

Bayesian Decision Theory

Bayesian Decision Theory

- Know probability distribution of the categories
 - Almost never the case in real life!
 - Nevertheless useful since other cases can be reduced to this one after some work
- Do not even need training data
- Can design optimal classifier

Bayesian Decision theory

Fish Example:

- Each fish is in one of 2 states: sea bass or salmon
- Let ω denote the **state of nature**
 - $\omega = \omega_1$ for sea bass
 - $\omega = \omega_2$ for salmon
- The state of nature is unpredictable ω is a variable that must be described probabilistically.
 - If the catch produced as much salmon as sea bass the next fish is equally likely to be sea bass or salmon.
- Define:
 - $P(\omega_1)$: **a priori** probability that the next fish is sea bass
 - $P(\omega_2)$: **a priori** probability that the next fish is salmon.

Bayesian Decision theory

- If other types of fish are irrelevant:

$$P(\omega_1) + P(\omega_2) = 1.$$

Prior probabilities reflect our prior knowledge (e.g. time of year, fishing area, ...)

- ***Simple decision Rule:***
 - *Make a decision without seeing the fish.*
 - *Decide ω_1 if $P(\omega_1) > P(\omega_2)$; ω_2 otherwise.*
 - *OK if deciding for one fish*
 - *If several fish, all assigned to same class*

In general, we have some features and more information.

Cats and Dogs

- Suppose we have these conditional probability mass functions for cats and dogs
 - $P(\text{small ears} \mid \text{dog}) = 0.1$, $P(\text{large ears} \mid \text{dog}) = 0.9$
 - $P(\text{small ears} \mid \text{cat}) = 0.8$, $P(\text{large ears} \mid \text{cat}) = 0.2$
- Observe an animal with large ears
 - Dog or a cat?
 - Makes sense to say dog because probability of observing large ears in a dog is much larger than probability of observing large ears in a cat
 - $\mathbf{Pr}[\text{large ears} \mid \text{dog}] = 0.9 > 0.2 = \mathbf{Pr}[\text{large ears} \mid \text{cat}] = 0.2$
 - We choose the event of larger probability, i.e. maximum likelihood event

Example: Fish Sorting

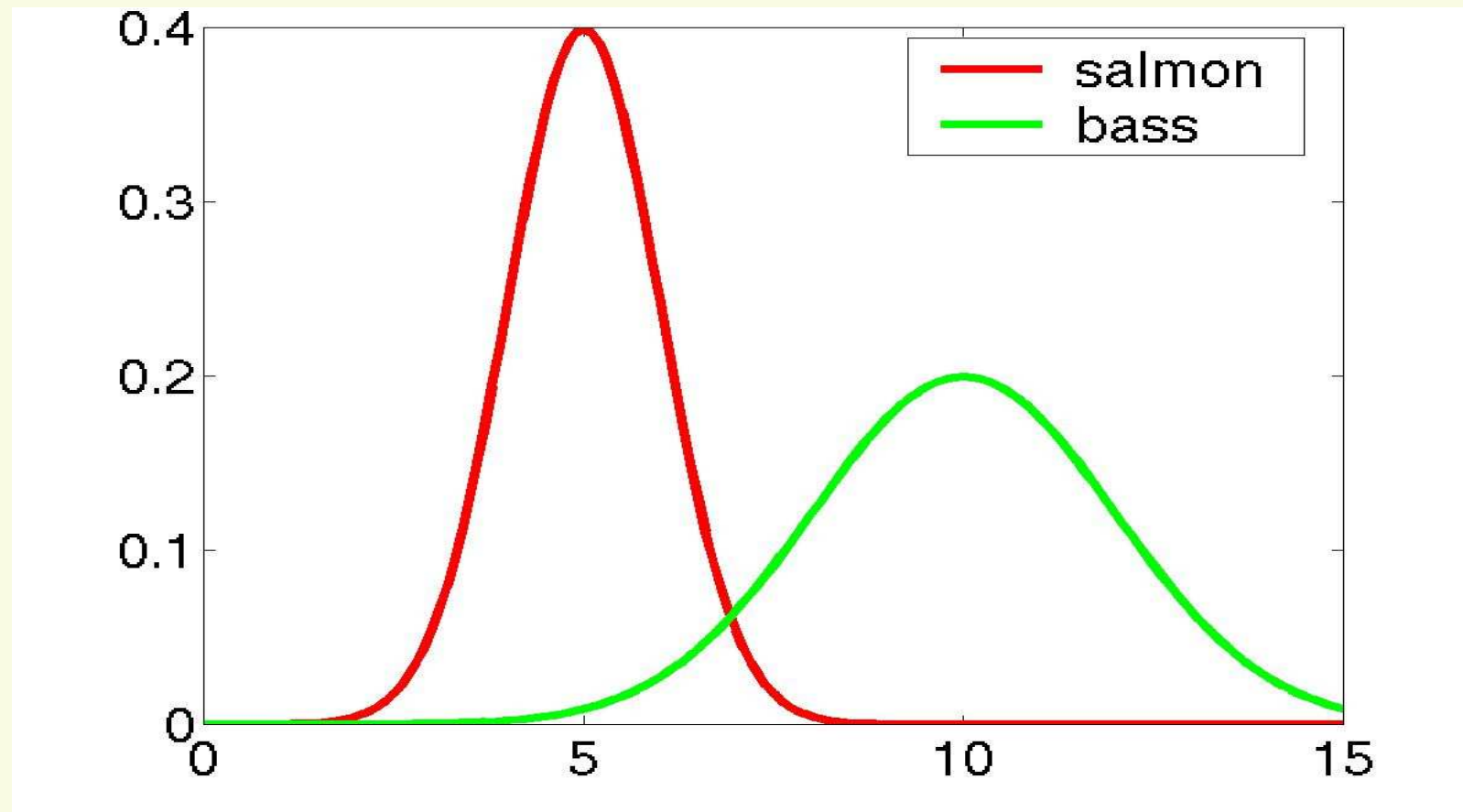
- Respected fish expert says that
 - Salmon' length has distribution $\mathbf{N}(5,1)$
 - Sea bass's length has distribution $\mathbf{N}(10,4)$
- Recall if r.v. is $\mathbf{N}(\mu, \sigma^2)$ then it's density is

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Class Conditional Densities

$$p(I | \text{salmon})_{\text{fixed}} = \frac{1}{\sqrt{2\pi}} e^{-\frac{(I-5)^2}{2}}$$

$$p(I | \text{bass})_{\text{fixed}} = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(I-10)^2}{2 \cdot 4}}$$

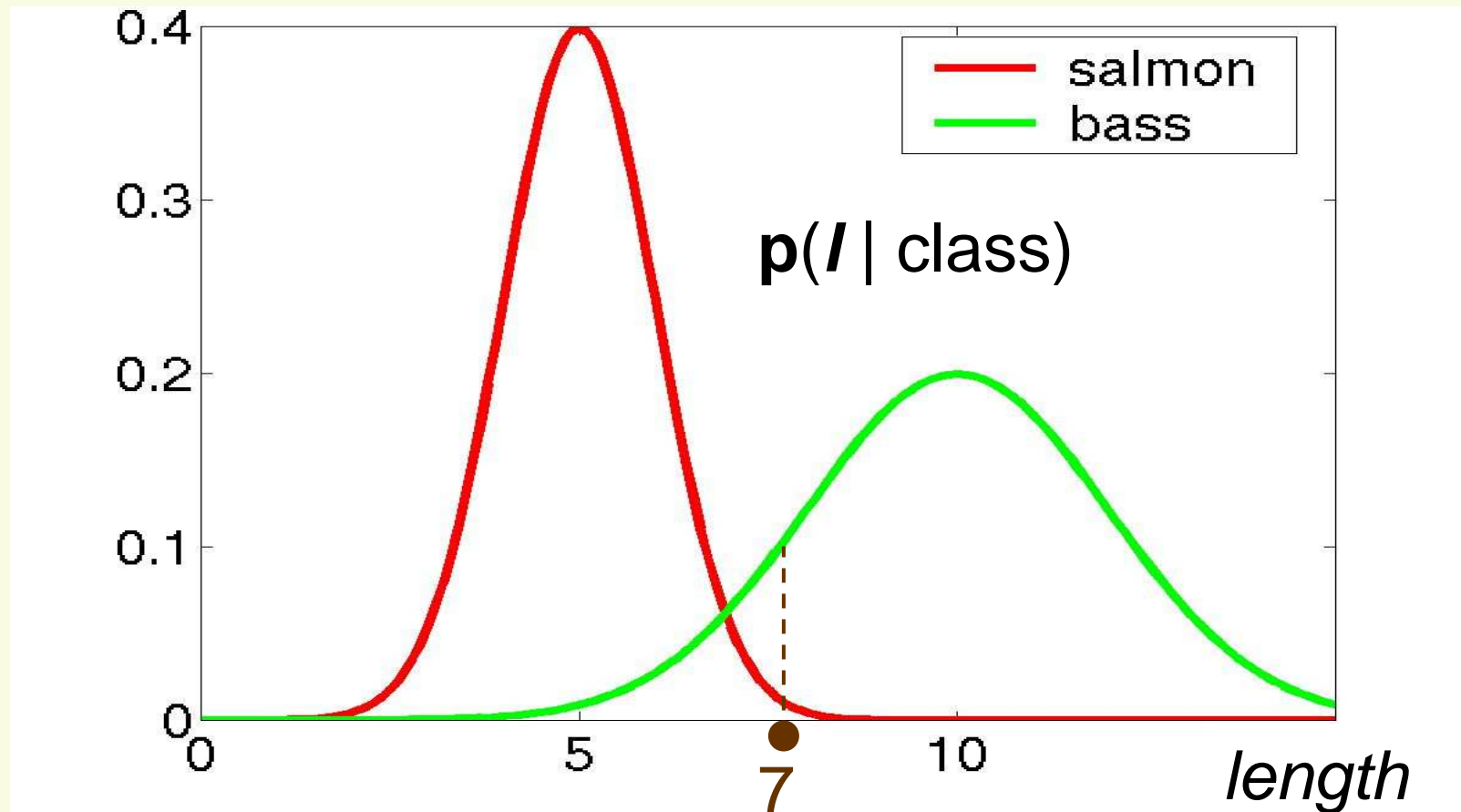


Likelihood function

- Fix length, let fish class vary. Then we get *likelihood function* (it is **not density** and **not probability mass**)

$$p(\underset{\substack{\uparrow \\ \text{fixed}}}{l} \mid \text{class}) = \begin{cases} \frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}} & \text{if class = salmon} \\ \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{8}} & \text{if class = bass} \end{cases}$$

Likelihood vs. Class Conditional Density



Suppose a fish has length 7. How do we classify it?

ML (maximum likelihood) Classifier

- We would like to choose salmon if

$$\Pr[\text{length}=7 \mid \text{salmon}] > \Pr[\text{length}=7 \mid \text{bass}]$$

- However, since *length* is a continuous r.v.,

$$\Pr[\text{length}=7 \mid \text{salmon}] = \Pr[\text{length}=7 \mid \text{bass}] = 0$$

- Instead, we choose class which maximizes likelihood

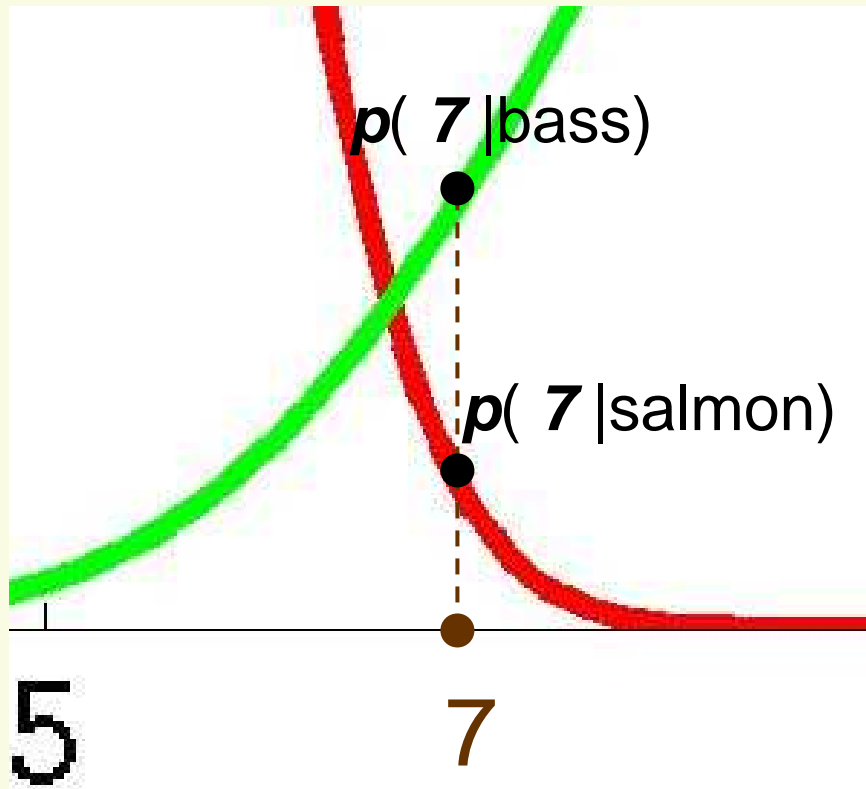
$$p(l \mid \text{salmon}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}} \quad p(l \mid \text{bass}) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{2 \cdot 4}}$$

- **ML classifier**: for an observed *l*:

$$p(l \mid \text{salmon}) \stackrel{\text{bass}}{<} ? p(l \mid \text{bass}) \\ > \text{salmon}$$

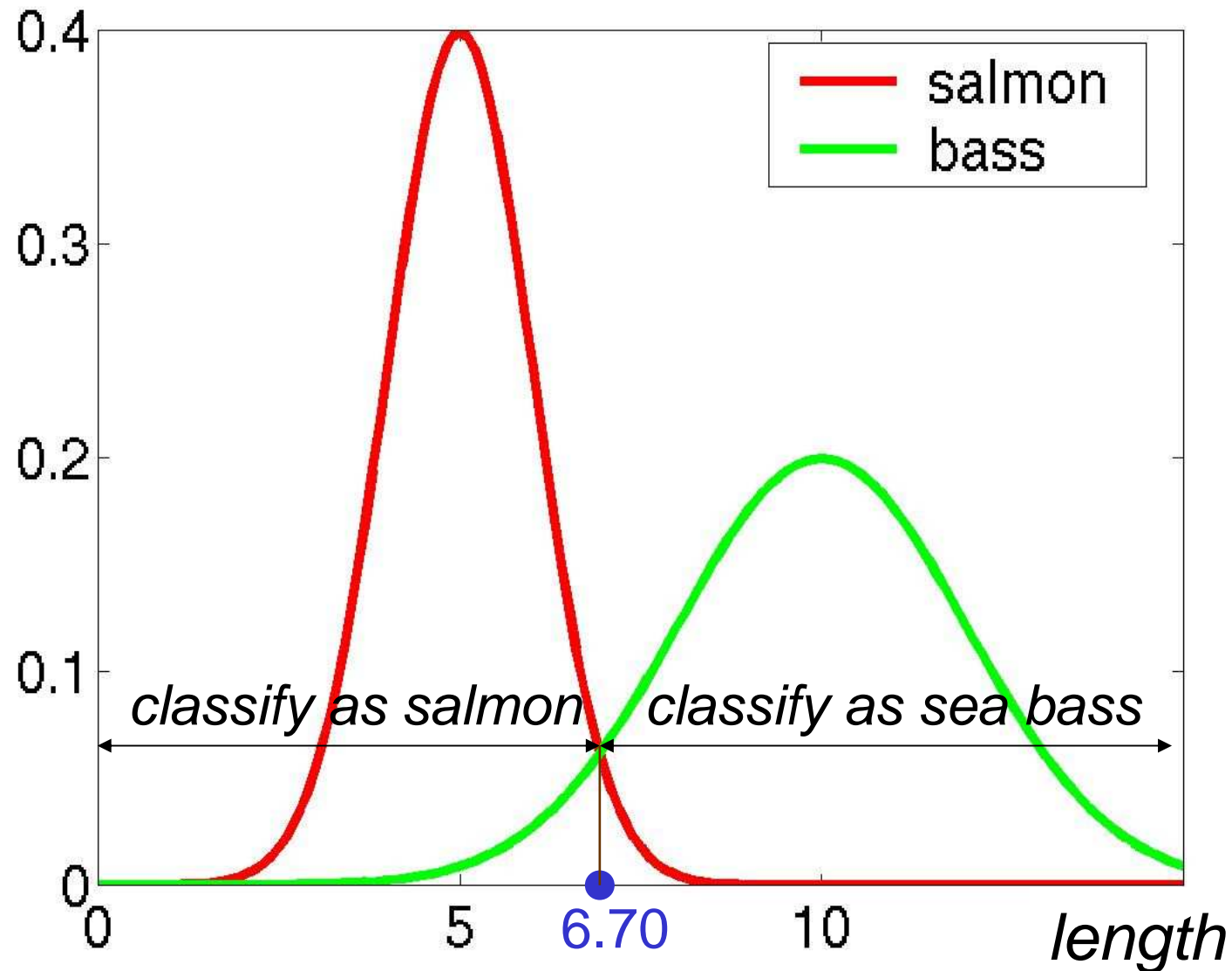
in words: if $p(l \mid \text{salmon}) > p(l \mid \text{bass})$,
classify as salmon, else classify as bass

ML (maximum likelihood) Classifier



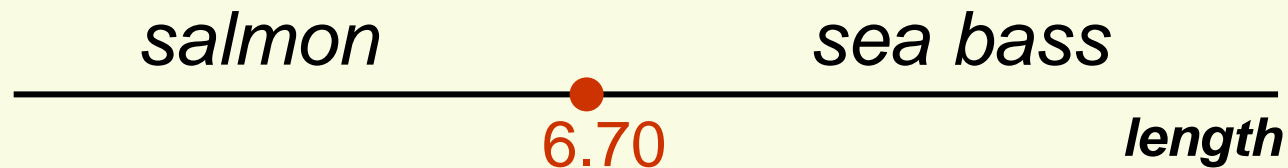
Thus we choose the class (bass) which is more likely to have given the observation

Decision Boundary

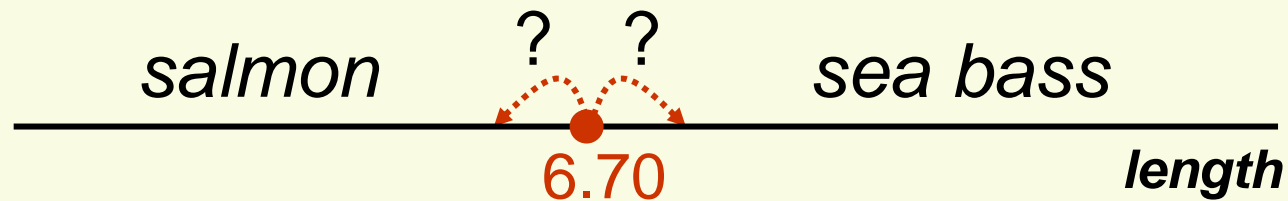


How Prior Changes Decision Boundary?

- Without priors



- How should this change with prior?
 - $P(\text{salmon}) = 2/3$
 - $P(\text{bass}) = 1/3$



Bayes Decision Rule

1. Have likelihood functions $p(\text{length} \mid \text{salmon})$ and $p(\text{length} \mid \text{bass})$
2. Have priors $P(\text{salmon})$ and $P(\text{bass})$
 - **Question:** Having observed fish of certain length, do we classify it as salmon or bass?
 - **Natural Idea:**
 - salmon if $P(\text{salmon} \mid \text{length}) > P(\text{bass} \mid \text{length})$
 - bass if $P(\text{bass} \mid \text{length}) > P(\text{salmon} \mid \text{length})$

Posterior

- $P(\text{salmon} \mid \text{length})$ and $P(\text{bass} \mid \text{length})$ are called **posterior** distributions, because the data (length) was revealed (post data)
- How to compute posteriors? Not obvious
- From Bayes rule:

$$P(\text{salmon} \mid \text{length}) = \frac{p(\text{length} \mid \text{salmon}) P(\text{salmon})}{p(\text{length})}$$

- Similarly:

$$P(\text{bass} \mid \text{length}) = \frac{p(\text{length} \mid \text{bass}) P(\text{bass})}{p(\text{length})}$$

MAP (maximum a posteriori) classifier

$$P(\text{salmon} | \text{length}) \stackrel{> \text{salmon}}{?} P(\text{bass} | \text{length})$$

bass <

$$\frac{p(\text{length} | \text{salmon})P(\text{salmon})}{p(\text{length})} \stackrel{> \text{salmon}}{?} \frac{p(\text{length} | \text{bass})P(\text{bass})}{p(\text{length})}$$

bass <

$$p(\text{length} | \text{salmon})P(\text{salmon}) \stackrel{> \text{salmon}}{?} p(\text{length} | \text{bass})P(\text{bass})$$

bass <

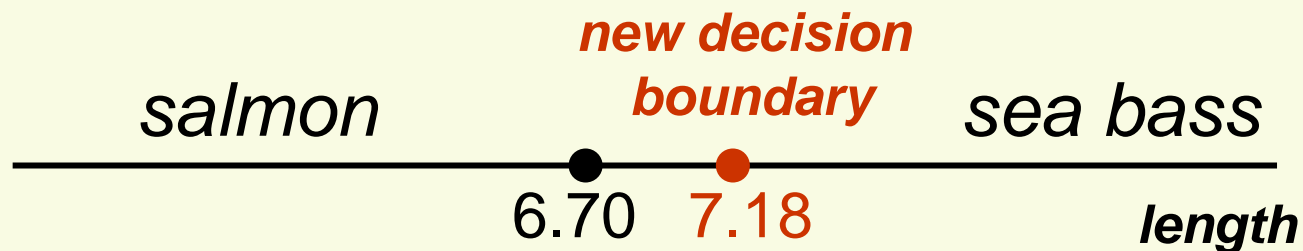
Back to Fish Sorting Example

- Likelihood

$$p(l | \text{salmon}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}} \quad p(l | \text{bass}) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{8}}$$

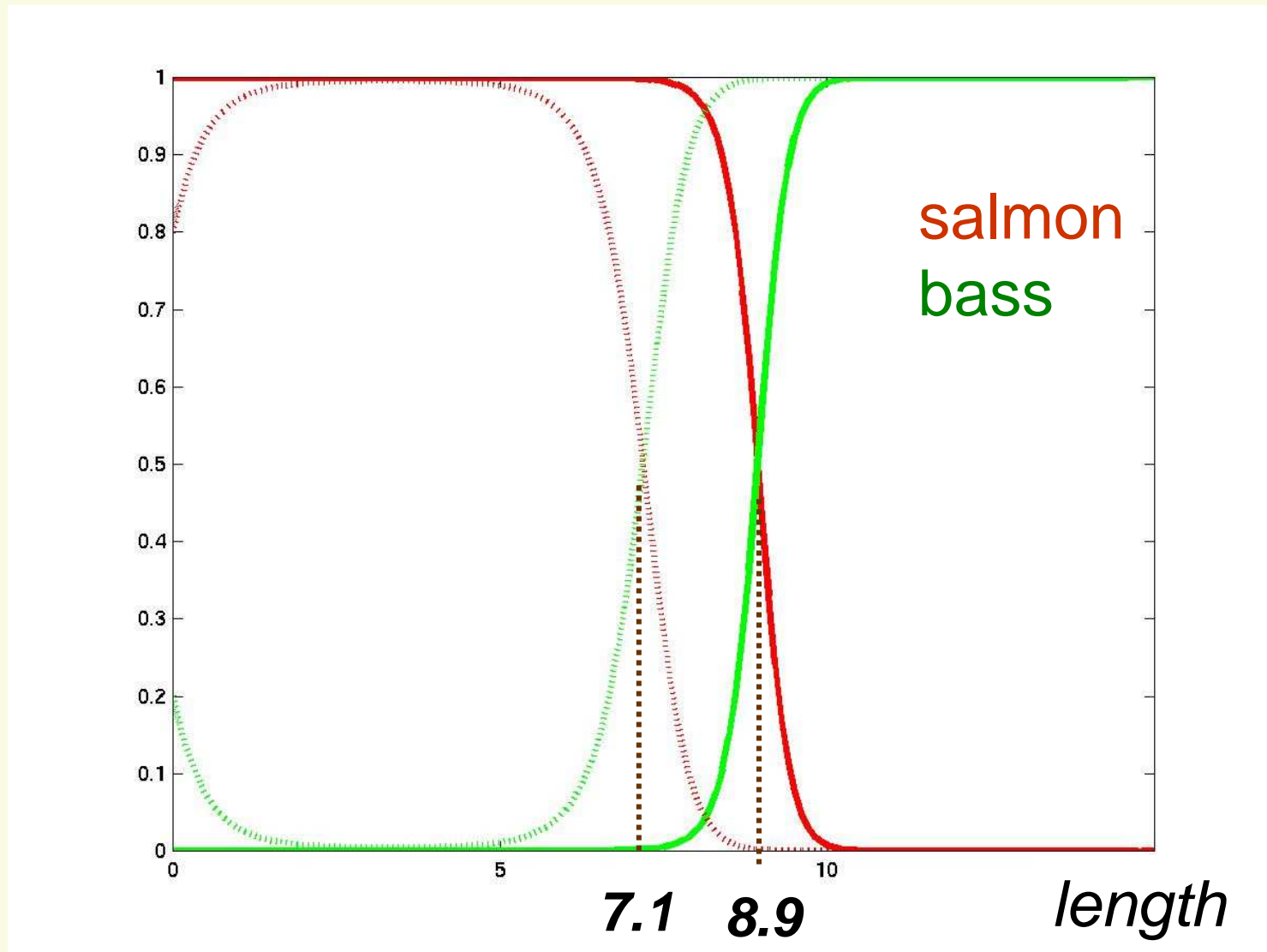
- Priors: $P(\text{salmon}) = 2/3$, $P(\text{bass}) = 1/3$

- Solve inequality $\frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}} * \frac{2}{3} > \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{8}} * \frac{1}{3}$

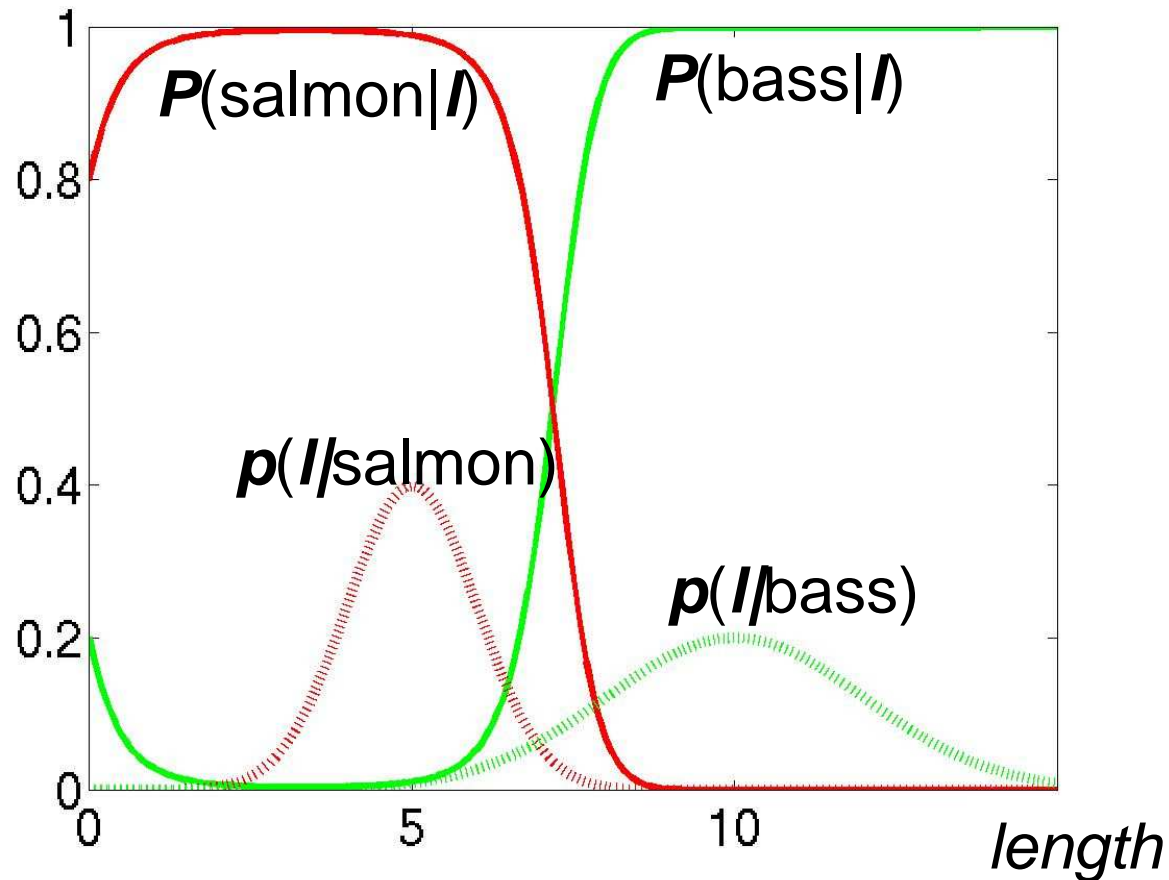


- New decision boundary makes sense since we expect to see more salmon

*Prior $P(\mathbf{s})=2/3$ and $P(\mathbf{b})= 1/3$ vs.
Prior $P(\mathbf{s})=0.999$ and $P(\mathbf{b})= 0.001$*



Likelihood vs Posteriors



likelihood
 $p(l|\text{fish class})$

density with respect to length, area under the curve is 1

posterior **$P(\text{fish class}|l)$**

mass function with respect to fish class, so for each l , $P(\text{salmon}|l) + P(\text{bass}|l) = 1$

More on Posterior

<i>posterior density (our goal)</i>	<i>likelihood (given)</i>	<i>Prior (given)</i>
$P(\mathbf{c} I)$	$P(I \mathbf{c})$	$P(\mathbf{c})$

$= \frac{\quad}{P(I)}$

normalizing factor, often do not even need it for classification since $P(I)$ does not depend on class \mathbf{c} . If we do need it, from the law of total probability:

$$P(I) = p(I | \text{salmon})p(\text{salmon}) + p(I | \text{bass})p(\text{bass})$$

Notice this formula consists of likelihoods and priors, which are given

More on Priors

- Prior comes from prior knowledge, no data has been seen yet
- If there is a reliable source prior knowledge, it should be used
- Some problems cannot even be solved reliably without a good prior

More on Map Classifier

$$\text{posterior } P(\mathbf{c} | I) = \frac{\text{likelihood } P(I | \mathbf{c}) \text{ prior } P(\mathbf{c})}{P(I)}$$

- Do not care about $P(I)$ when maximizing $P(\mathbf{c} | I)$

$$P(\mathbf{c} | I) \stackrel{\text{proportional}}{\propto} P(I | \mathbf{c}) P(\mathbf{c})$$

- If $P(\text{salmon}) = P(\text{bass})$ (uniform prior) MAP classifier becomes ML classifier $P(\mathbf{c} | I) \propto P(I | \mathbf{c})$
- If for some observation I , $P(I | \text{salmon}) = P(I | \text{bass})$, then this observation is uninformative and decision is based solely on the prior $P(\mathbf{c} | I) \propto P(\mathbf{c})$

Justification for MAP Classifier

- Let's compute probability of error for the MAP estimate:

$$P(\text{salmon} | I) \stackrel{\text{salmon}}{>} ? P(\text{bass} | I) \stackrel{\text{bass}}{<}$$

- For any particular I , probability of error

$$\Pr[\text{error} | I] = \begin{cases} P(\text{bass} | I) & \text{if we decide salmon} \\ P(\text{salmon} | I) & \text{if we decide bass} \end{cases}$$

Thus MAP classifier is optimal for each individual I !

Justification for MAP Classifier

- We are interested to minimize error not just for one I , we really want to minimize the average error over all I

$$\Pr[\text{error}] = \int_{-\infty}^{\infty} p(\text{error}, I) dI = \int_{-\infty}^{\infty} \Pr[\text{error} | I] p(I) dI$$

- If $\Pr[\text{error} | I]$ is as small as possible, the integral is small as possible
- But Bayes rule makes $\Pr[\text{error} | I]$ as small as possible

Thus MAP classifier minimizes the probability of error!

More General Case

- Let's generalize a little bit
 - Have more than one feature $\mathbf{x} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_d]$
 - Have more than 2 classes $\{\mathbf{c}_1, \mathbf{c}_2, \dots, \mathbf{c}_m\}$

More General Case

- As before, for each j we have
 - $p(\mathbf{x} / \mathbf{c}_j)$ is likelihood of observation \mathbf{x} given that the true class is \mathbf{c}_j
 - $P(\mathbf{c}_j)$ is prior probability of class \mathbf{c}_j
 - $P(\mathbf{c}_j / \mathbf{x})$ is posterior probability of class \mathbf{c}_j given that we observed data \mathbf{x}
- Evidence, or probability density for data

$$p(\mathbf{x}) = \sum_{j=1}^m p(\mathbf{x} / \mathbf{c}_j) P(\mathbf{c}_j)$$

Minimum Error Rate Classification

- Want to minimize average probability of error

$$Pr[\text{error}] = \int p(\text{error}, \mathbf{x}) d\mathbf{x} = \int Pr[\text{error} / \mathbf{x}] p(\mathbf{x}) d\mathbf{x}$$

need to make this as small as possible

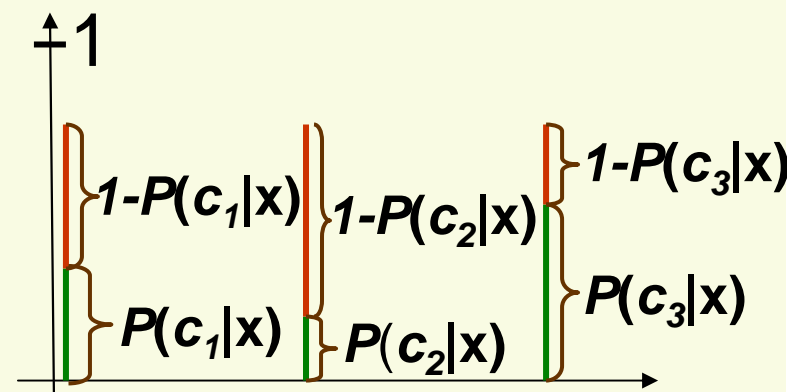
- $Pr[\text{error} / \mathbf{x}] = 1 - P(\mathbf{c}_i / \mathbf{x})$ if we decide class \mathbf{c}_i

- $Pr[\text{error} / \mathbf{x}]$ is minimized with MAP classifier

- Decide on class \mathbf{c}_i if

$$P(\mathbf{c}_i / \mathbf{x}) > P(\mathbf{c}_j / \mathbf{x}) \quad \forall j \neq i$$

*MAP classifier is optimal
If we want to minimize the
probability of error*



General Bayesian Decision Theory

- In close cases we may want to refuse to make a decision (let human expert handle tough case)
 - allow actions $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$
- Suppose some mistakes are more costly than others (classifying a benign tumor as cancer is not as bad as classifying cancer as benign tumor)
 - Allow loss functions $\lambda(\alpha_i / \mathbf{c}_j)$ describing loss occurred when taking action α_i when the true class is \mathbf{c}_j

Conditional Risk

- Suppose we observe \mathbf{x} and wish to take action α_i
- If the true class is \mathbf{c}_j , by definition, we incur loss $\lambda(\alpha_i / \mathbf{c}_j)$
- Probability that the true class is \mathbf{c}_j after observing \mathbf{x} is $P(\mathbf{c}_j / \mathbf{x})$
- The expected loss associated with taking action α_i is called **conditional risk** and it is:

$$R(\alpha_i / \mathbf{x}) = \sum_{j=1}^m \lambda(\alpha_i / \mathbf{c}_j) P(\mathbf{c}_j / \mathbf{x})$$

Conditional Risk

*sum over disjoint events
(different classes)*

*probability of
class \mathbf{c}_j given
observation x*

$$\underbrace{R(\alpha_i | \mathbf{x})}_{\text{penalty for taking action } \alpha_i \text{ if observe } x} = \sum_{j=1}^m \underbrace{\lambda(\alpha_i | \mathbf{c}_j) P(\mathbf{c}_j | \mathbf{x})}_{\text{part of overall penalty which comes from event that true class is } \mathbf{c}_j}$$

*penalty for
taking action α_i
if observe x*

*part of overall penalty
which comes from event
that true class is \mathbf{c}_j*

Example: Zero-One loss function

- action α_i is decision that true class is \mathbf{c}_i

$$\lambda(\alpha_i | \mathbf{c}_j) = \begin{cases} \mathbf{0} & \text{if } i = j & \text{(no mistake)} \\ \mathbf{1} & \text{otherwise} & \text{(mistake)} \end{cases}$$

$$\begin{aligned} R(\alpha_i | \mathbf{x}) &= \sum_{j=1}^m \lambda(\alpha_i | \mathbf{c}_j) P(\mathbf{c}_j | \mathbf{x}) = \sum_{i \neq j} P(\mathbf{c}_j | \mathbf{x}) = \\ &= 1 - P(\mathbf{c}_i | \mathbf{x}) = \text{Pr}[\text{error if decide } \mathbf{c}_i] \end{aligned}$$

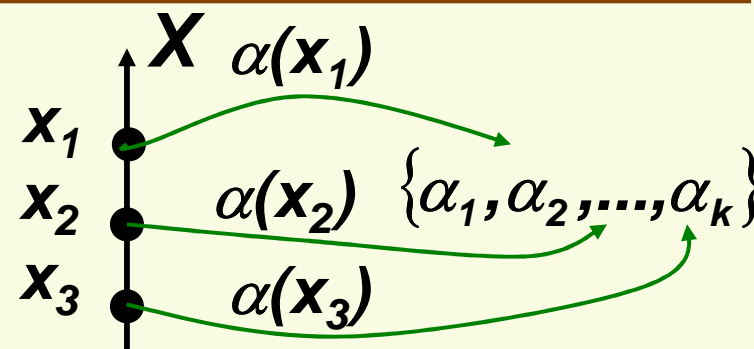
- Thus MAP classifier optimizes $R(\alpha_i | \mathbf{x})$

$$P(\mathbf{c}_i | \mathbf{x}) > P(\mathbf{c}_j | \mathbf{x}) \quad \forall j \neq i$$

- MAP classifier is Bayes decision rule under zero-one loss function

Overall Risk

- Decision rule is a function $\alpha(\mathbf{x})$ which for every \mathbf{x} specifies action out of $\{\alpha_1, \alpha_2, \dots, \alpha_k\}$



- The average risk for $\alpha(\mathbf{x})$

$$R(\alpha) = \int R(\alpha(\mathbf{x}) / \mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

need to make this as small as possible

- Bayes decision rule** $\alpha(\mathbf{x})$ for every \mathbf{x} is the action which minimizes the conditional risk

$$R(\alpha_i / \mathbf{x}) = \sum_{j=1}^m \lambda(\alpha_i / \mathbf{c}_j) P(\mathbf{c}_j / \mathbf{x})$$

- Bayes decision rule $\alpha(\mathbf{x})$ is **optimal**, i.e. gives the minimum possible overall risk R^*

Bayes Risk: Example

- Salmon is more tasty and expensive than sea bass

$\lambda_{sb} = \lambda(\mathbf{salmon} | \mathbf{bass}) = 2$ *classify bass as salmon*

$\lambda_{bs} = \lambda(\mathbf{bass} | \mathbf{salmon}) = 1$ *classify salmon as bass*

$\lambda_{ss} = \lambda_{bb} = 0$ *no mistake, no loss*

- Likelihoods $p(I | \mathbf{salmon}) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(I-5)^2}{2}}$ $p(I | \mathbf{bass}) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(I-10)^2}{2 \cdot 4}}$

- Priors $P(\mathbf{salmon}) = P(\mathbf{bass})$

- Risk $R(\alpha | \mathbf{x}) = \sum_{j=1}^m \lambda(\alpha | \mathbf{c}_j) P(\mathbf{c}_j | \mathbf{x}) = \lambda_{\alpha s} P(\mathbf{s} | I) + \lambda_{\alpha b} P(\mathbf{b} | I)$

$$R(\mathbf{salmon} | I) = \lambda_{ss} P(\mathbf{s} | I) + \lambda_{sb} P(\mathbf{b} | I) = \lambda_{sb} P(\mathbf{b} | I)$$

$$R(\mathbf{bass} | I) = \lambda_{bs} P(\mathbf{s} | I) + \lambda_{bb} P(\mathbf{b} | I) = \lambda_{bs} P(\mathbf{s} | I)$$

Bayes Risk: Example

$$R(\text{salmon} | I) = \lambda_{sb} P(\mathbf{b} | I) \quad R(\text{bass} | I) = \lambda_{bs} P(\mathbf{s} | I)$$

- Bayes decision rule (**optimal** for our loss function)

$$\lambda_{sb} P(\mathbf{b} | I) \begin{matrix} < \text{salmon} \\ ? \\ > \text{bass} \end{matrix} \lambda_{bs} P(\mathbf{s} | I)$$

- Need to solve $\frac{P(\mathbf{b} | I)}{P(\mathbf{s} | I)} < \frac{\lambda_{bs}}{\lambda_{sb}}$

- Or, equivalently, since priors are equal:

$$\frac{P(I | \mathbf{b})P(\mathbf{b})p(I)}{p(I)P(I | \mathbf{s})P(\mathbf{s})} = \frac{P(I | \mathbf{b})}{P(I | \mathbf{s})} < \frac{\lambda_{bs}}{\lambda_{sb}}$$

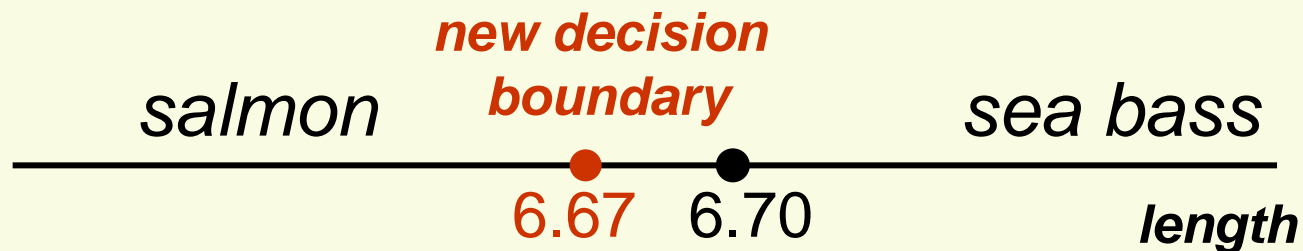
Bayes Risk: Example

- Need to solve $\frac{P(I|b)}{P(I|s)} < \frac{\lambda_{bs}}{\lambda_{sb}}$

- Substituting likelihoods and losses

$$\frac{2 \cdot \sqrt{2\pi} \exp\left[-\frac{(l-10)^2}{8}\right]}{1 \cdot 2\sqrt{2\pi} \exp\left[-\frac{(l-5)^2}{2}\right]} < 1 \Leftrightarrow \frac{\exp\left[-\frac{(l-10)^2}{8}\right]}{\exp\left[-\frac{(l-5)^2}{2}\right]} < 1 \Leftrightarrow \ln\left(\frac{\exp\left[-\frac{(l-10)^2}{8}\right]}{\exp\left[-\frac{(l-5)^2}{2}\right]}\right) < \ln(1) \Leftrightarrow$$

$$\Leftrightarrow -\frac{(l-10)^2}{8} + \frac{(l-5)^2}{2} < 0 \Leftrightarrow 3l^2 - 20l < 0 \Leftrightarrow 0 \leq l < 6.6667$$



Likelihood Ratio Rule

- In 2 category case, use likelihood ratio rule

$$\frac{P(\mathbf{x} | \mathbf{c}_1)}{P(\mathbf{x} | \mathbf{c}_2)} > \frac{\lambda_{12} - \lambda_{22}}{\lambda_{21} - \lambda_{11}} \frac{P(\mathbf{c}_2)}{P(\mathbf{c}_1)}$$

likelihood ratio *fixed number Independent of x*

- If above inequality holds, decide \mathbf{c}_1
- Otherwise decide \mathbf{c}_2

Discriminant Functions

- All decision rules have the same structure:
at observation \mathbf{x} choose class \mathbf{c}_i s.t.

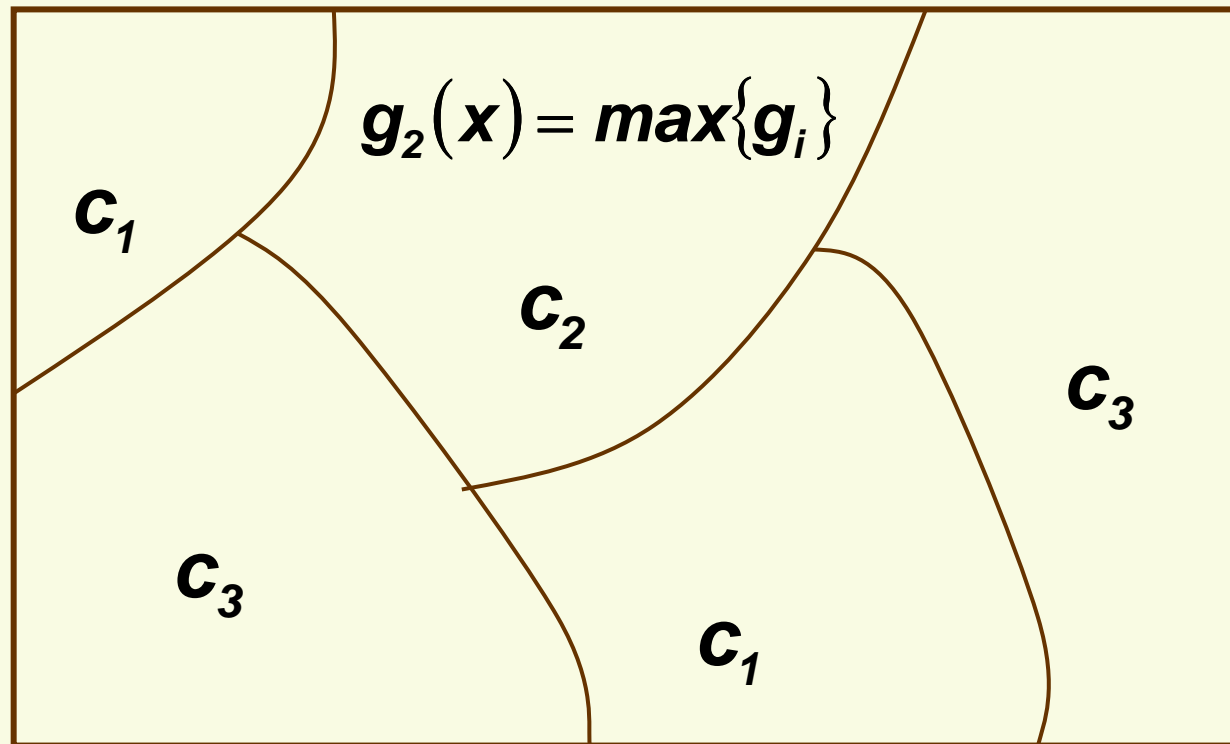
$$\mathbf{g}_i(\mathbf{x}) > \mathbf{g}_j(\mathbf{x}) \quad \forall j \neq i$$

*discriminant
function*

- ML decision rule: $\mathbf{g}_i(\mathbf{x}) = P(\mathbf{x} / \mathbf{c}_i)$
- MAP decision rule: $\mathbf{g}_i(\mathbf{x}) = P(\mathbf{c}_i / \mathbf{x})$
- Bayes decision rule: $\mathbf{g}_i(\mathbf{x}) = -R(\mathbf{c}_i / \mathbf{x})$

Decision Regions

- Discriminant functions split the feature vector space X into decision regions



Important Points

- If we know probability distributions for the classes, we can design the **optimal classifier**
- Definition of “optimal” depends on the chosen loss function
 - Under the minimum error rate (zero-one loss function)
 - No prior: ML classifier is optimal
 - Have prior: MAP classifier is optimal
 - More general loss function
 - General Bayes classifier is optimal