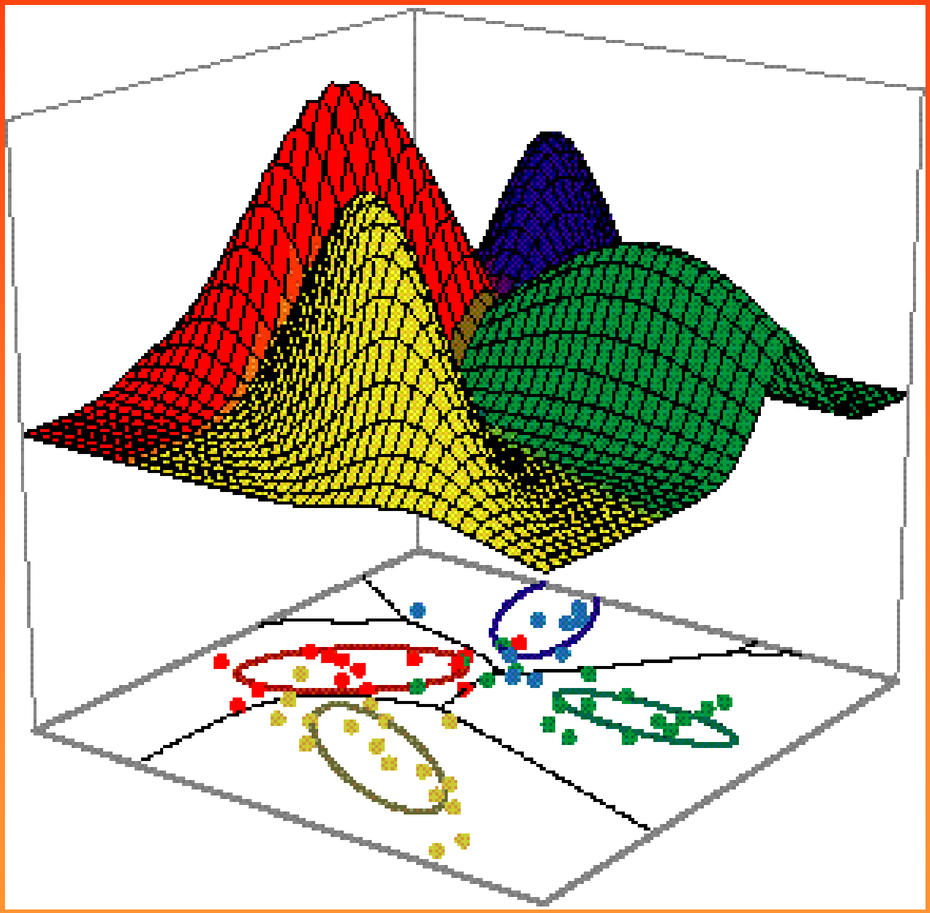


Pattern Classification

Most of the material in these slides was taken from the figures in *Pattern Classification (2nd ed) by R. O. Duda, P. E. Hart and D. G. Stork, John Wiley & Sons, 2001*

Recall the Fish!

- Recall our example from the first lecture on classifying two fish as salmon or sea bass.
- And recall our agreement that any given fish is either a salmon or a sea bass; DHS call this the **state of nature** of the fish.
- Let's define a (probabilistic) variable ω that describes the state of nature.

$$\omega = \omega_1 \quad \text{for sea bass} \tag{1}$$
$$\omega = \omega_2 \quad \text{for salmon} \tag{2}$$

- Let's assume this two class case.

Salmon

Sea Bass

Prior Probability

- The *a priori* or **prior** probability reflects our knowledge of how likely we expect a certain state of nature before we can actually observe said state of nature.
- In the fish example, it is the probability that we will see either a salmon or a sea bass next on the conveyor belt.
- Note: The prior may vary depending on the situation.
  - If we get equal numbers of salmon and sea bass in a catch, then the priors are equal, or **uniform**.
  - Depending on the season, we may get more salmon than sea bass, for example.
- We write $P(\omega = \omega_1)$ or just $P(\omega_1)$ for the prior the next is a sea bass.
- The priors must exhibit exclusivity and exhaustivity. For c states of nature, or classes:

$$1 = \sum_{i=1}^{c} P(\omega_i) \tag{3}$$

Decision Rule From Only Priors

- A **decision rule** prescribes what action to take based on observed input.
- IDEA CHECK: What is a reasonable Decision Rule if
  - the only available information is the prior, and
  - the cost of any incorrect classification is equal?
- Decide ω_1 if $P(\omega_1) > P(\omega_2)$; otherwise decide ω_2.
- What can we say about this decision rule?
  - Seems reasonable, but it will **always** choose the same fish.
  - If the priors are uniform, this rule will behave poorly.
  - Under the given assumptions, no other rule can do better! (We will see this later on.)

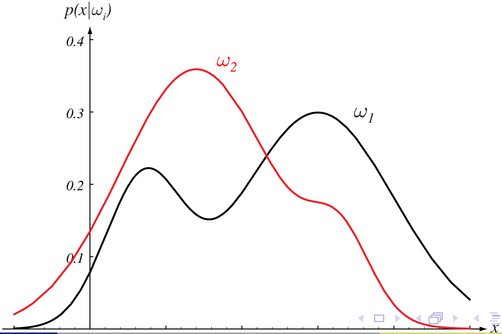
Class-Conditional Density

or Likelihood

- The **class-conditional probability density** function is the probability density function for $\mathbf{x}$, our feature, given that the state of nature is ω:

$$p(\mathbf{x}|\omega) \tag{4}$$

- Here is the hypothetical class-conditional density $p(x|\omega)$ for lightness values of sea bass and salmon.



Posterior Probability

Bayes Formula

- If we know the prior distribution and the class-conditional density, how does this affect our decision rule?
- **Posterior probability** is the probability of a certain state of nature given our observables: $P(\omega|\mathbf{x})$.
- Use Bayes Formula:

$$P(\omega, \mathbf{x}) = P(\omega|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\omega)P(\omega) \tag{5}$$

$$\begin{aligned} P(\omega|\mathbf{x}) &= \frac{p(\mathbf{x}|\omega)P(\omega)}{p(\mathbf{x})} && (6) \\ &= \frac{p(\mathbf{x}|\omega)P(\omega)}{\sum_i p(\mathbf{x}|\omega_i)P(\omega_i)} && (7) \end{aligned}$$

Probability of Error

- For a given observation x, we would be inclined to let the posterior govern our decision:

$$\omega^* = \arg\max_i P(\omega_i|\mathbf{x}) \tag{8}$$

- What is our **probability of error**?
- For the two class situation, we have

$$P(\text{error}|\mathbf{x}) = \begin{cases} P(\omega_1|\mathbf{x}) & \text{if we decide } \omega_2 \\ P(\omega_2|\mathbf{x}) & \text{if we decide } \omega_1 \end{cases} \tag{9}$$

Probability of Error

- We can minimize the probability of error by following the posterior:

$$\text{Decide } \omega_1 \text{ if } P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x}) \tag{10}$$

Loss Functions

- A **loss function** states exactly how costly each action is.
- As earlier, we have c classes $\{\omega_1, \dots, \omega_c\}$.
- We also have a possible actions $\{\alpha_1, \dots, \alpha_a\}$.
- The loss function $\lambda(\alpha_i|\omega_j)$ is the loss incurred for taking action α_i when the class is ω_j.
- The **Zero-One Loss Function** is a particularly common one:

$$\lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \qquad i, j = 1, 2, \dots, c \tag{13}$$

It assigns no loss to a correct decision and uniform unit loss to an incorrect decision.

Expected Loss
a.k.a. Conditional Risk

- We can consider the loss that would be incurred from taking each possible action in our set.
- The **expected loss** or conditional risk is by definition

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \quad (14)$$

- The **zero-one conditional risk** is

$$R(\alpha_i|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) \quad (15)$$
$$= 1 - P(\omega_i|\mathbf{x}) \quad (16)$$

- Hence, for an observation x, we can minimize the expected loss by selecting the action that minimizes the conditional risk.
- (Teaser) You guessed it: this is what Bayes Decision Rule does!

Bayes Risk

The Minimum Overall Risk

- Bayes Decision Rule gives us a method for minimizing the overall risk.
- Select the action that minimizes the conditional risk:

$$\alpha* = \arg\min_{\alpha_i} R\left(\alpha_i|\mathbf{x}\right) \tag{18}$$

$$= \arg\min_{\alpha_i} \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \tag{19}$$

- The Bayes Risk is the best we can do.

Two-Category Classification Examples

- Consider two classes and two actions, α_1 when the true class is ω_1 and α_2 for ω_2.
-
- Fundamental rule is decide ω_1 if

$$R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x}) .$$

-

Pattern Classifiers Version 1: Discriminant Functions

- **Discriminant Functions** are a useful way of representing pattern classifiers.
- Let's say $g_i(\mathbf{x})$ is a discriminant function for the ith class.
- This classifier will assign a class ω_i to the feature vector $\mathbf{x}$ if

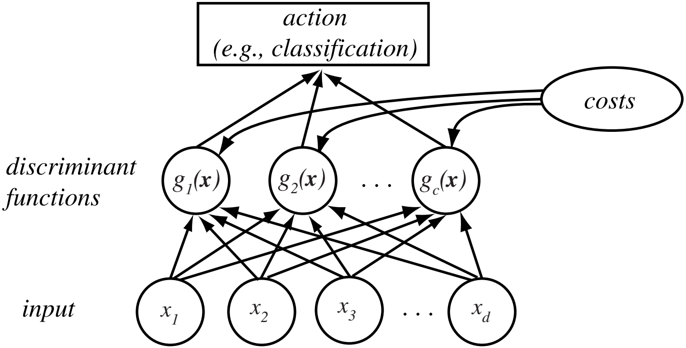
$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \qquad \forall j \neq i \ , \tag{26}$$

or, equivalently

$$i^* = \arg\max_i g_i(x) \ , \quad \text{decide} \quad \omega_{i^*} \ .$$

Discriminants as a Network

- We can view the discriminant classifier as a network (for c classes and a d-dimensional input vector).



Bayes Discriminants

Minimum Conditional Risk Discriminant

- General case with risks

$$g_i(\mathbf{x}) = -R(\alpha_i|\mathbf{x}) \tag{27}$$

$$= -\sum_{j=1}^{c} \lambda(\alpha_i|\omega_j)P(\omega_j|\mathbf{x}) \tag{28}$$

- Can we prove that this is correct?
- **Yes!** The minimum conditional risk corresponds to the maximum discriminant.

Minimum Error-Rate Discriminant

- In the case of zero-one loss function, the Bayes Discriminant can be further simplified:

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) \quad . \tag{29}$$

Uniqueness Of Discriminants

- Is the choice of discriminant functions unique?
- **No!**
- Multiply by some positive constant.
- Shift them by some additive constant.
- For monotonically increasing function $f(\cdot)$, we can replace each $g_i(\mathbf{x})$ by $f(g_i(\mathbf{x}))$ without affecting our classification accuracy.
  - These can help for ease of understanding or computability.
  - The following all yield the same exact classification results for minimum-error-rate classification.

$$g_i(\mathbf{x}) = P(\omega_i|\mathbf{x}) = \frac{p(\mathbf{x}|\omega_i)P(\omega_i)}{\sum_j p(\mathbf{x}|\omega_j)P(\omega_j)} \tag{30}$$

$$g_i(\mathbf{x}) = p(\mathbf{x}|\omega_i)P(\omega_i) \tag{31}$$

$$g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i) \tag{32}$$

Visualizing Discriminants

Decision Regions

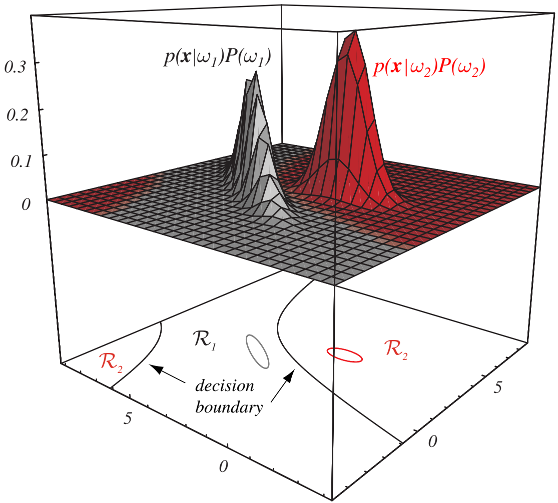
- The effect of any decision rule is to divide the feature space into decision regions.
- Denote a decision region $\mathcal{R}_i$ for ω_i.
- One not necessarily connected region is created for each category and assignments is according to:

$$\text{If } g_i(\mathbf{x}) > g_j(\mathbf{x}) \; \forall j \neq i, \text{ then } \mathbf{x} \text{ is in } \mathcal{R}_i \quad . \tag{33}$$

- **Decision boundaries** separate the regions; they are ties among the discriminant functions.

Visualizing Discriminants

Decision Regions



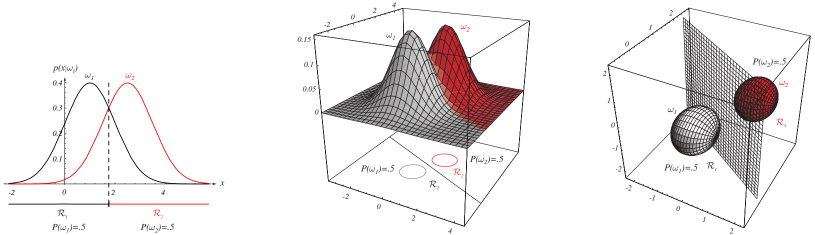
General Discriminant for Normal Densities

- Recall the minimum error rate discriminant, $g_i(\mathbf{x}) = \ln p(\mathbf{x}|\omega_i) + \ln P(\omega_i)$.
- If we assume normal densities, i.e., if $p(\mathbf{x}|\omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, then the general discriminant is of the form

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^{\mathsf{T}} \boldsymbol{\Sigma}_i^{-1} (\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln|\boldsymbol{\Sigma}_i| + \ln P(\omega_i) \tag{50}$$

Simple Case: Statistically Independent Features with Same Variance

- What do the decision boundaries look like if we assume $\mathbf{\Sigma}_i = \sigma^2 \mathbf{I}$?
- They are hyperplanes.



-

Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

- But, we don't need to actually compute the distances.
- Expanding the quadratic form $(\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}}(\mathbf{x} - \boldsymbol{\mu})$ yields

$$g_i(\mathbf{x}) = -\frac{1}{2\sigma^2}\left[\mathbf{x}^{\mathsf{T}}\mathbf{x} - 2\boldsymbol{\mu}_i^{\mathsf{T}}\mathbf{x} + \boldsymbol{\mu}_i^{\mathsf{T}}\boldsymbol{\mu}_i\right] + \ln P(\omega_i) \ . \tag{52}$$

- The quadratic term $\mathbf{x}^{\mathsf{T}}\mathbf{x}$ is the same for all i and can thus be ignored.
- This yields the equivalent **linear discriminant functions**

$$g_i(\mathbf{x}) = \mathbf{w}_i^{\mathsf{T}}\mathbf{x} + w_{i0} \tag{53}$$

$$\mathbf{w}_i = \frac{1}{\sigma^2}\boldsymbol{\mu}_i \tag{54}$$

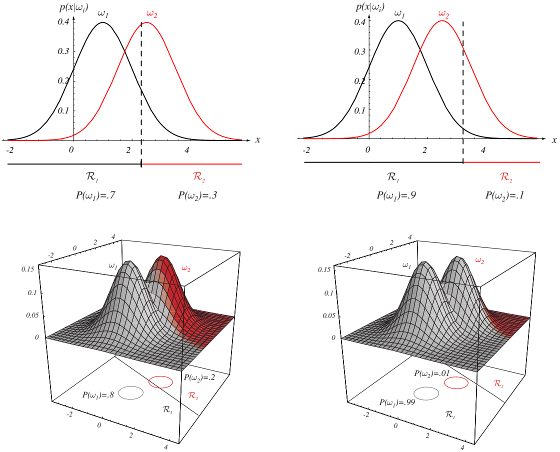
$$w_{i0} = -\frac{1}{2\sigma^2}\boldsymbol{\mu}_i^{\mathsf{T}}\boldsymbol{\mu}_i + \ln P(\omega_i) \tag{55}$$

- w_{i0} is called the **bias**.

Simple Case: $\Sigma_i = \sigma^2 \mathbf{I}$

Decision Boundary Equation

- The decision boundary changes with the prior.



General Case: Arbitrary Σ_i

- The discriminant functions are quadratic (the only term we can drop is the $\ln 2\pi$ term):

$$g_i(\mathbf{x}) = \mathbf{x}^{\mathsf{T}} \mathbf{W}_i \mathbf{x} + \mathbf{w}_i^{\mathsf{T}} \mathbf{x} + w_{i0} \tag{59}$$

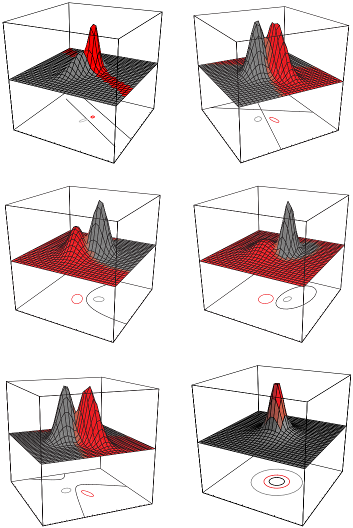
$$\mathbf{W}_i = -\frac{1}{2} \mathbf{\Sigma}_i^{-1} \tag{60}$$

$$\mathbf{w}_i = \mathbf{\Sigma}_i^{-1} \boldsymbol{\mu}_i \tag{61}$$

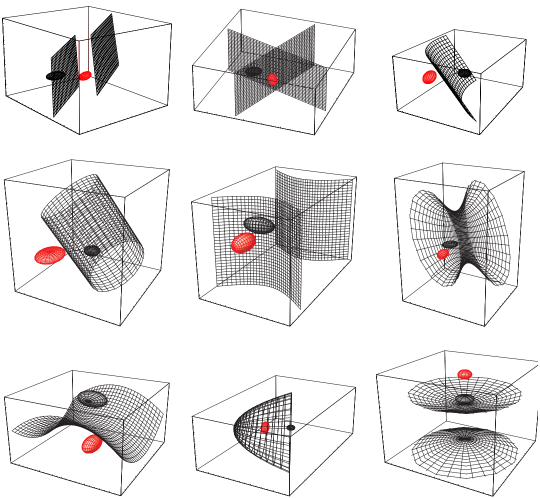
$$w_{i0} = -\frac{1}{2} \boldsymbol{\mu}_i^{\mathsf{T}} \mathbf{\Sigma}_i^{-1} \boldsymbol{\mu}_i - \frac{1}{2} \ln |\mathbf{\Sigma}_i| + \ln P(\omega_i) \tag{62}$$

- The decision surface between two categories are **hyperquadrics**.

General Case: Arbitrary Σ_i



General Case: Arbitrary Σ_i



General Case for Multiple Categories