SYDE 372 - Winter 2011
Introduction to Pattern Recognition

Probability Measures for Classification: Part II

Alexander Wong

Department of Systems Design Engineering
University of Waterloo

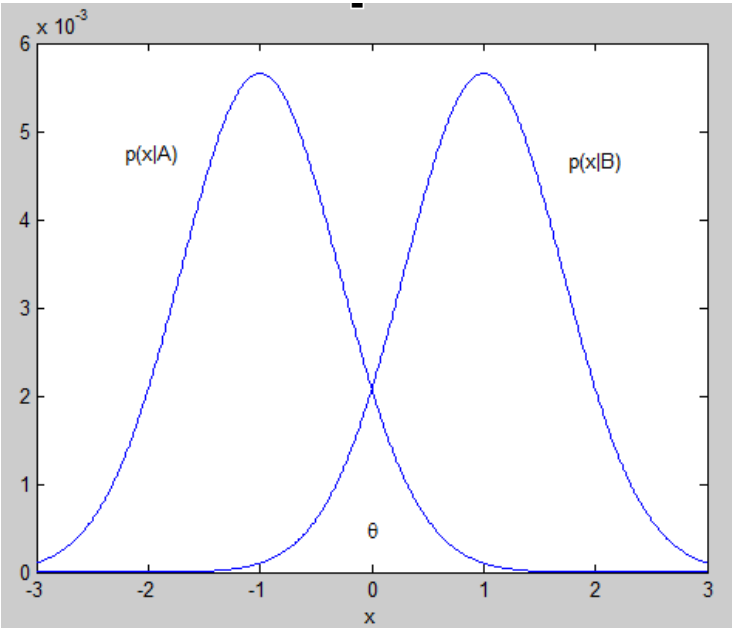
Outline

MAP Classifier for Normal Distributions

- By far the most popular conditional class distribution model is the Gaussian distribution:

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

and $p(x|B) = \mathcal{N}(\mu_B, \sigma_B^2)$.



MAP Classifier for Normal Distributions

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

$$\frac{\mathcal{N}(\mu_A, \sigma_A^2)}{\mathcal{N}(\mu_B, \sigma_B^2)} \underset{B}{\overset{A}{\gtrless}} \frac{P(B)}{P(A)} \tag{2}$$

$$\frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu_A}{\sigma_A}\right)^2\right]}{\exp\left[-\frac{1}{2}\left(\frac{x-\mu_B}{\sigma_B}\right)^2\right]} \underset{B}{\overset{A}{\gtrless}} \frac{\sigma_A P(B)}{\sigma_B P(A)} \tag{3}$$

MAP Classifier for Normal Distributions

- In log-likelihood form:

$$\frac{\exp\left[-\frac{1}{2}(\frac{x-\mu_A}{\sigma_A})^2\right]}{\exp\left[-\frac{1}{2}(\frac{x-\mu_A}{\sigma_A})^2\right]} \underset{B}{\overset{A}{\gtrless}} \frac{\sigma_A P(B)}{\sigma_B P(A)} \tag{4}$$

$$\left[-\frac{1}{2}(\frac{x-\mu_A}{\sigma_A})^2\right] - \left[-\frac{1}{2}(\frac{x-\mu_B}{\sigma_B})^2\right] \underset{B}{\overset{A}{\gtrless}} \ln\left[\sigma_A P(B)\right] - \ln\left[\sigma_B P(A)\right] \tag{5}$$

$$\left[(\frac{x-\mu_B}{\sigma_B})^2\right] - \left[(\frac{x-\mu_A}{\sigma_A})^2\right] \underset{B}{\overset{A}{\gtrless}} 2\left[\ln\left[\sigma_A P(B)\right] - \ln\left[\sigma_B P(A)\right]\right] \tag{6}$$

MAP Classifier for Normal Distributions

- Giving us the final form:

$$\left[\left(\frac{x-\mu_B}{\sigma_B}\right)^2\right] - \left[\left(\frac{x-\mu_A}{\sigma_A}\right)^2\right] \underset{B}{\overset{A}{\gtrless}} 2\left[\ln\left[\sigma_A P(B)\right] - \ln\left[\sigma_B P(A)\right]\right] \tag{7}$$

- Does this look familiar?

MAP Classifier for Normal Distributions

- The decision boundary (threshold) for the MAP classifier where $P(x|A)$ and $P(x|B)$ are Gaussian distributions can be find by solving the following expression for x:

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

$$x^2\left[\frac{1}{\sigma_B^2} - \frac{1}{\sigma_A^2}\right] - 2x\left[\frac{\mu_B}{\sigma_B^2} - \frac{\mu_A}{\sigma_A^2}\right] + \frac{\mu_B^2}{\sigma_B^2} - \frac{\mu_A^2}{\sigma_A^2} = 2\ln\left[\frac{\sigma_A P(B)}{\sigma_B P(A)}\right] \quad (9)$$

MAP Classifier for Normal Distributions

- For case where $\sigma_A = \sigma_B$, $P(A) = P(B) = \frac{1}{2}$:

$$x^2\left[\frac{1}{\sigma_B^2} - \frac{1}{\sigma_A^2}\right] - 2x\left[\frac{\mu_B}{\sigma_B^2} - \frac{\mu_A}{\sigma_A^2}\right] + \frac{\mu_B^2}{\sigma_B^2} - \frac{\mu_A^2}{\sigma_A^2} = 2\ln\left[\frac{\sigma_A P(B)}{\sigma_B P(A)}\right] \tag{10}$$

$$x^2(\sigma_A^2 - \sigma_B^2) - 2x(\mu_B\sigma_A^2 - \mu_A\sigma_B^2) + (\mu_B^2\sigma_A^2 - \mu_A^2\sigma_B^2) = 2\ln[1] \tag{11}$$

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

$$x = \frac{(\mu_B^2\sigma_A^2 - \mu_A^2\sigma_A^2)}{2(\mu_B\sigma_A^2 - \mu_A\sigma_A^2)} \tag{12}$$

$$x = \frac{(\mu_B^2 - \mu_A^2)}{2(\mu_B - \mu_A)} \tag{13}$$

MAP Classifier for Normal Distributions

- Since $(a^2 - b^2) = (a - b)(a + b)$:

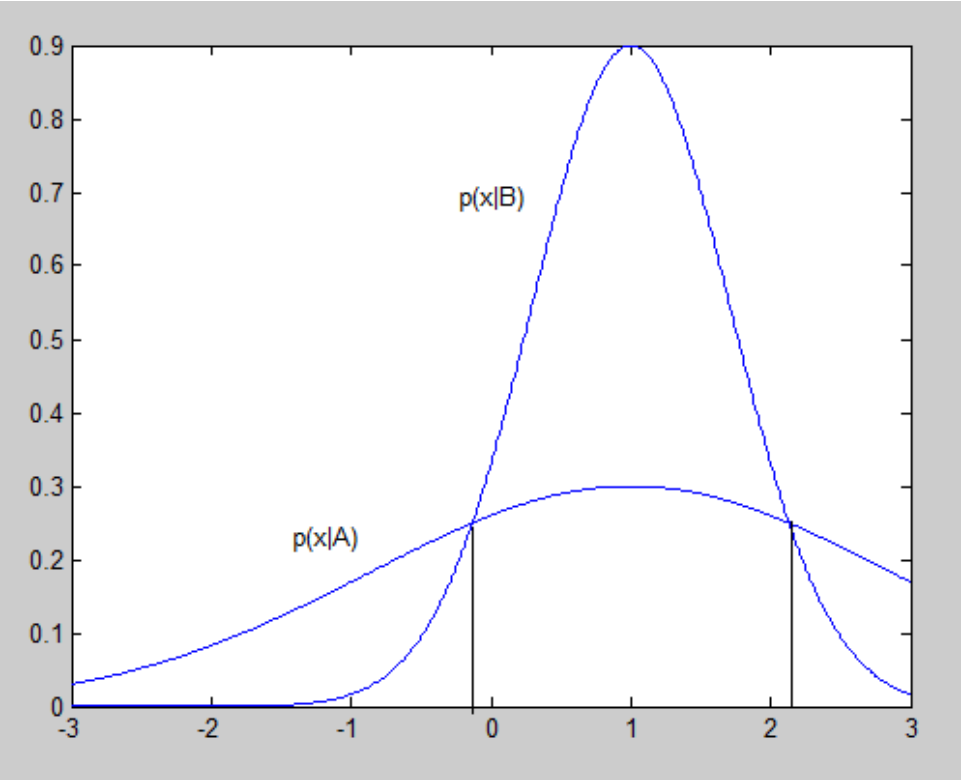
$$x = \frac{(\mu_B - \mu_A)(\mu_B + \mu_A)}{2(\mu_B - \mu_A)} \tag{14}$$

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

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

MAP Classifier for Normal Distributions

- 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.



MAP Classifier for Normal Distributions

- Example of a 1-D case:
- Suppose that, given a pattern $\underline{x}$, we wish to classify it as one of two classes: class A and class B.
- Suppose the two classes have patterns $\underline{x}$ which are normally distributed as follows:

$$p(x|A) = \mathcal{N}(\mu_A, \sigma_A^2) = \frac{1}{\sqrt{2\pi}\sigma_A} \exp\left[-\frac{1}{2}(\frac{x-\mu_A}{\sigma_A})^2\right] \tag{16}$$

$$p(x|B) = \mathcal{N}(\mu_B, \sigma_B^2) = \frac{1}{\sqrt{2\pi}\sigma_B} \exp\left[-\frac{1}{2}(\frac{x-\mu_B}{\sigma_B})^2\right] \tag{17}$$

$\mu_A = 130$, $\mu_B = 150$.

MAP Classifier for Normal Distributions

- **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:

$$\frac{\mathcal{N}(\mu_A, \sigma_A^2)}{\mathcal{N}(\mu_B, \sigma_B^2)} \underset{B}{\overset{A}{\gtrless}} \frac{P(B)}{P(A)} \tag{18}$$

$$\frac{\exp\left[-\frac{1}{2}\left(\frac{x-\mu_A}{\sigma_A}\right)^2\right]}{\exp\left[-\frac{1}{2}\left(\frac{x-\mu_B}{\sigma_B}\right)^2\right]} \underset{B}{\overset{A}{\gtrless}} \frac{\sigma_A P(B)}{\sigma_B P(A)} \tag{19}$$

MAP Classifier for Normal Distributions

- Plugging in μ_A, μ_B, and $\sigma_A = \sigma_B = \sigma$:

$$\frac{\exp\left[-\frac{1}{2}\left(\frac{x-130}{20}\right)^2\right]}{\exp\left[-\frac{1}{2}\left(\frac{x-150}{20}\right)^2\right]} \underset{B}{\overset{A}{\gtrless}} \frac{P(B)}{P(A)} \tag{20}$$

- Taking the log:

$$\left[-\frac{1}{2}(x-130)^2\right] - \left[-\frac{1}{2}(x-150)^2\right] \underset{B}{\overset{A}{\gtrless}} 2(20^2)\ln\left[P(B)\right] - \ln\left[P(A)\right] \tag{21}$$

$$\left[(x-150)^2\right] - \left[(x-130)^2\right] \underset{B}{\overset{A}{\gtrless}} 800\left[\ln\left[P(B)\right] - \ln\left[P(A)\right]\right] \tag{22}$$

MAP Classifier for Normal Distributions

- The prior probability $P(A)$ and $P(B)$ can be determined as:

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

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

$$\left[(x-150)^2\right] - \left[(x-130)^2\right] \underset{B}{\overset{A}{\gtrless}} 800\left[\ln\left[0.6/0.4\right]\right] \tag{24}$$

$$\left[(x-150)^2\right] - \left[(x-130)^2\right] \underset{B}{\overset{A}{\gtrless}} 800\left[\ln\left[1.5\right]\right] \tag{25}$$

MAP Classifier for Normal Distributions

- Expanding and simplifying:

$$\left[(x-150)^2\right] - \left[(x-130)^2\right] \underset{B}{\overset{A}{\gtrless}} 800\,[\ln[1.5]] \tag{26}$$

$$(x^2 - 300x + (150)^2) - (x^2 - 260x + (130)^2) \underset{B}{\overset{A}{\gtrless}} 800\,[\ln[1.5]] \tag{27}$$

$$-40x \underset{B}{\overset{A}{\gtrless}} 800\,[\ln[1.5]] - (150)^2 + (130)^2 \tag{28}$$

MAP Classifier for Normal Distributions

- Expanding and simplifying:

$$-40x \underset{B}{\overset{A}{\gtrless}} 800\left[\ln\left[1.5\right]\right] - (150)^2 + (130)^2 \tag{29}$$

$$x \underset{A}{\overset{B}{\gtrless}} \frac{800\left[\ln\left[1.5\right]\right] - 5600}{-40} \tag{30}$$

$$x \underset{A}{\overset{B}{\gtrless}} 131.9 \tag{31}$$

MAP Classifier for Normal Distributions

- For the n-d case, where $p(\underline{x}|A) = \mathcal{N}(\underline{\mu}_A, \Sigma_A^2)$ and $p(\underline{x}|B) = \mathcal{N}(\underline{\mu}_B, \Sigma_B^2)$,

$$\frac{P(A)\exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_A)^T\Sigma_A^{-1}(\underline{x}-\underline{\mu}_A)\right]}{(2\pi)^{\frac{n}{2}}|\Sigma_A|^{1/2}} \underset{B}{\overset{A}{\gtrless}} \frac{P(B)\exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_B)^T\Sigma_B^{-1}(\underline{x}-\underline{\mu}_B)\right]}{(2\pi)^{\frac{n}{2}}|\Sigma_B|^{1/2}} \tag{32}$$

$$\frac{\exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_A)^T\Sigma_A^{-1}(\underline{x}-\underline{\mu}_A)\right]}{\exp\left[-\frac{1}{2}(\underline{x}-\underline{\mu}_B)^T\Sigma_B^{-1}(\underline{x}-\underline{\mu}_B)\right]} \underset{B}{\overset{A}{\gtrless}} \frac{|\Sigma_A|^{1/2}P(B)}{|\Sigma_B|^{1/2}P(A)} \tag{33}$$

MAP Classifier for Normal Distributions

- Taking the log and simplifying:

$$(\underline{x}-\underline{\mu}_B)^T\Sigma_B^{-1}(\underline{x}-\underline{\mu}_B)-(\underline{x}-\underline{\mu}_A)^T\Sigma_A^{-1}(\underline{x}-\underline{\mu}_A) \underset{B}{\overset{A}{\gtrless}} 2\ln\left[\frac{|\Sigma_A|^{1/2}P(B)}{|\Sigma_B|^{1/2}P(A)}\right] \tag{34}$$

$$(\underline{x}-\underline{\mu}_B)^T\Sigma_B^{-1}(\underline{x}-\underline{\mu}_B)-(\underline{x}-\underline{\mu}_A)^T\Sigma_A^{-1}(\underline{x}-\underline{\mu}_A) \underset{B}{\overset{A}{\gtrless}} 2\ln\left[\frac{P(B)}{P(A)}\right]+\ln\left[\frac{|\Sigma_A|}{|\Sigma_B|}\right] \tag{35}$$

- Looks familiar?

MAP Decision Boundaries for Normal Distribution

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

$$\underline{x}^T Q_0 \underline{x} + Q_1 \underline{x} + Q_2 + 2Q_3 + Q_4 = 0, \tag{36}$$

where,

$$Q_0 = S_A^{-1} - S_B^{-1} \tag{37}$$

$$Q_1 = 2[\underline{m}_B^T S_B^{-1} - \underline{m}_A^T S_A^{-1}] \tag{38}$$

$$Q_2 = \underline{m}_A^T S_A^{-1} \underline{m}_A - \underline{m}_B^T S_B^{-1} \underline{m}_B \tag{39}$$

$$Q_3 = -\ln\left[\frac{P(B)}{P(A)}\right] \tag{40}$$

$$Q_4 = -\ln\left[\frac{|S_A|}{|S_B|}\right] \tag{41}$$

MAP Classifier: Example

- Suppose we are given the following statistical information about the classes:
  - Class A: $\underline{m}_A = [0 \;\; 0]^T, S_A = \begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$, P(A)=0.6.
  - Class B: $\underline{m}_B = [0 \;\; 0]^T, S_B = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$, P(B)=0.4.
- Suppose we wish to build a MAP classifier.
  - Compute the decision boundary.

MAP Classifier: Example

- Step 1: Compute ${S_A}^{-1}$ and ${S_B}^{-2}$:

$$ {S_A}^{-1} = \begin{bmatrix} 1/4 & 0 \\ 0 & 1/4 \end{bmatrix} \; {S_B}^{-1} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{42} $$

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

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

$$ Q_1 = 2[\underline{m}_B^T S_B^{-1} - \underline{m}_A^T S_A^{-1}] = 0 \tag{44} $$

$$ Q_2 = \underline{m}_A^T S_A^{-1} \underline{m}_A - \underline{m}_B^T S_B^{-1} \underline{m}_B = 0 \tag{45} $$

MAP Classifier: Example

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

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

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

- Step 3: Plugging in Q_0, Q_1, Q_2, Q_3 gives us:

$$\underline{x}^T Q_0 \underline{x} + Q_1 \underline{x} + Q_2 + 2Q_3 + Q_4 = 0, \tag{48}$$

$$\underline{x}^T \begin{bmatrix} -3/4 & 0 \\ 0 & -3/4 \end{bmatrix} \underline{x} - 2\ln(4/6) - \ln(1/16) = 0, \tag{49}$$

MAP Classifier: Example

- Simplifying gives us:

$$\underline{(}[x_1 \;\; x_2]^T)^T \begin{bmatrix} -3/4 & 0 \\ 0 & -3/4 \end{bmatrix} [x_1 \;\; x_2]^T - 2\ln(4/6) - \ln(1/16) = 0, \tag{50}$$

$$\underline{[} -3/4x_1 - 3/4x_2][x_1 x_2]^T + 549/677 + 2731/985 = 0, \tag{51}$$

$$-3/4x_1^2 - 3/4x_2^2 + 1609/449 = 0, \tag{52}$$

The final MAP decision boundary is:

$$x_1^2 + x_2^2 = 2131/446, \tag{53}$$

- 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!

$$(\underline{x}-\underline{\mu}_B)^T\Sigma_B^{-1}(\underline{x}-\underline{\mu}_B)-(\underline{x}-\underline{\mu}_A)^T\Sigma_A^{-1}(\underline{x}-\underline{\mu}_A) \underset{B}{\overset{A}{\gtrless}} 2\ln\left[\frac{P(B)}{P(A)}\right]+\ln\left[\frac{|\Sigma_A|}{|\Sigma_B|}\right] \tag{54}$$

$$d^2_{MICD}(\underline{x},\underline{\mu}_B,\Sigma_B) - d^2_{MICD}(\underline{x},\underline{\mu}_A,\Sigma_A) \underset{B}{\overset{A}{\gtrless}} 2\ln\left[\frac{P(B)}{P(A)}\right] + \ln\left[\frac{|\Sigma_A|}{|\Sigma_B|}\right] \tag{55}$$

- 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!

Relationship between MICD and MAP Classifiers for Normal Distributions

- Therefore, the MICD is only optimal in terms of probability of error only if we have multivariate Normal distributions $\mathcal{N}(\underline{\mu}, \Sigma)$ 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 ($|\Sigma|$)

Relationship between MICD and MAP Classifiers for Normal Distributions

- So under what circumstance does MAP classifier perform better than MICD?
- Recall the case where we have only one feature ($n = 1$), $\underline{m} = 0$, and $s_A \neq s_B$.
- The MICD classification rule for this case is:

$$(1/s_B^2 - 1/s_A^2)x^2 > 0 \tag{56}$$

$$(1/s_A^2)x^2 < (1/s_B^2)x^2 \tag{57}$$

$$s_A^2 > s_B^2 \tag{58}$$

- The MICD classification rule decides in favor of the class with the largest variance, regardless of x

Relationship between MICD and MAP Classifiers for Normal Distributions

- 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 $P(A) = P(B)$

$$(1/s_B^2 - 1/s_A^2)x^2 > \ln\left[\frac{s_A^2}{s_B^2}\right] \tag{60}$$

Relationship between MICD and MAP Classifiers for Normal Distributions

- Looking at the MAP classification rule:

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

- At the mean $m = 0$,

$$0 > \ln\left[\frac{s_A^2}{s_B^2}\right] \tag{62}$$

- 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.

Performance of the Bayes Classifier

- 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.

Probability of error given pattern

- For any pattern $\underline{x}$ such that $P(A|\underline{x}) > P(B|\underline{x})$:
  - $\underline{x}$ is classified as part of class A
  - The probability of error of classifying $\underline{x}$ as A is $P(B|\underline{x})$
- Therefore, naturally, for any given $\underline{x}$ the probability of error $P(\epsilon|\underline{x})$ is:

$$P(\epsilon|\underline{x}) = \min\left[P(A|\underline{x}), P(B|\underline{x})\right] \tag{63}$$

- **Rationale**: Since we always chose the maximum posterior probability as our class, the minimum posterior probability would be the probability of choosing incorrectly.

Probability of error given pattern

- Recall our previous example of a 1-D case:

$$p(x|A) = \mathcal{N}(\mu_A, \sigma_A^2) = \frac{1}{\sqrt{2\pi}\sigma_A}\exp\left[-\frac{1}{2}(\frac{x-\mu_A}{\sigma_A})^2\right] \quad (64)$$

$$p(x|B) = \mathcal{N}(\mu_B, \sigma_B^2) = \frac{1}{\sqrt{2\pi}\sigma_B}\exp\left[-\frac{1}{2}(\frac{x-\mu_B}{\sigma_B})^2\right] \quad (65)$$

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

- For $\underline{x} = 140$, what is the probability of error $P(\epsilon|\underline{x})$?

Probability of error given pattern

- Recall the MAP classifier for this scenario:

$$x \underset{A}{\overset{B}{\gtrless}} 131.9 \tag{66}$$

- Based on this MAP classifier, the pattern $\underline{x} = 140$ belongs to class A.
- Given the probability of error $P(\epsilon|\underline{x})$ is:

$$P(\epsilon|\underline{x}) = \min\left[P(A|\underline{x}), P(B|\underline{x})\right] \tag{67}$$

- Since B gives the maximum probability, the minimum probability would be $P(A|\underline{x})$.

Probability of error given pattern

- Therefore, $P(\epsilon|\underline{x})$ for $\underline{x} = 140$ is:

$$P(\epsilon|x)|_{x=140} = P(A|x)|_{x=140} = \frac{P(x|A)P(A)}{P(x|A)P(A) + P(x|B)P(B)}|_{x=140} \tag{68}$$

$$P(\epsilon|x)|_{x=140} = \frac{26/1477(0.4)}{(26/1477)0.4 + (26/1477)(0.6)} \tag{69}$$

$$P(\epsilon|x)|_{x=140} = 0.4. \tag{70}$$

Expected probability of error

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

$$P(\epsilon) = \int P(\epsilon|\underline{x})p(\underline{x})d\underline{x} \tag{71}$$

$$P(\epsilon) = \int \min\left[P(A|\underline{x}), P(B|\underline{x})\right] p(\underline{x})d\underline{x} \tag{72}$$

- In terms of class PDFs:

$$P(\epsilon) = \int \min\left[P(\underline{x}|A)P(A), P(\underline{x}|B)P(B)\right]d\underline{x} \tag{73}$$

Expected probability of error

- Now if we were to define decision regions R_A and R_B:
  - $R_A = \underline{x}$ such that $P(A|\underline{x}) > P(B|\underline{x})$
  - $R_B = \underline{x}$ such that $P(B|\underline{x}) > P(A|\underline{x})$
- The expected probability of error can be defined as:

$$P(\epsilon) = \int_{R_A} P(\underline{x}|B)P(B)d\underline{x} + \int_{R_B} P(\underline{x}|A)P(A)d\underline{x} \qquad (74)$$

- 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.

Expected probability of error

- 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:

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

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

- 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.

$$x = \frac{(\mu_B + \mu_A)}{2} \tag{77}$$

- 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 $x > \frac{(\mu_B + \mu_A)}{2}$

Expected probability of error

- Based on decision regions R_A, R_B, $P(A)$, $P(B)$, $P(\underline{x}|A)$, $P(\underline{x}|B)$, μ_B, μ_A, the expected probability of error $P(\epsilon)$ becomes

$$P(\epsilon) = \int_{R_A} P(B)P(\underline{x}|B)d\underline{x} + \int_{R_B} P(A)P(\underline{x}|A)d\underline{x} \tag{78}$$

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

$$P(\epsilon) = \frac{1}{2}\int_{-\infty}^{\frac{(\mu_B+\mu_A)}{2}} \mathcal{N}(\mu_B, \sigma^2)dx + \frac{1}{2}\int_{\frac{(\mu_B+\mu_A)}{2}}^{\infty} \mathcal{N}(\mu_A, \sigma^2)dx \tag{80}$$

Expected probability of error

- Since the two classes are symmetric ($P(\epsilon|A) = P(\epsilon|B)$),

$$P(\epsilon) = \frac{1}{2}\int_{-\infty}^{\frac{(\mu_B+\mu_A)}{2}} \mathcal{N}(\mu_B, \sigma^2)dx + \frac{1}{2}\int_{\frac{(\mu_B+\mu_A)}{2}}^{\infty} \mathcal{N}(\mu_A, \sigma^2)dx \tag{81}$$

$$P(\epsilon) = \int_{\frac{(\mu_B+\mu_A)}{2}}^{\infty} \mathcal{N}(\mu_A, \sigma^2)dx \tag{82}$$

$$P(\epsilon) = \int_{\frac{(\mu_B+\mu_A)}{2}}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_A}\exp\left[-\frac{1}{2}(\frac{x-\mu_A}{\sigma_A})^2\right]dx \tag{83}$$

Expected probability of error

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

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

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

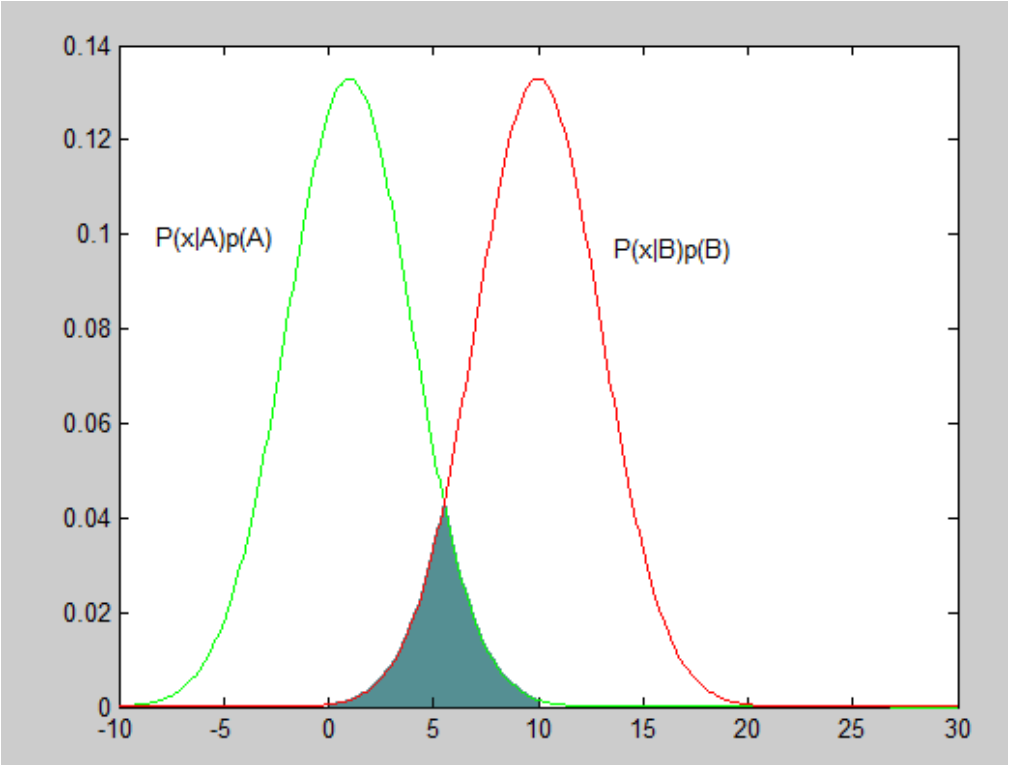
$$Q(\alpha) = \int_{\alpha}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left[-\frac{1}{2}y^2\right] dy \tag{85}$$

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

$$P(\epsilon) = Q(\frac{\mu_B - \mu_A}{2\sigma}) \tag{86}$$

Expected probability of error

- Visualization of $P(\epsilon)$:



$P(\epsilon)$ is essentially the shaded area.

Expected probability of error

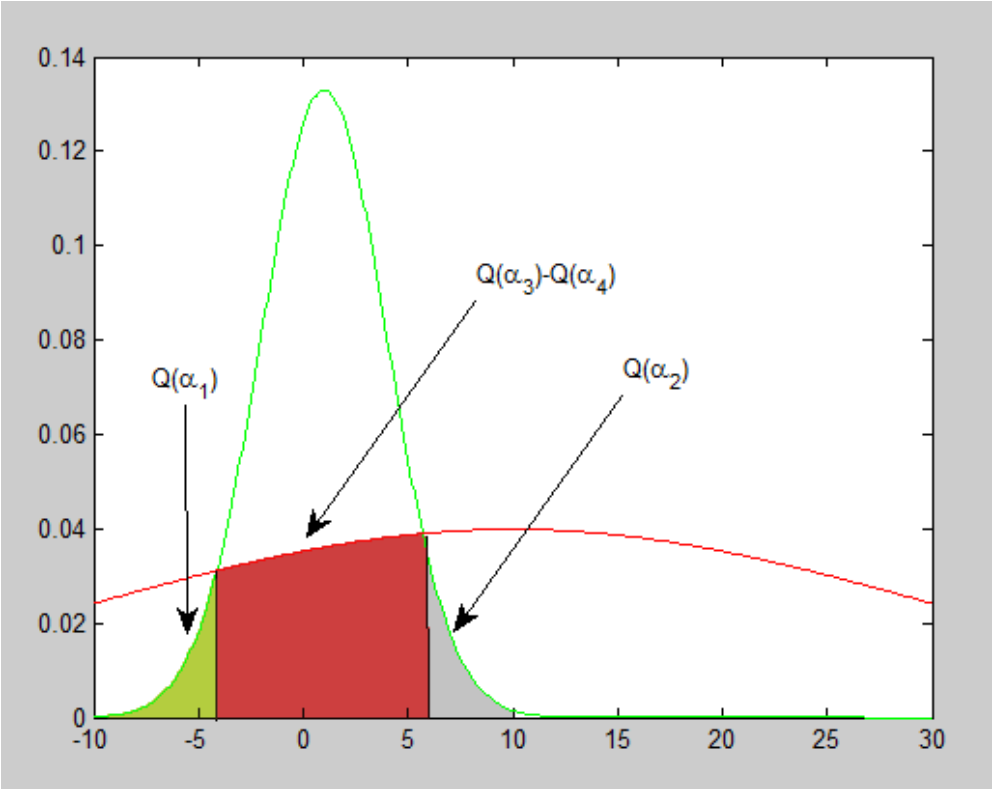
- Observations:
  - As the distance between the means increase, the shaded area becomes monotonically smaller and the expected probability of error $P(\epsilon)$ monotonically decreases.
  - At $\alpha = 0$, $\mu_A = \mu_B = 0$ and $P(\epsilon) = 1/2$ (makes sense since the distributions completely overlap, 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.

Expected probability of error

- Example:



$P(\epsilon)$ is essentially the shaded area.

$$P(\epsilon) = P(A)Q(\alpha_1) + P(B)[Q(\alpha_3) - Q(\alpha_4)] + P(A)Q(\alpha_2)$$

Expected probability of error

- Let's take a look at the multivariate case (n>1)
- For $p(\underline{x}|A) = \mathcal{N}(\underline{\mu}_A, \Sigma)$, $p(\underline{x}|B) = \mathcal{N}(\underline{\mu}_B, \Sigma)$, $P(A) = P(B)$, it can be shown that:

$$P(\epsilon) = Q(d_M(\underline{\mu}_A, \underline{\mu}_B)/2) \tag{87}$$

where $d_M(\underline{\mu}_A, \underline{\mu}_B)$ is the Mahalanobis distance between the classes.

$$d_M(\underline{\mu}_A, \underline{\mu}_B) = [(\underline{\mu}_A - \underline{\mu}_B)^T \Sigma^{-1} (\underline{\mu}_A - \underline{\mu}_B)]^{1/2} \tag{88}$$

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 Σ (since our orthonormal whitening transform is identical for both classes)
    - intersects with the midpoint of the line segment between $\underline{mu}_A$ and $\underline{mu}_B$
  - The probability of error is just the area under $P(\underline{x}|A)p(A)$ on the class B side of this decision boundary PLUS the area under $P(\underline{x}|B)p(B)$ on the class A side of this decision boundary.

Expected probability of error

- 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.

Expected probability of error

- Therefore, the expected probability of error is:

$$P(\epsilon) = \int \min\left[P(\underline{x}|C_1)P(C_1), P(\underline{x}|C_2)P(C_2)\right]d\underline{x} \tag{92}$$

$$P(\epsilon) = \int_{R_{C_1}} P(\underline{x}|C_2)P(C_2)d\underline{x} + \int_{R_{C_2}} P(\underline{x}|C_1)P(C_1)d\underline{x} \tag{93}$$

$$P(\epsilon) = \int_0^{0.5} 0.5P(\underline{x}|C_2)d\underline{x} + \int_{0.5}^{1.0} 0.5P(\underline{x}|C_1)d\underline{x} \tag{94}$$

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

$$P(\epsilon) = \int_{0.5}^{1.0} ce^{-\lambda x}d\underline{x} \tag{95}$$

$$P(\epsilon) = \frac{c}{\lambda}\left[e^{-\lambda/2} - e^{-\lambda}\right] \tag{96}$$

Expected probability of error

- (b) Find $P(\epsilon|x)$:
- From the decision boundary and decision regions we determined in (a),

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

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

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

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 **bounds** on $P(\epsilon)$, which are:
  - Easier to compute
  - Leads to estimates of classifier performance

Bhattacharrya bound

- Using the following inequality:

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

- The following holds true:

$$P(\epsilon) = \int \min\left[P(\underline{x}|A)P(A), P(\underline{x}|B)P(B)\right] d\underline{x} \tag{101}$$

$$P(\epsilon) \leq \sqrt{P(A)P(B)} \int \sqrt{P(\underline{x}|A)P(\underline{x}|B)} d\underline{x} \tag{102}$$

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

Bhattacharrya bound

- Since $P(A) + P(B) = 1$ and the Bhattacharrya coefficient ρ can be defined as:

$$\rho = \int \sqrt{P(\underline{x}|A)P(\underline{x}|B)}d\underline{x} \tag{103}$$

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

$$P(\epsilon) \leq \frac{1}{2}\rho \tag{104}$$

Bhattacharrya bound: Example

- 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

$$P(\epsilon) = 0.3 \leq \sqrt{P(A)P(B)} \int \sqrt{P(x|A)P(x|B)} d\underline{x} \quad (105)$$

$$\frac{0.3}{\sqrt{P(A)P(B)}} \leq \int \sqrt{P(x|A)P(x|B)} d\underline{x} \quad (106)$$

$$\rho = \int \sqrt{P(x|A)P(x|B)} d\underline{x} \geq \frac{0.3}{\sqrt{0.5 \times 0.5}} = 0.6 \quad (107)$$

Bhattacharrya bound Example

- Step 2: Based on Bhattacharrya coefficient ρ and new priors $P(A) = 0.2$ and $P(B) = 0.8$, the new upper bound can be computed as:

$$P(\epsilon) \leq \sqrt{P(A)P(B)} \int \sqrt{P(x|A)P(x|B)} d\underline{x} \tag{108}$$

$$P(\epsilon) \leq \sqrt{0.8 \times 0.2} \times \rho \tag{109}$$

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