = \sigma_B$,

$$x = \frac{(\mu_B^2 \sigma_A^2 - \mu_A^2 \sigma_A^2)}{2(\mu_B \sigma_A^2 - \mu_A \sigma_A^2)}$$

(12)

$$x = \frac{(\mu_B^2 - \mu_A^2)}{2(\mu_B - \mu_A)}$$

(13)
Since \((a^2 - b^2) = (a - b)(a + b)\):

\[
    x = \frac{(\mu_B - \mu_A)(\mu_B + \mu_A)}{2(\mu_B - \mu_A)} \tag{14}
\]

\[
    x = \frac{(\mu_B + \mu_A)}{2} \tag{15}
\]

Therefore, for the case of equally likely, equi-variance classes, the MAP rule reduces to a threshold midway between the means.
For case where $P(A) \neq P(B)$ and $\sigma_A \neq \sigma_B$, the threshold shifts and a second threshold appears as the second solution to the quadratic expression.
Example of a 1-D case:

Suppose that, given a pattern $x$, we wish to classify it as one of two classes: class $A$ and class $B$.

Suppose the two classes have patterns $x$ which are normally distributed as follows:

$$p(x|A) = \mathcal{N}(\mu_A, \sigma_A^2) = \frac{1}{\sqrt{2\pi\sigma_A}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \right]$$

$$p(x|B) = \mathcal{N}(\mu_B, \sigma_B^2) = \frac{1}{\sqrt{2\pi\sigma_B}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_B}{\sigma_B} \right)^2 \right]$$

$\mu_A = 130, \mu_B = 150$. 
Question: If we know that in a previous case that 4 patterns belong to class $A$ and 6 patterns belong to class $B$, and both classes have the same standard deviation of 20, what is the MAP classifier?

For the two-class case where both distributions are Gaussian, the following MAP classifier can be defined as:

$$\frac{\mathcal{N}(\mu_A, \sigma_A^2)}{\mathcal{N}(\mu_B, \sigma_B^2)} \begin{cases} A < B & \frac{P(B)}{P(A)} \\ A > B & \frac{P(A)}{P(B)} \end{cases}$$  \hspace{1cm} (18)

$$\exp \left[ -\frac{1}{2} \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \right] \begin{cases} A < B & \frac{\sigma_A P(B)}{\sigma_B P(A)} \\ A > B & \frac{\sigma_B P(A)}{\sigma_A P(B)} \end{cases}$$  \hspace{1cm} (19)
MAP Classifier for Normal Distributions

- Plugging in $\mu_A$, $\mu_B$, and $\sigma_A = \sigma_B = \sigma$:

$$\frac{\exp \left[ -\frac{1}{2} \left( \frac{x-130}{20} \right)^2 \right]}{\exp \left[ -\frac{1}{2} \left( \frac{x-150}{20} \right)^2 \right]} \begin{cases} A \\
B 
\end{cases} \frac{P(B)}{P(A)}$$  \hspace{1cm} (20)

- Taking the log:

$$\begin{align*}
\left[ -\frac{1}{2} (x - 130)^2 \right] & - \left[ -\frac{1}{2} (x - 150)^2 \right] \\
\begin{cases} A \\
B \end{cases} & > 2(20^2) \ln \left[ P(B) \right] - \ln \left[ P(A) \right] \\
\begin{cases} A \\
B \end{cases} & \leq 800 \left[ \ln \left[ P(B) \right] - \ln \left[ P(A) \right] \right] \hspace{1cm} (22)
\end{align*}$$
The prior probability $P(A)$ and $P(B)$ can be determined as:

$$P(A) = \frac{4}{6 + 4} = 0.4 \quad P(B) = \frac{6}{6 + 4} = 0.6 \quad (23)$$

Plugging in $P(A)$ and $P(B)$:

$$\begin{align*}
A & \quad \left( (x - 150)^2 \right) - \left( (x - 130)^2 \right) > 800 \ln \left[ \frac{0.6}{0.4} \right] \quad (24) \\
B & \quad \left( (x - 150)^2 \right) - \left( (x - 130)^2 \right) < 800 \ln \left[ 1.5 \right] \quad (25)
\end{align*}$$
Expanding and simplifying:

\[
\left( x - 150 \right)^2 - \left( x - 130 \right)^2 \geq 800 \ln \left[ 1.5 \right] \quad (26)
\]

\[
(x^2 - 300x + (150)^2) - (x^2 - 260x + (130)^2) \geq 800 \ln \left[ 1.5 \right] \quad (27)
\]

\[
-40x < 800 \ln \left[ 1.5 \right] - (150)^2 + (130)^2 \quad (28)
\]
Expanding and simplifying:

\[
A - 40x > 800 \ln[1.5] - (150)^2 + (130)^2 \quad (29)
\]

\[
B < x < A
\]

\[
B > \frac{800 \ln[1.5] - 5600}{-40} \quad (30)
\]

\[
B \quad x > 131.9 \quad (31)
\]

\[
A < x < B
\]
For the n-d case, where
\[ p(x|A) = \mathcal{N}(\mu_A, \Sigma_A^2) \]
and
\[ p(x|B) = \mathcal{N}(\mu_B, \Sigma_B^2), \]

\[
P(A) \exp \left[ -\frac{1}{2} (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) \right] \frac{\exp \left[ -\frac{1}{2} (x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) \right]}{P(B) \exp \left[ -\frac{1}{2} (x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) \right]} \]}

\[
\frac{\exp \left[ -\frac{1}{2} (x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) \right]}{\exp \left[ -\frac{1}{2} (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) \right]}
\]

\[
A > \frac{\exp \left[ -\frac{1}{2} (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) \right]}{\exp \left[ -\frac{1}{2} (x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) \right]} B \]

\[
P(A) \exp \left[ -\frac{1}{2} (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) \right] > \frac{\exp \left[ -\frac{1}{2} (x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) \right]}{\exp \left[ -\frac{1}{2} (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) \right]} P(B)
\]

\[
P(B) \exp \left[ -\frac{1}{2} (x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) \right] < \frac{\exp \left[ -\frac{1}{2} (x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) \right]}{\exp \left[ -\frac{1}{2} (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) \right]} P(A)
\]
Taking the log and simplifying:

\[
(x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) - (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) > 2 \ln \frac{|\Sigma_A|^{1/2} P(B)}{|\Sigma_B|^{1/2} P(A)}
\]

(34)

\[
(x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) - (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) < 2 \ln \frac{P(B)}{P(A)} + \ln \left( \frac{|\Sigma_A|}{|\Sigma_B|} \right)
\]

(35)

Looks familiar?
What is the MAP decision boundaries if our classes can be characterized by normal distributions?

\[ x^T Q_0 x + Q_1 x + Q_2 + 2Q_3 + Q_4 = 0, \]  

where,

\[ Q_0 = S_A^{-1} - S_B^{-1} \]  

\[ Q_1 = 2[m_B^T S_B^{-1} - m_A^T S_A^{-1}] \]  

\[ Q_2 = m_A^T S_A^{-1} m_A - m_B^T S_B^{-1} m_B \]  

\[ Q_3 = - \ln \left( \frac{P(B)}{P(A)} \right) \]  

\[ Q_4 = - \ln \left( \frac{|S_A|}{|S_B|} \right) \]
MAP Classifier: Example

- Suppose we are given the following statistical information about the classes:
  - Class A: \( \mathbf{m}_A = [0 \ 0]^T \), \( \mathbf{S}_A = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix} \), \( P(A)=0.6 \).
  - Class B: \( \mathbf{m}_B = [0 \ 0]^T \), \( \mathbf{S}_B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \), \( P(B)=0.4 \).

- Suppose we wish to build a MAP classifier.
  - Compute the decision boundary.
Step 1: Compute $S_A^{-1}$ and $S_B^{-2}$:

$$S_A^{-1} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/4 \end{bmatrix} \quad S_B^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad (42)$$

Step 2: Compute $Q_0, Q_1, Q_2, Q_3$:

$$Q_0 = S_A^{-1} - S_B^{-1} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/4 \end{bmatrix} - \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \begin{bmatrix} -3/4 & 0 \\ 0 & -3/4 \end{bmatrix} \quad (43)$$

$$Q_1 = 2[m_B^T S_B^{-1} - m_A^T S_A^{-1}] = 0 \quad (44)$$

$$Q_2 = m_A^T S_A^{-1} m_A - m_B^T S_B^{-1} m_B = 0 \quad (45)$$
MAP Classifier: Example

Step 2: Compute \( Q_0, Q_1, Q_2, Q_3: \)

\[
Q_3 = - \ln \left[ \frac{P(B)}{P(A)} \right] = \ln \left[ \frac{0.4}{0.6} \right] = - \ln(4/6) \quad (46)
\]

\[
Q_4 = - \ln \left[ \frac{|S_A|}{|S_B|} \right] = \ln \left[ \frac{(1/4)(1/4) - (0)(0)}{(1)(1) - (0)(0)} \right] = - \ln(1/16). \quad (47)
\]

Step 3: Plugging in \( Q_0, Q_1, Q_2, Q_3 \) gives us:

\[
x^T Q_0 x + Q_1 x + Q_2 + 2Q_3 + Q_4 = 0, \quad (48)
\]

\[
x^T \begin{bmatrix} -3/4 & 0 \\ 0 & -3/4 \end{bmatrix} x - 2 \ln(4/6) - \ln(1/16) = 0, \quad (49)
\]
MAP Classifier: Example

Simplifying gives us:

\[
([x_1 \ x_2]^T)^T \begin{bmatrix} -3/4 & 0 \\ 0 & -3/4 \end{bmatrix} [x_1 \ x_2]^T - 2 \ln(4/6) - \ln(1/16) = 0,
\]

\[
\begin{bmatrix} -3/4 x_1 - 3/4 x_2 \\ -3/4 x_1 - 3/4 x_2 \end{bmatrix}^T [x_1 \ x_2]^T + 549/677 + 2731/985 = 0,
\]

\[
-3/4 x_1^2 - 3/4 x_2^2 + 1609/449 = 0,
\]

The final MAP decision boundary is:

\[
x_1^2 + x_2^2 = 2131/446,
\]

This is just a circle centered at \((x_1, x_2) = (0, 0)\) with a radius of 2011/920.
Relationship between MICD and MAP Classifiers for Normal Distributions

You will notice that the terms on the right has the same form as the MICD distance metric!

\[
(x - \mu_B)^T \Sigma_B^{-1} (x - \mu_B) - (x - \mu_A)^T \Sigma_A^{-1} (x - \mu_A) > \frac{A}{B} \ln \left( \frac{P(B)}{P(A)} \right) + \ln \left( \frac{|\Sigma_A|}{|\Sigma_B|} \right)
\]

(54)

\[
d_{MICD}^2(x, \mu_B, \Sigma_B) - d_{MICD}^2(x, \mu_A, \Sigma_A) > \frac{A}{B} \ln \left( \frac{P(B)}{P(A)} \right) + \ln \left( \frac{|\Sigma_A|}{|\Sigma_B|} \right)
\]

(55)

If \(2 \ln \left( \frac{P(B)}{P(A)} \right) + \ln \left( \frac{|\Sigma_A|}{|\Sigma_B|} \right) = 0\), then the MAP classifier becomes just the MICD classifier!
Therefore, the MICD is only optimal in terms of probability of error only if we have multivariate Normal distributions $\mathcal{N}(\mu, \Sigma)$ that have:

- Equal a priori probabilities ($P(A) = P(B)$)
- Equal volume cases ($|\Sigma_A| = |\Sigma_B|$)

If that is the case, what’s so special about

$$2 \ln \left[ \frac{P(B)}{P(A)} \right] + \ln \left[ \frac{|\Sigma_A|}{|\Sigma_B|} \right]?$$

- First term $2 \ln \left[ \frac{P(B)}{P(A)} \right]$ biases decision in favor of more likely class according to a priori probabilities
- Second term $\ln \left[ \frac{|\Sigma_A|}{|\Sigma_B|} \right]$ biases decision in favor of class with smaller volume ($|\Sigma|$)
So under what circumstance does MAP classifier perform better than MICD?

Recall the case where we have only one feature \( n = 1 \), \( m = 0 \), and \( s_A \neq s_B \).

The MICD classification rule for this case is:

\[
\left( \frac{1}{s^2_B} - \frac{1}{s^2_A} \right) x^2 > 0 \quad \text{(56)}
\]

\[
\left( \frac{1}{s^2_A} \right) x^2 < \left( \frac{1}{s^2_B} \right) x^2 \quad \text{(57)}
\]

\[
s^2_A > s^2_B \quad \text{(58)}
\]

The MICD classification rule decides in favor of the class with the largest variance, regardless of \( x \).
The MAP classification rule for this case is:

\[
(1/s_B^2 - 1/s_A^2)x^2 > 2 \ln \left( \frac{P(A)}{P(B)} \right) + \ln \left( \frac{s_A^2}{s_B^2} \right) \tag{59}
\]

If \( P(A) = P(B) \)

\[
(1/s_B^2 - 1/s_A^2)x^2 > \ln \left( \frac{s_A^2}{s_B^2} \right) \tag{60}
\]
Looking at the MAP classification rule:

\[(1/s_B^2 - 1/s_A^2)x^2 > \ln \left[ \frac{s_A^2}{s_B^2} \right] \quad (61)\]

At the mean \( m = 0 \),

\[0 > \ln \left[ \frac{s_A^2}{s_B^2} \right] \quad (62)\]

- If \( s_A^2 < s_B^2 \), the log term is negative and favors class A
- If \( s_B^2 < s_A^2 \), the log term is positive and favors class B

Therefore, the MAP classification rule decides in favor of class with the lowest variance close to the mean, and favors the class with highest variance beyond a certain point in both directions.
How do we quantify how well the Bayes classifier works?
Since the Bayes classifier minimizes the probability of error, one way to analyze how well it does is to compute the probability of error $P(\epsilon)$ itself.
Allows us to see the theoretical limit on the expected performance, under the assumption of known probability density functions.
For any pattern $\mathbf{x}$ such that $P(A|\mathbf{x}) > P(B|\mathbf{x})$:
- $\mathbf{x}$ is classified as part of class $A$
- The probability of error of classifying $\mathbf{x}$ as $A$ is $P(B|\mathbf{x})$

Therefore, naturally, for any given $\mathbf{x}$ the probability of error $P(\epsilon|\mathbf{x})$ is:

$$P(\epsilon|\mathbf{x}) = \min [P(A|\mathbf{x}), P(B|\mathbf{x})]$$

(63)

**Rationale:** Since we always chose the maximum posterior probability as our class, the minimum posterior probability would be the probability of choosing incorrectly.
Recall our previous example of a 1-D case:

\[ p(x|A) = \mathcal{N}(\mu_A, \sigma_A^2) = \frac{1}{\sqrt{2\pi\sigma_A}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \right] \]  

(64)

\[ p(x|B) = \mathcal{N}(\mu_B, \sigma_B^2) = \frac{1}{\sqrt{2\pi\sigma_B}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_B}{\sigma_B} \right)^2 \right] \]  

(65)

\[ \mu_A = 130, \ \mu_B = 150, \ P(A) = 0.4, \ P(B) = 0.6, \ \sigma_A = \sigma_B = 20. \]

For \( x = 140 \), what is the probability of error \( P(\epsilon|x) \)?
Recall the MAP classifier for this scenario:

\begin{equation}
\begin{aligned}
B \\
\quad x > 131.9 \\
\quad A
\end{aligned}
\end{equation}

Based on this MAP classifier, the pattern $x = 140$ belongs to class $A$.

Given the probability of error $P(\epsilon|x)$ is:

\begin{equation}
P(\epsilon|x) = \min [P(A|x), P(B|x)]
\end{equation}

Since $B$ gives the maximum probability, the minimum probability would be $P(A|x)$. 
Therefore, \( P(\epsilon | x) \) for \( x = 140 \) is:

\[
P(\epsilon | x) \bigg|_{x=140} = P(A | x) \bigg|_{x=140} = \frac{P(x | A)P(A)}{P(x | A)P(A) + P(x | B)P(B)} \bigg|_{x=140}
\]

\[
P(\epsilon | x) \bigg|_{x=140} = \frac{26/1477(0.4)}{(26/1477)0.4 + (26/1477)(0.6)}
\]

\[
P(\epsilon | x) \bigg|_{x=140} = 0.4.
\]
Expected probability of error

Now that we know the probability of error for a given $x$, denoted as $P(\epsilon|x)$, the expected probability of error $P(\epsilon)$ can be found as:

$$P(\epsilon) = \int P(\epsilon|x)p(x)dx$$  \hspace{1cm} (71)

$$P(\epsilon) = \int \min [P(A|x), P(B|x)] p(x)dx$$  \hspace{1cm} (72)

In terms of class PDFs:

$$P(\epsilon) = \int \min [P(x|A)P(A), P(x|B)P(B)]dx$$  \hspace{1cm} (73)
Expected probability of error

Now if we were to define decision regions $R_A$ and $R_B$:

- $R_A = x$ such that $P(A|x) > P(B|x)$
- $R_B = x$ such that $P(B|x) > P(A|x)$

The expected probability of error can be defined as:

$$P(\epsilon) = \int_{R_A} P(x|B)P(B)dx + \int_{R_B} P(x|A)P(A)dx \quad (74)$$

Rationale: For all patterns in $R_A$, the probability of $A$ will be the maximum between $A$ and $B$, so the probability of error of patterns in $R_A$ is just the minimum probability (in this case, the probability of $B$), and vice versa.
Example 1: univariate Normal, equal variance, equally likely two class problem:

- \( n = 1, \ P(A) = P(B) = 0.5, \ \sigma_A = \sigma_B = \sigma, \ \mu_A < \mu_B \)
- Likelihood:

\[
p(x|A) = \mathcal{N}(\mu_A, \sigma_A^2) = \frac{1}{\sqrt{2\pi\sigma_A}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \right] \tag{75}
\]
\[
p(x|B) = \mathcal{N}(\mu_B, \sigma_B^2) = \frac{1}{\sqrt{2\pi\sigma_B}} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_B}{\sigma_B} \right)^2 \right] \tag{76}
\]

Find \( p(\epsilon) \)
Expected probability of error

- Recall for the case of equally likely, equi-variance classes, the MAP decision boundary reduces to a threshold midway between the means.

\[ x = \frac{(\mu_B + \mu_A)}{2} \]  \hspace{1cm} (77)

- Since \( \mu_A < \mu_B \), this gives us the following decision regions \( R_A \) and \( R_B \):
  - \( R_A = x \) such that \( x < \frac{(\mu_B + \mu_A)}{2} \)
  - \( R_B = x \) such that \( x > \frac{(\mu_B + \mu_A)}{2} \)
Based on decision regions $R_A$, $R_B$, $P(A)$, $P(B)$, $P(x|A)$, $P(x|B)$, $\mu_B$, $\mu_A$, the expected probability of error $P(\epsilon)$ becomes

$$P(\epsilon) = \int_{R_A} P(B)P(x|B)dx + \int_{R_B} P(A)P(x|A)dx \quad (78)$$

$$P(\epsilon) = \frac{1}{2} \int_{-\infty}^{\frac{(\mu_B+\mu_A)}{2}} P(x|B)dx + \frac{1}{2} \int_{(\mu_B+\mu_A)}^{\infty} P(x|A)dx \quad (79)$$

$$P(\epsilon) = \frac{1}{2} \int_{-\infty}^{\frac{(\mu_B+\mu_A)}{2}} \mathcal{N}(\mu_B, \sigma^2)dx + \frac{1}{2} \int_{(\mu_B+\mu_A)}^{\infty} \mathcal{N}(\mu_A, \sigma^2)dx \quad (80)$$
Since the two classes are symmetric ($P(\epsilon|A) = P(\epsilon|B)$),

$$P(\epsilon) = \frac{1}{2} \int_{-\infty}^{\frac{(\mu_B + \mu_A)}{2}} \mathcal{N}(\mu_B, \sigma^2) dx + \frac{1}{2} \int_{\frac{(\mu_B + \mu_A)}{2}}^{\infty} \mathcal{N}(\mu_A, \sigma^2) dx$$

(81)

$$P(\epsilon) = \int_{\frac{(\mu_B + \mu_A)}{2}}^{\infty} \mathcal{N}(\mu_A, \sigma^2) dx$$

(82)

$$P(\epsilon) = \int_{\frac{(\mu_B + \mu_A)}{2}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_A} \exp \left[ -\frac{1}{2} \left( \frac{x - \mu_A}{\sigma_A} \right)^2 \right] dx$$

(83)
Expected probability of error

- Doing a change of variables, where \( y = \frac{x - \mu_A}{\sigma} \), \( dx = \sigma dy \),

\[
P(\epsilon) = \int_{(\mu_B - \mu_A) / 2\sigma}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2}y^2 \right] dy \quad (84)
\]

- This corresponds to an integral over a normalized \((\mathcal{N}(0, 1))\) Normal random variable:

\[
Q(\alpha) = \int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp \left[ -\frac{1}{2}y^2 \right] dy \quad (85)
\]

- Plugging \( Q \) in gives us the final expected probability of error \( P(\epsilon) \):

\[
P(\epsilon) = Q\left( \frac{\mu_B - \mu_A}{2\sigma} \right) \quad (86)
\]
Visualization of $P(\epsilon)$:

$P(\epsilon)$ is essentially the shaded area.
Expected probability of error

- Observations:
  - As the distance between the means increase, the shaded area becomes monotonically smaller and the expected probability of error $P(\epsilon)$ monotonically decreases.
  - At $\alpha = 0$, $\mu_A = \mu_B = 0$ and $P(\epsilon) = 1/2$ (makes sense since the distributions completely overlap, and you have a 50/50 chance of either class)
  - $\lim_{\alpha \to \infty} P(\epsilon) = 0$. 
Expected probability of error

- For cases where $P(A) \neq P(B)$ or $\sigma_A \neq \sigma_B$, the decision boundary change AND an additional boundary is introduced!

- Luckily, $P(\epsilon)$ can still be expressed using the $Q(\alpha)$ function with appropriate change of variables.
Example:

\[ P(\epsilon) \text{ is essentially the shaded area.} \]

\[ P(\epsilon) = P(A)Q(\alpha_1) + P(B)[Q(\alpha_3) - Q(\alpha_4)] + P(A)Q(\alpha_2) \]
Let’s take a look at the multivariate case (n>1)

For $p(x|A) = \mathcal{N}(\mu_A, \Sigma)$, $p(x|B) = \mathcal{N}(\mu_B, \Sigma)$, $P(A) = P(B)$, it can be shown that:

$$P(\epsilon) = Q(d_M(\mu_A, \mu_B)/2)$$  \hspace{1cm} (87)

where $d_M(\mu_A, \mu_B)$ is the Mahalanobis distance between the classes.

$$d_M(\mu_A, \mu_B) = \left[(\mu_A - \mu_B)^T \Sigma^{-1} (\mu_A - \mu_B)\right]^{1/2}$$  \hspace{1cm} (88)
**Expected probability of error**

- Why is $P(\epsilon)$ like that for this case?
  - Remember that for all cases where the covariance matrices AND the prior probabilities are the same, the decision boundary between the classes is always a straight line in hyperspace that is:
    - sloped based on $\Sigma$ (since our orthonormal whitening transform is identical for both classes)
    - intersects with the midpoint of the line segment between $\mu_A$ and $\mu_B$
  - The probability of error is just the area under $P(x|A)p(A)$ on the class $B$ side of this decision boundary PLUS the area under $P(x|B)p(B)$ on the class $A$ side of this decision boundary.
Example of non-Gaussian density functions:

Suppose two classes have density functions and a priori probabilities:

\[
p(x|C_1) = \begin{cases} 
    ce^{-\lambda x} & 0 \leq x \leq 1 \\
    0 & \text{else}
\end{cases} \tag{89}
\]

\[
p(x|C_2) = \begin{cases} 
    ce^{-\lambda(1-x)} & 0 \leq x \leq 1 \\
    0 & \text{else}
\end{cases} \tag{90}
\]

\[
P(C_1) = P(C_2) = \frac{1}{2} \tag{91}
\]

where \( c = \frac{\lambda}{1-e^{-\lambda}} \) is just the appropriate constant to normalize the PDF.
Therefore, the expected probability of error is:

\[ P(\epsilon) = \int \min \left[ P(x|C_1)P(C_1), P(x|C_2)P(C_2) \right] dx \]  

(92)

\[ P(\epsilon) = \int_{R_{C_1}} P(x|C_2)P(C_2)dx + \int_{R_{C_2}} P(x|C_1)P(C_1)dx \]  

(93)

\[ P(\epsilon) = \int_0^{0.5} 0.5 P(x|C_2)dx + \int_{0.5}^{1.0} 0.5 P(x|C_1)dx \]  

(94)

Because of symmetry between the two classes \( P(\epsilon|C_1) = P(\epsilon|C_2) \),

\[ P(\epsilon) = \int_{0.5}^{1.0} ce^{-\lambda x} dx \]  

(95)

\[ P(\epsilon) = \frac{C}{\lambda} \left[ e^{-\lambda/2} - e^{-\lambda} \right] \]  

(96)
(b) Find $P(\epsilon|x)$:

From the decision boundary and decision regions we determined in (a),

$$p(\epsilon|x) = \begin{cases} 
P(C_2|x) & 0 \leq x \leq 1/2 \\
P(C_1|x) & 1/2 \leq x \leq 1 
\end{cases}$$  (97)

$$p(\epsilon|x) = \begin{cases} 
\frac{P(x|C_2)P(C_2)}{P(x)} & 0 \leq x \leq 1/2 \\
\frac{P(x|C_1)P(C_1)}{P(x)} & 1/2 \leq x \leq 1 
\end{cases}$$  (98)

$$p(\epsilon|x) = \begin{cases} 
\frac{e^{-\lambda x}0.5}{e^{-\lambda x}+e^{-\lambda(1-x)}} & 0 \leq x \leq 1/2 \\
\frac{e^{-\lambda(1-x)}0.5}{e^{-\lambda x}+e^{-\lambda(1-x)}} & 1/2 \leq x \leq 1 
\end{cases}$$  (99)
Error bounds

- In practice, the exact $P(\epsilon)$ is only easy to compute for simple cases as shown before.
- So how can we quantify the probability of error in such cases?
- Instead of finding the exact $P(\epsilon)$, we determine the \textbf{bounds} on $P(\epsilon)$, which are:
  - Easier to compute
  - Leads to estimates of classifier performance
Bhattacharyya bound

Using the following inequality:

\[ \min[a, b] \leq \sqrt{(a, b)} \] (100)

The following holds true:

\[ P(\epsilon) = \int \min [P(x|A)P(A), P(x|B)P(B)] dx \] (101)

\[ P(\epsilon) \leq \sqrt{P(A)P(B)} \int \sqrt{P(x|A)P(x|B)} dx \] (102)

What’s so special about this?
Answer: You don’t need the actual decision regions to compute this!
Bhattacharyya bound

- Since $P(A) + P(B) = 1$ and the Bhattacharyya coefficient $\rho$ can be defined as:

\[
\rho = \int \sqrt{P(x|A)P(x|B)} \, dx
\]  

(103)

- The upper bound (Bhattacharyya bound) of $P(\epsilon)$ can be written as

\[
P(\epsilon) \leq \frac{1}{2} \rho
\]  

(104)
Example: Consider a classifier for a two class problem. Both classes are multivariate normal. When both classes are a priori equally likely, the Bhattacharrya bound is $P(\epsilon) \leq 0.3$.

New information is specified, such that we are told that the a priori probabilities of the two classes are 0.2 and 0.8, for $A$ and $B$ respectively.

What is the new upper bound for the probability of error?
Bhattacharrya bound Example

Step 1: Based on old bound, compute the Bhattacharrya coefficient

\[ P(\epsilon) = 0.3 \leq \sqrt{P(A)P(B)} \int \sqrt{P(x|A)P(x|B)} \, dx \]  \hfill (105)

\[ \frac{0.3}{\sqrt{P(A)P(B)}} \leq \int \sqrt{P(x|A)P(x|B)} \, dx \]  \hfill (106)

\[ \rho = \int \sqrt{P(x|A)P(x|B)} \, dx \geq \frac{0.3}{\sqrt{0.5 \times 0.5}} = 0.6 \]  \hfill (107)
Step 2: Based on Bhattacharrya coefficient $\rho$ and new priors $P(A) = 0.2$ and $P(B) = 0.8$, the new upper bound can be computed as:

$$P(\epsilon) \leq \sqrt{P(A)P(B)} \int \sqrt{P(x|A)P(x|B)} \, dx$$  \hspace{1cm} (108)

$$P(\epsilon) \leq \sqrt{0.8 \times 0.2 \times \rho}$$  \hspace{1cm} (109)

$$P(\epsilon) \leq \sqrt{0.8 \times 0.2 \times 0.6} = 0.24$$  \hspace{1cm} (110)