L4: Bayesian Decision Theory

- Likelihood ratio test
- Probability of error
- Bayes risk
- Bayes, MAP and ML criteria
- Multi-class problems
- Discriminant functions

Likelihood ratio test (LRT)

Assume we are to classify an object based on the evidence provided by feature vector x

– Would the following decision rule be reasonable?
  - "**Choose the class that is most probable given observation x**"
  - More formally: Evaluate the posterior probability of each class $P(\omega_i|x)$ and choose the class with largest $P(\omega_i|x)$

Let's examine this rule for a 2-class problem

– In this case the decision rule becomes

$$\text{if } P(\omega_1|x) > P(\omega_2|x) \text{ choose } \omega_1 \text{ else choose } \omega_2$$

– Or, in a more compact form

$$P(\omega_1|x) \underset{\omega_2}{\overset{\omega_1}{\gtrless}} P(\omega_2|x)$$

– Applying Bayes rule

$$\frac{p(x|\omega_1)P(\omega_1)}{p(x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{p(x|\omega_2)P(\omega_2)}{p(x)}$$

– Since $p(x)$ does not affect the decision rule, it can be eliminated*
– Rearranging the previous expression

$$\Lambda(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)}$$

– The term $\Lambda(x)$ is called the likelihood ratio, and the decision rule is known as the **likelihood ratio test**

**$p(x)$ can be disregarded in the decision rule since it is constant regardless of class ω_i. However, $p(x)$ will be needed if we want to estimate the posterior $P(\omega_i|x)$ which, unlike $p(x|\omega_1)P(\omega_1)$, is a true probability value and, therefore, gives us an estimate of the "goodness" of our decision*

Likelihood ratio test: an example

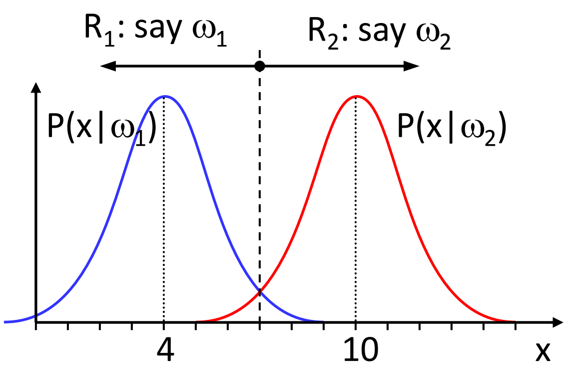
Problem

- Given the likelihoods below, derive a decision rule based on the LRT (assume equal priors)

$$p(x|\omega_1) = N(4,1); \qquad p(x|\omega_2) = N(10,1)$$

Solution

- Substituting into the LRT expression $\Lambda(x) = \dfrac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-4)^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(x-10)^2}} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \dfrac{1}{1}$
- Simplifying the LRT expression $\Lambda(x) = e^{-\frac{1}{2}(x-4)^2 + \frac{1}{2}(x-10)^2} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$
- Changing signs and taking logs $(x-4)^2 - (x-10)^2 \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 0$
- Which yields $x \underset{\omega_2}{\overset{\omega_1}{\lessgtr}} 7$
- This LRT result is intuitive since the likelihoods differ only in their mean
- How would the LRT decision rule change if the priors were such that $P(\omega_1) = 2P(\omega_2)$?



Probability of error

The performance of any decision rule can be measured by $P[error]$

- Making use of the Theorem of total probability (L2):

$$P[error] = \sum_{i=1}^{C} P[error|\omega_i] P[\omega_i]$$

- The class conditional probability $P[error|\omega_i]$ can be expressed as

$$P[error|\omega_i] = P\left[choose\ \omega_j \middle| \omega_i\right] = \int_{R_j} p(x|\omega_i)dx = \epsilon_i$$

- So, for our 2-class problem, $P[error]$ becomes

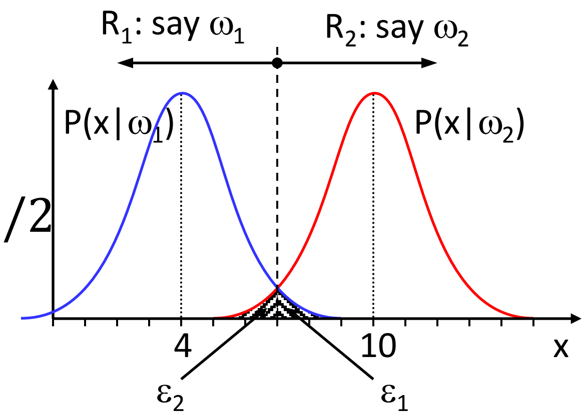
$$P[error] = P[\omega_1] \underbrace{\int_{R_2} p(x|\omega_1)dx}_{\epsilon_1} + P[\omega_2] \underbrace{\int_{R_1} p(x|\omega_2)dx}_{\epsilon_2}$$

  - where ϵ_i is the integral of $p(x|\omega_i)$ over region R_j where we choose ω_j

- For the previous example, since we assumed equal priors, then

$$P[error] = (\epsilon_1 + \epsilon_2)/2$$

- How would you compute $P[error]$ numerically?



How good is the LRT decision rule?

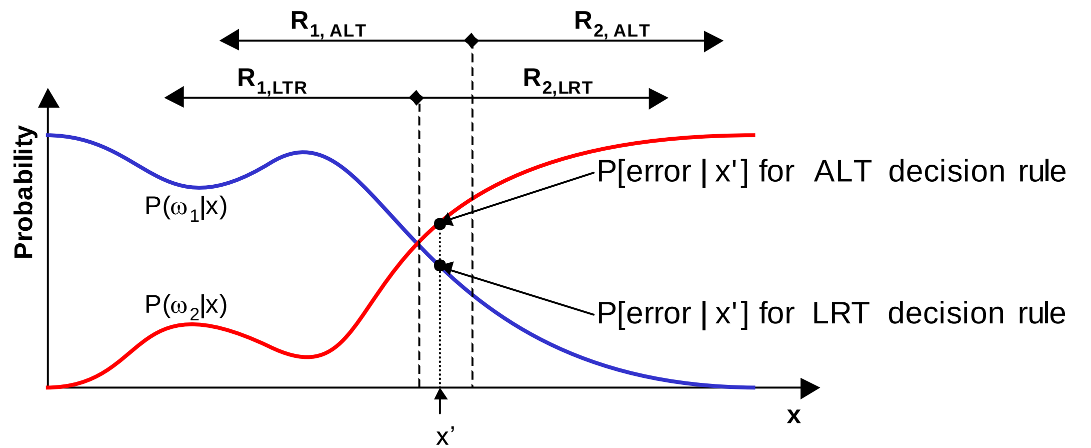
- To answer this question, it is convenient to express $P[error]$ in terms of the posterior $P[error|x]$

$$P[error] = \int_{-\infty}^{\infty} P[error|x]p(x)dx$$

- The optimal decision rule will minimize $P[error|x]$ at every value of x in feature space, so that the integral above is minimized

– At each x', $P[error|x']$ is equal to $P[\omega_i|x']$ when we choose ω_j

- This is illustrated in the figure below



– From the figure it becomes clear that, for any value of x', the LRT will always have a lower $P[error|x']$

- Therefore, when we integrate over the real line, the LRT decision rule will yield a lower $P[error]$

For any given problem, the minimum probability of error is achieved by the LRT decision rule; this probability of error is called the **Bayes Error Rate** and is the **best** any classifier can do.

Bayes risk

So far we have assumed that the penalty of misclassifying $\mathbf{x} \in \boldsymbol{\omega}_1$ as $\boldsymbol{\omega}_2$ is the same as the reciprocal error

- In general, this is not the case
- For example, misclassifying a cancer sufferer as a healthy patient is a much more serious problem than the other way around
- This concept can be formalized in terms of a cost function C_{ij}
  - C_{ij} represents the cost of choosing class ω_i when ω_j is the true class

We define the Bayes Risk as the expected value of the cost

$$\Re = E[C] = \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} P\left[choose\ \omega_i and\ x \in \omega_j\right] =$$
$$= \sum_{i=1}^{2}\sum_{j=1}^{2} C_{ij} P\left[x \in R_i|\omega_j\right] P\left[\omega_j\right]$$

What is the decision rule that minimizes the Bayes Risk?

- First notice that

$$P\left[x \in \mathrm{R}_i \,\middle|\, \omega_j\right] = \int_{R_i} p\left(x \middle| \omega_j\right) dx$$

- We can express the Bayes Risk as

$$\mathfrak{R} = \int_{R_1} [C_{11} P[\omega_1] p(x|\omega_1) + C_{12} P[\omega_2] p(x|\omega_2] dx + \int_{R_2} [C_{21} P[\omega_1] p(x|\omega_1) + C_{22} P[\omega_2] p(x|\omega_2] dx$$

- Then we note that, for either likelihood, one can write:

$$\int_{R_1} p(x|\omega_i) dx + \int_{R_2} p(x|\omega_i) dx = \int_{R_1 \cup R_2} p(x|\omega_i) dx = 1$$

– Merging the last equation into the Bayes Risk expression yields

$$\begin{aligned}\Re = & \, C_{11}P_1\int_{R_1} p(x|\omega_1)dx + C_{12}P_2\int_{R_1} p(x|\omega_2)dx \\ & + C_{21}P_1\int_{R_2} p(x|\omega_1)dx + C_{22}P_2\int_{R_2} p(x|\omega_2)dx \\ & + C_{21}P_1\int_{R_1} p(x|\omega_1)dx + C_{22}P_2\int_{R_1} p(x|\omega_2)dx \\ & - C_{21}P_1\int_{R_1} p(x|\omega_1)dx - C_{22}P_2\int_{R_1} p(x|\omega_2)dx\end{aligned}$$

– Now we cancel out all the integrals over R_2

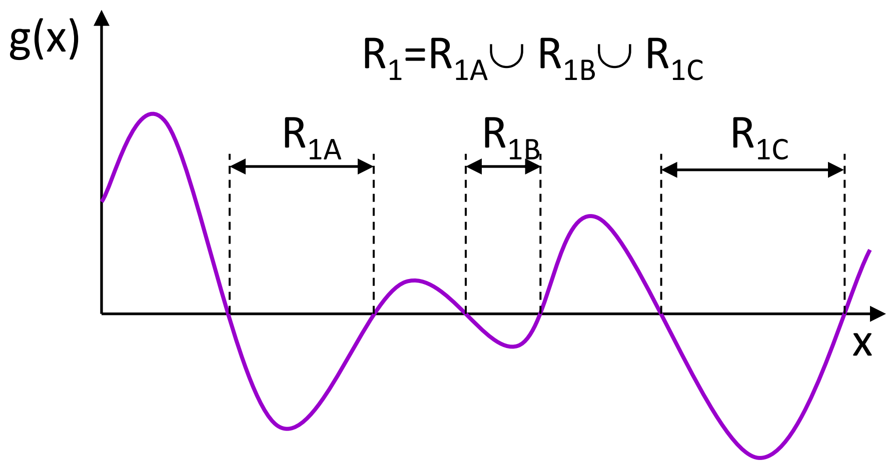
$$\Re = C_{21}P_1 + C_{22}P_2 + \underbrace{(C_{12} - C_{22})}_{>0}P_2\int_{R_1} p(x|\omega_2)dx - \underbrace{(C_{21} - C_{11})}_{>0}P_1\int_{R_1} p(x|\omega_1)dx$$

– The first two terms are constant w.r.t. R_1 so they can be ignored

– Thus, we seek a decision region R_1 that minimizes

$$\begin{aligned}R_1 &= argmin\int_{R_1} [(C_{12} - C_{22})P_2 p(x|\omega_2) - (C_{21} - C_{11})P_1 p(x|\omega_1)]dx \\ &= argmin\int_{R_1} g(x)\end{aligned}$$

- Let's forget about the actual expression of $g(x)$ to develop some intuition for what kind of decision region R_1 we are looking for
  - Intuitively, we will select for R_1 those regions that minimize $\int_{R_1} g(x)$
  - In other words, those regions where $g(x) < 0$



- So we will choose R_1 such that

$$(C_{21} - C_{11})P_1 p(x|\omega_1) > (C_{12} - C_{22})P_2 p(x|\omega_2)$$

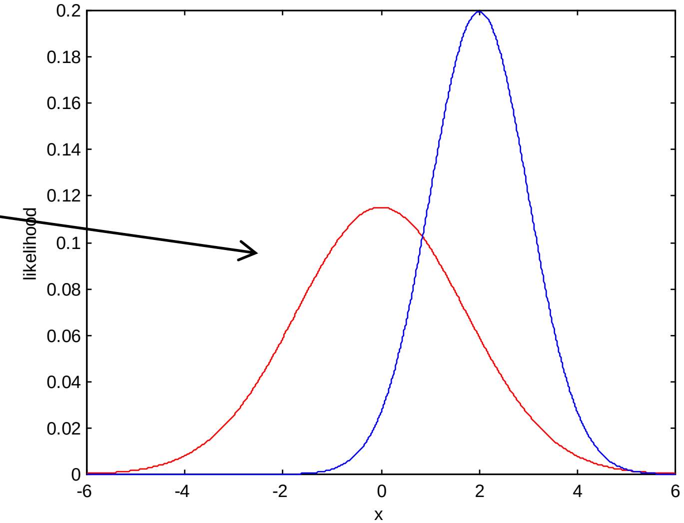
- And rearranging

$$\frac{P(x|\omega_1)}{P(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{(C_{12} - C_{22})P(\omega_2)}{(C_{21} - C_{11})\, P(\omega_1)}$$

- **Therefore, minimization of the Bayes Risk also leads to an LRT**

The Bayes risk: an example

- Consider a problem with likelihoods $L_1 = N\left(0, \sqrt{3}\right)$ and $L_2 = N(2,1)$
  - Sketch the two densities
  - What is the likelihood ratio?
  - Assume $P_1 = P_2$, $C_{ii} = 0$, $C_{12} = 1$ and $C_{21} = 3^{1/2}$
  - Determine a decision rule to minimize $P[error]$

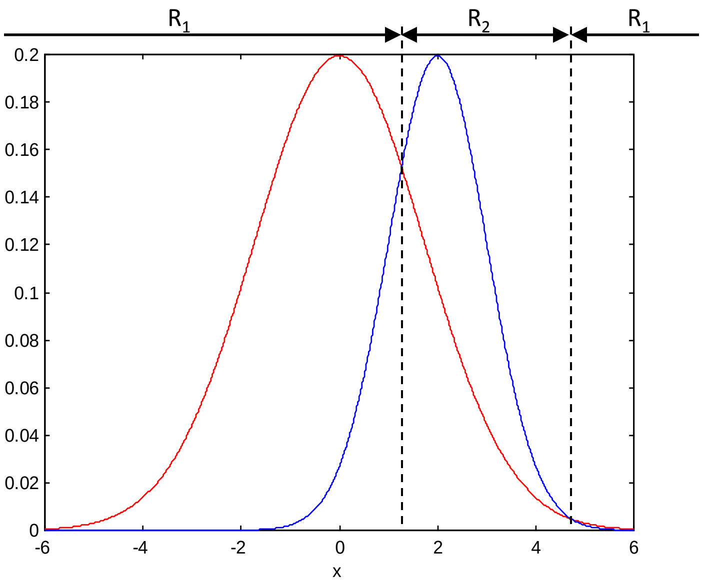


$$\Lambda(x) = \frac{N\left(0,\sqrt{3}\right)}{N(2,1\,)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{1}{\sqrt{3}} \Rightarrow$$

$$\Rightarrow -\frac{1}{2}\frac{x^2}{3} + \frac{1}{2}(x-2)^2 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0 \Rightarrow$$

$$\Rightarrow 2x^2 - 12x + 12 \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 0 \Rightarrow$$

$$\Rightarrow x = 4.73, 1.27$$



LRT variations

Bayes criterion

– This is the LRT that minimizes the Bayes risk

$$\Lambda_{\text{Bayes}}(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{(C_{12} - C_{22})P(\omega_2)}{(C_{21} - C_{11})\,P(\omega_1)}$$

Maximum A Posteriori criterion

– Sometimes we may be interested in minimizing $P[error]$

– A special case of $\Lambda_{\text{Bayes}}(x)$ that uses a zero-one cost $\mathrm{C_{ij}} = \begin{cases} 0; i = j \\ 1; i \neq j \end{cases}$

– Known as the MAP criterion, since it seeks to maximize $P(\omega_i|x)$

$$\Lambda_{\text{MAP}}(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} \frac{P(\omega_2)}{P(\omega_1)} \Rightarrow \frac{P(\omega_1|x)}{P(\omega_2|x)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

Maximum Likelihood criterion

– For equal priors $P[\omega_i] = 1/2$ and 0/1 loss function, the LTR is known as a ML criterion, since it seeks to maximize $P(x|\omega_i)$

$$\Lambda_{\text{ML}}(x) = \frac{p(x|\omega_1)}{p(x|\omega_2)} \underset{\omega_2}{\overset{\omega_1}{\gtrless}} 1$$

Two more decision rules are commonly cited in the literature

- The **Neyman-Pearson Criterion**, used in Detection and Estimation Theory, which also leads to an LRT, fixes one class error probabilities, say $\epsilon_1 < \alpha$, and seeks to minimize the other
  - For instance, for the sea-bass/salmon classification problem of L1, there may be some kind of government regulation that we must not misclassify more than 1% of salmon as sea bass
  - The Neyman-Pearson Criterion is very attractive since it does not require knowledge of priors and cost function
- The **Minimax Criterion**, used in Game Theory, is derived from the Bayes criterion, and seeks to <u>mini</u>mize the <u>max</u>imum Bayes Risk
  - The Minimax Criterion does nor require knowledge of the priors, but it needs a cost function
- For more information on these methods, refer to "*Detection, Estimation and Modulation Theory*", by H.L. van Trees

Minimum $P[error]$ for multi-class problems

Minimizing $P[error]$ generalizes well for multiple classes

- For clarity in the derivation, we express $P[error]$ in terms of the probability of making a correct assignment

$$P[error] = 1 - P[correct]$$

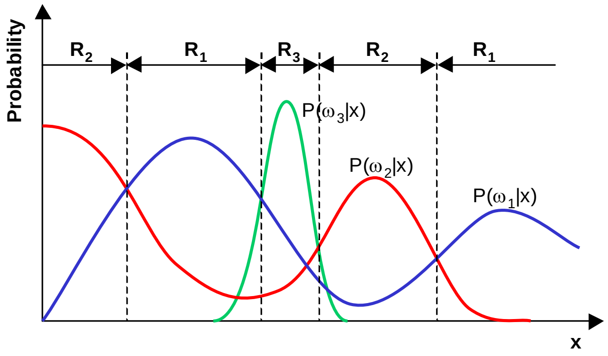
  - The probability of making a correct assignment is

$$P[correct] = \Sigma_{i=1}^{C} P[\omega_i] \int_{R_i} p(x|\omega_i)dx$$

  - Minimizing $P[error]$ is equivalent to maximizing $P[correct]$, so expressing the latter in terms of posteriors

$$P[correct] = \Sigma_{i=1}^{C} \int_{R_i} p(x)P(\omega_i|x)dx$$

  - To maximize $P[correct]$, we must maximize each integral $\int_{R_i}$, which we achieve by choosing the class with largest posterior
  - So each R_i is the region where $P(\omega_i|x)$ is maximum, and <u>the decision rule that minimizes P[error] is the MAP criterion</u>



Minimum Bayes risk for multi-class problems

Minimizing the Bayes risk also generalizes well

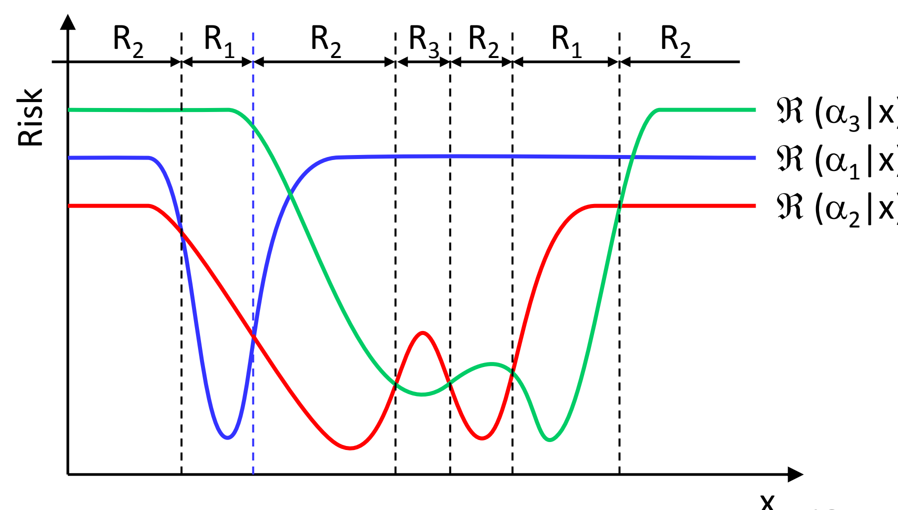
- As before, we use a slightly different formulation
  - We denote by α_i the decision to choose class ω_i
  - We denote by $\alpha(x)$ the overall decision rule that maps feature vectors x into classes ω_i, $\alpha(x) \to \{\alpha_1, \alpha_2, \dots \alpha_C\}$
- The (conditional) risk $\Re(\alpha_i|x)$ of assigning x to class ω_i is

$$\Re(\alpha(x) \to \alpha_i) = \Re(\alpha_i|x) = \sum_{j=1}^{C} C_{ij} P(\omega_j|x)$$

- And the Bayes Risk associated with decision rule $\alpha(x)$ is

$$\Re(\alpha(x)) = \int \Re(\alpha(x)|x) p(x) dx$$

- To minimize this expression, we must minimize the conditional risk $\Re(\alpha(x)|x)$ at each x, which is equivalent to choosing ω_i such that $\Re(\alpha_i|x)$ is minimum



Discriminant functions

All the decision rules shown in L4 have the same structure

- At each point x in feature space, choose class ω_i that maximizes (or minimizes) some measure $g_i(x)$
- This structure can be formalized with a set of discriminant functions $g_i(x), i = 1..C$, and the decision rule

"assign x to class ω_i if $g_i(x) > g_j(x) \ \ \forall j \neq i$"

- Therefore, we can visualize the decision rule as a network that computes C df's and selects the class with highest discriminant
- And the three decision rules can be summarized as

Criterion	Discriminant Function
Bayes	$g_i(x) = -\Re(\alpha_i\|x)$
MAP	$g_i(x) = P(\omega_i\|x)$
ML	$g_i(x) = P(x\|\omega_i)$