# Bayesian Decision Theory

Jason Corso

SUNY at Buffalo

# Overview and Plan

- Covering Chapter 2 of DHS.
- Bayesian Decision Theory is a fundamental statistical approach to the problem of pattern classification.
- Quantifies the tradeoffs between various classifications using probability and the costs that accompany such classifications.
- Assumptions:
  - Decision problem is posed in probabilistic terms.
  - All relevant probability values are known.

# 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 $\omega$ 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.

# 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.

# 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?

# 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 $\omega_1$ if $P(\omega_1) > P(\omega_2)$; otherwise decide $\omega_2$.
- What can we say about this decision rule?

# 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 $\omega_1$ if $P(\omega_1) > P(\omega_2)$; otherwise decide $\omega_2$.
- 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.)

# Features and Feature Spaces

- A **feature** is an observable variable.
- A **feature space** is a set from which we can sample or observe values.
- Examples of features:
  - Length
  - Width
  - Lightness
  - Location of Dorsal Fin
- For simplicity, let's assume that our features are all continuous values.
- Denote a scalar feature as $x$ and a vector feature as $\mathbf{x}$. For a $d$-dimensional feature space, $\mathbf{x} \in \mathbb{R}^d$.

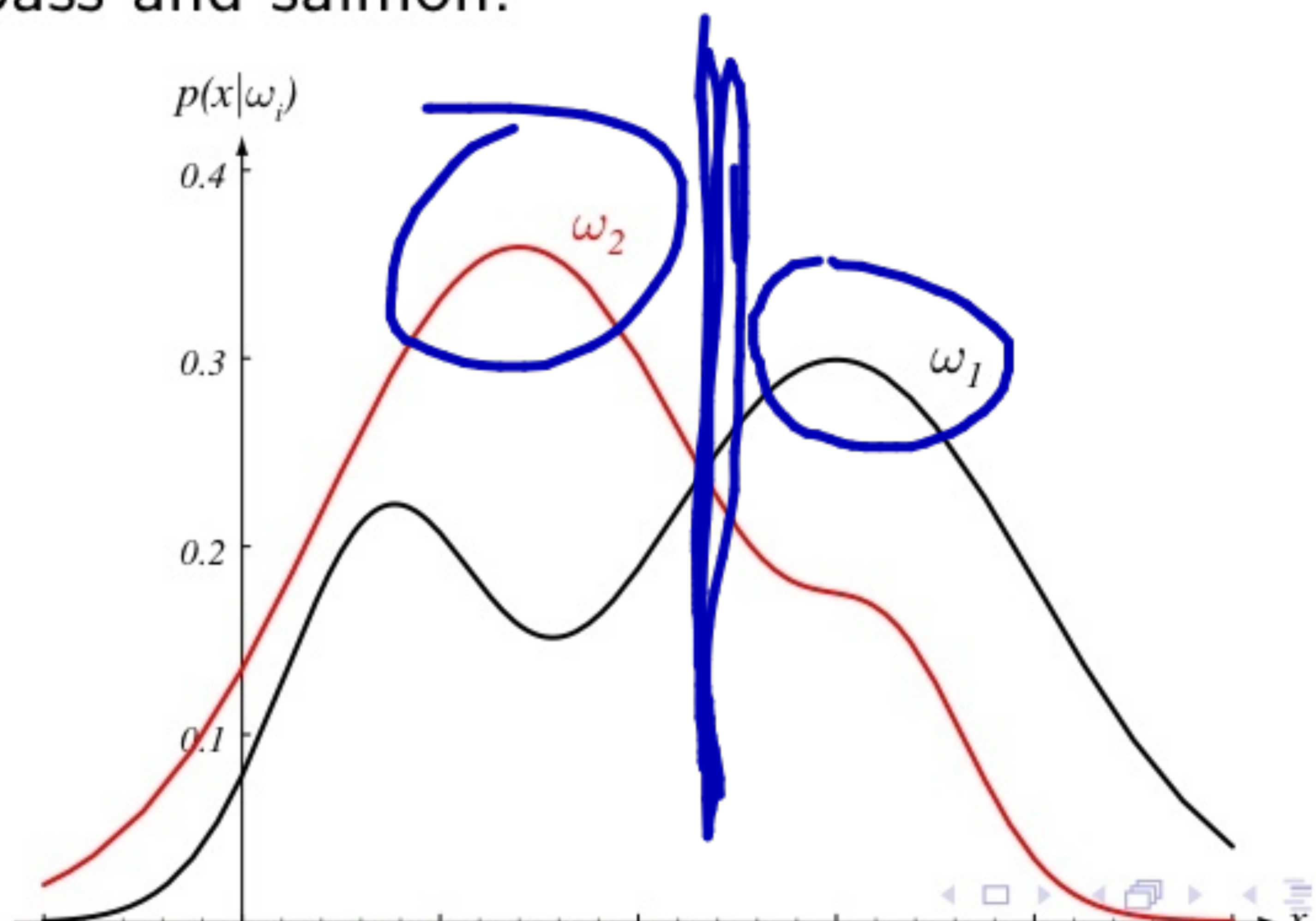
# Class-Conditional Density

## 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 $\omega$:

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

- 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:

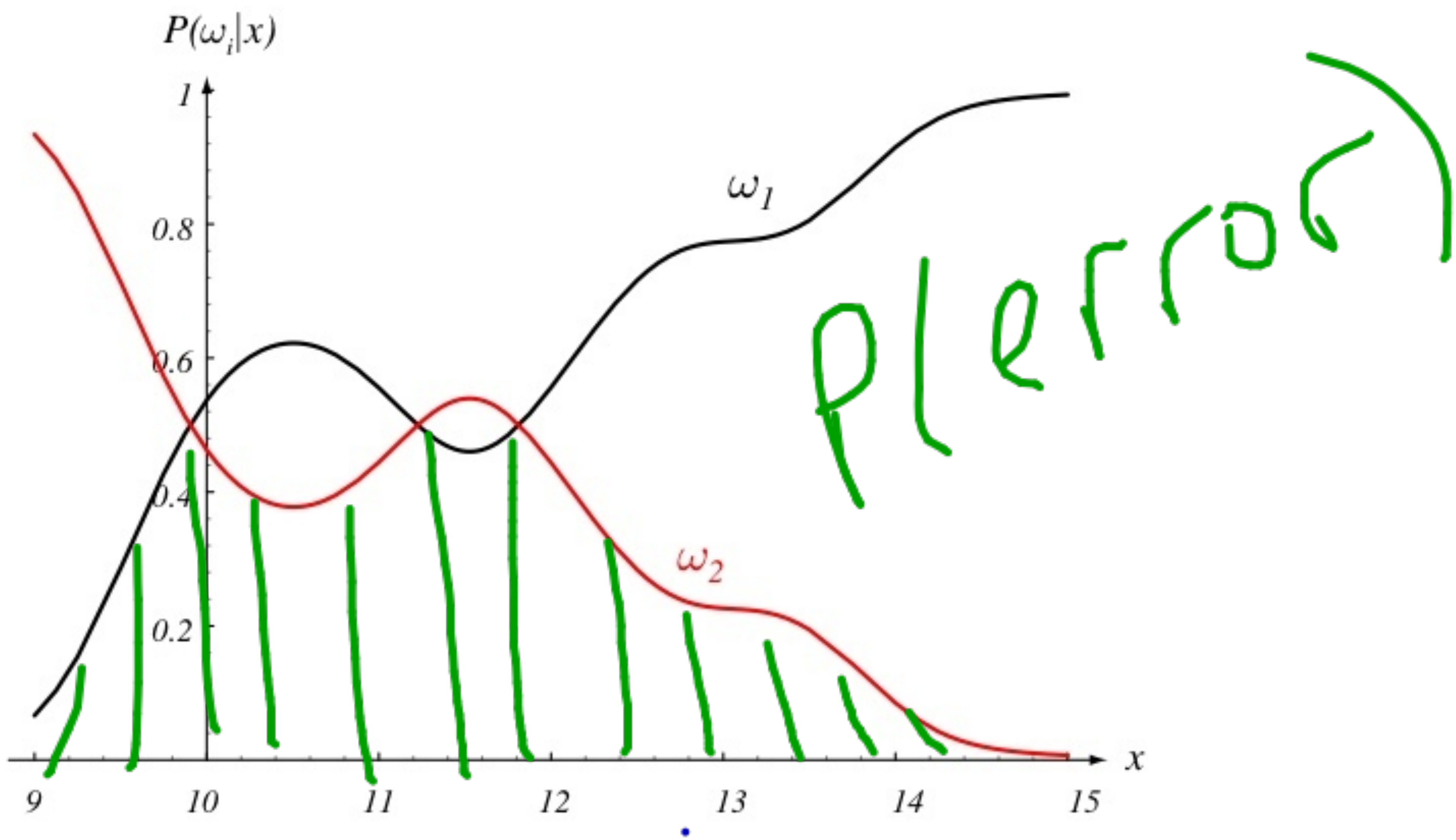
$$P(\omega, \mathbf{x}) = P(\omega|\mathbf{x})p(\mathbf{x}) = p(\mathbf{x}|\omega)P(\omega) \tag{5}$$

$$P(\omega|\mathbf{x}) = \frac{p(\mathbf{x}|\omega)P(\omega)}{p(\mathbf{x})} \tag{6}$$

$$= \frac{p(\mathbf{x}|\omega)P(\omega)}{\sum_i p(\mathbf{x}|\omega_i)P(\omega_i)} \tag{7}$$

# Posterior Probability

- Notice the likelihood and the prior govern the posterior. The $p(x)$ evidence term is a scale-factor to normalize the density.
- For the case of $P(\omega_1) = 2/3$ and $P(\omega_2) = 1/3$ the posterior is



# 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**?

# 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:

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

# Probability of Error

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

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

- And, this minimizes the average probability of error too:

$$P(\text{error}) = \int_{-\infty}^{\infty} P(\text{error}|\mathbf{x})p(\mathbf{x})d\mathbf{x} \tag{11}$$

(Because the integral will be minimized when we can ensure each $P(\text{error}|\mathbf{x})$ is as small as possible.)

# Bayes Decision Rule (with Equal Costs)

- Decide $\omega_1$ if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$; otherwise decide $\omega_2$
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min\left[P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})\right] \tag{12}$$

# Bayes Decision Rule (with Equal Costs)

- Decide $\omega_1$ if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$; otherwise decide $\omega_2$
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min\left[P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})\right] \quad (12)$$

- Equivalently, Decide $\omega_1$ if $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$; otherwise decide $\omega_2$
- I.e., the evidence term is not used in decision making.

# Bayes Decision Rule (with Equal Costs)

- Decide $\omega_1$ if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$; otherwise decide $\omega_2$
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min\left[P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})\right] \tag{12}$$

- Equivalently, Decide $\omega_1$ if $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$; otherwise decide $\omega_2$
- I.e., the evidence term is not used in decision making.
- If we have $p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2)$, then the decision will rely exclusively on the priors.
- Conversely, if we have uniform priors, then the decision will rely exclusively on the likelihoods.

# Bayes Decision Rule (with Equal Costs)

- Decide $\omega_1$ if $P(\omega_1|\mathbf{x}) > P(\omega_2|\mathbf{x})$; otherwise decide $\omega_2$
- Probability of error becomes

$$P(\text{error}|\mathbf{x}) = \min\left[P(\omega_1|\mathbf{x}), P(\omega_2|\mathbf{x})\right] \tag{12}$$

- Equivalently, Decide $\omega_1$ if $p(\mathbf{x}|\omega_1)P(\omega_1) > p(\mathbf{x}|\omega_2)P(\omega_2)$; otherwise decide $\omega_2$
- I.e., the evidence term is not used in decision making.
- If we have $p(\mathbf{x}|\omega_1) = p(\mathbf{x}|\omega_2)$, then the decision will rely exclusively on the priors.
- Conversely, if we have uniform priors, then the decision will rely exclusively on the likelihoods.
- Take Home Message: **Decision making relies on both the priors and the likelihoods and Bayes Decision Rule combines them to achieve the minimum probability of error.**

# 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 $\alpha_i$ when the class is $\omega_j$.

# Loss Functions

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

$$\lambda_{ij} \doteq \quad \lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases} \quad i, j = 1, 2, \ldots, c \tag{13}$$

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

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

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

$$R(\alpha_i|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) \tag{15}$$

$$= 1 - P(\omega_i|\mathbf{x}) \tag{16}$$

- Hence, for an observation $x$, we can minimize the expected loss by selecting the action that minimizes the conditional risk.

# 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

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

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

$$R(\alpha_i|\mathbf{x}) = \sum_{j \neq i} P(\omega_j|\mathbf{x}) \tag{15}$$

$$= 1 - P(\omega_i|\mathbf{x}) \tag{16}$$

- 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!

# Overall Risk

- Let $\alpha(x)$ denote a decision rule, a mapping from the input feature space to an action, $\mathbb{R}^d \mapsto \{\alpha_1, \ldots, \alpha_a\}$.
  - This is what we want to learn.

# Overall Risk

- Let $\alpha(x)$ denote a decision rule, a mapping from the input feature space to an action, $\mathbb{R}^d \mapsto \{\alpha_1, \ldots, \alpha_a\}$.
  - This is what we want to learn.
- The **overall risk** is the expected loss associated with a given decision rule.

$$R = \oint R(\alpha(\mathbf{x})|\mathbf{x})\, p(\mathbf{x})\, d\mathbf{x} \tag{17}$$

Clearly, we want the rule $\alpha(\cdot)$ that minimizes $R(\alpha(\mathbf{x})|\mathbf{x})$ for all $\mathbf{x}$.

# 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, $\alpha_1$ when the true class is $\omega_1$ and $\alpha_2$ for $\omega_2$.
- Writing out the conditional risks gives:

$$R(\alpha_1|\mathbf{x}) = \lambda_{11}P(\omega_1|\mathbf{x}) + \lambda_{12}P(\omega_2|\mathbf{x}) \quad (20)$$

$$R(\alpha_2|\mathbf{x}) = \lambda_{21}P(\omega_1|\mathbf{x}) + \lambda_{22}P(\omega_2|\mathbf{x}) \quad . \quad (21)$$

- Fundamental rule is decide $\omega_1$ if

$$R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x}) \quad . \quad (22)$$

- In terms of posteriors, decide $\omega_1$ if

$$(\lambda_{21} - \lambda_{11})P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})P(\omega_2|\mathbf{x}) \quad . \quad (23)$$

The more likely state of nature is scaled by the differences in loss (which are generally positive).

# Two-Category Classification Examples

- Or, expanding via Bayes Rule, decide $\omega_1$ if

$$(\lambda_{21} - \lambda_{11}) p(\mathbf{x}|\omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) p(\mathbf{x}|\omega_2) P(\omega_2) \tag{24}$$

- Or, assuming $\lambda_{21} > \lambda_{11}$, decide $\omega_1$ if

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\omega_2)}{P(\omega_1)} \tag{25}$$

  - LHS is called the **likelihood ratio**.
- Thus, we can say the Bayes Decision Rule says to decide $\omega_1$ if the likelihood ratio exceeds a threshold that is independent of the observation $\mathbf{x}$.

# 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 $i$th class.
- This classifier will assign a class $\omega_i$ to the feature vector $\mathbf{x}$ if

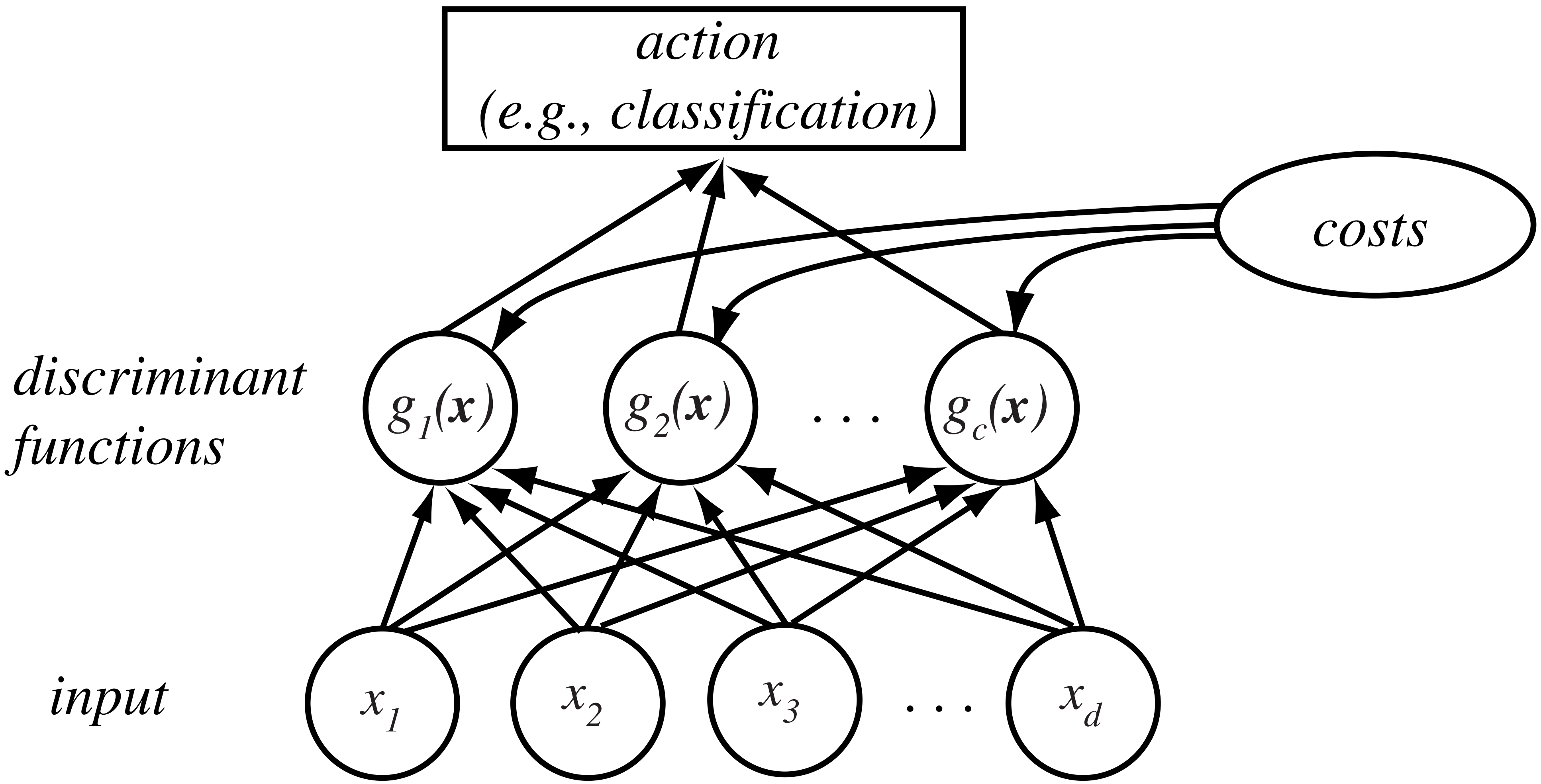
$$g_i(\mathbf{x}) > g_j(\mathbf{x}) \qquad \forall j \neq i \ , \tag{26}$$

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?

# 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?