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(small ears | dog) = 0.1, P(large ears | dog) = 0.9
  - P(small ears | cat) = 0.8, P(large ears | 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
    - ***P***[large ears | dog] = 0.9 > 0.2= ***P***[large ears | cat] = 0.2
  - We choose the event of larger probability, i.e. maximum likelihood event

Example: Fish Sorting

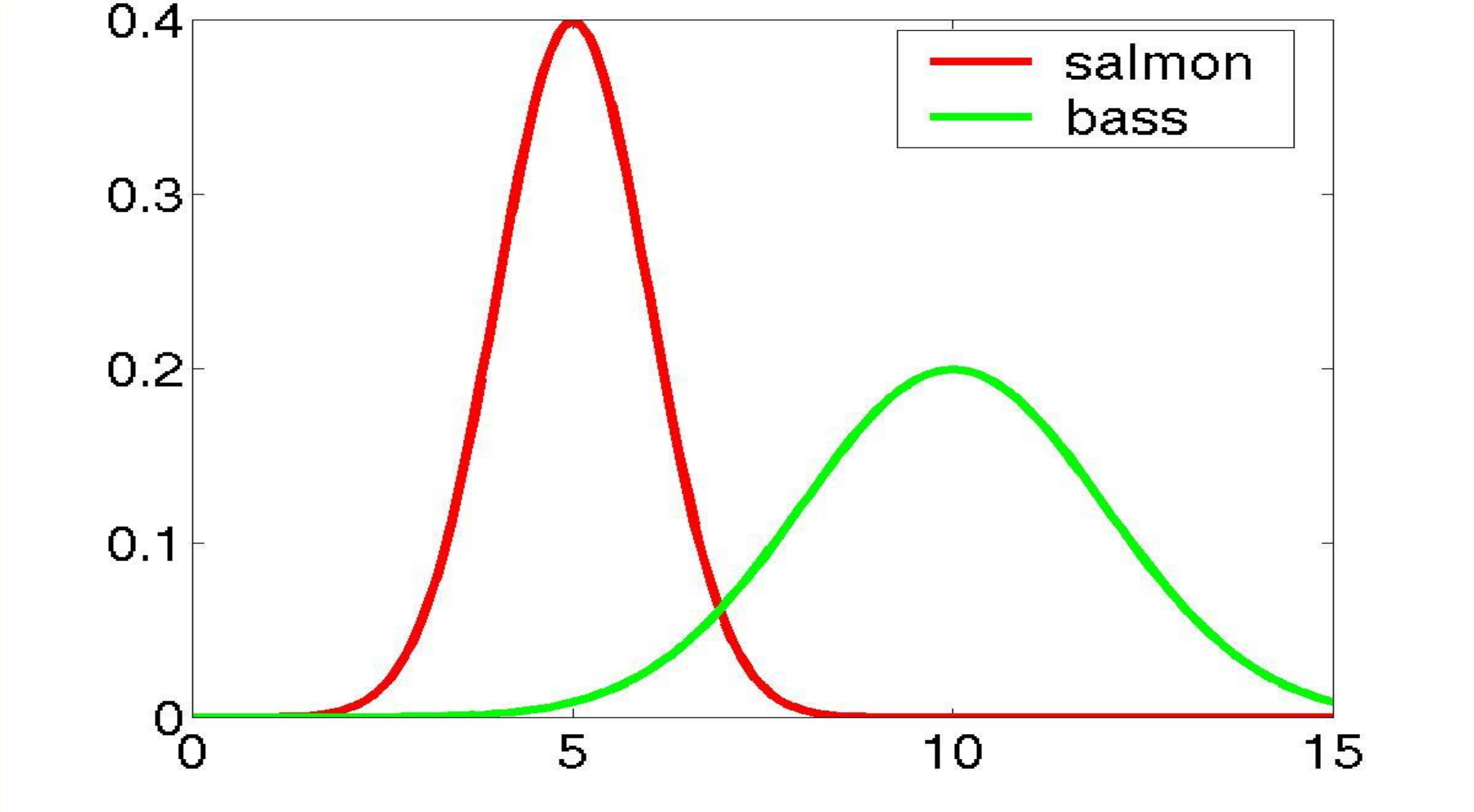
- Respected fish expert says that
  - Salmon' length has distribution ***N***(5,1)
  - Sea bass's length has distribution ***N***(10,4)

- Recall if r.v. is $\boldsymbol{N(\mu,\sigma^2)}$ then it's density is

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

Class Conditional Densities

$$p\left(\underset{\textit{fixed}}{l \mid salmon}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}} \qquad p\left(\underset{\textit{fixed}}{l \mid bass}\right) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{2*4}}$$



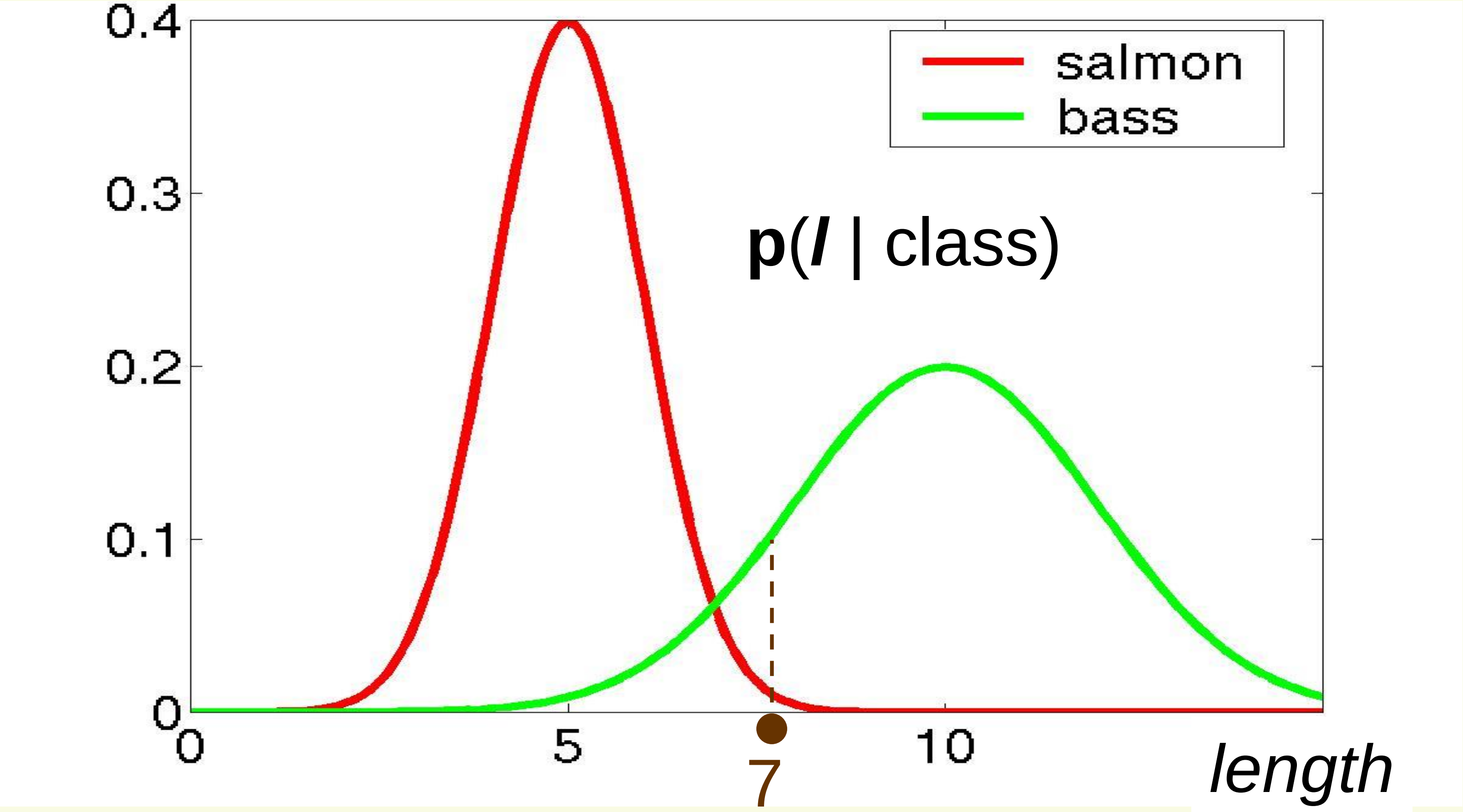
Likelihood function

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

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

fixed

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

$$\Pr\left[\,length = 7 \mid salmon\,\right] > \Pr\left[\,length = 7 \mid bass\,\right]$$

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

$$\Pr\left[\,length = 7 \mid salmon\,\right] = \Pr\left[\,length = 7 \mid bass\,\right] = 0$$

- Instead, we choose class which maximizes likelihood

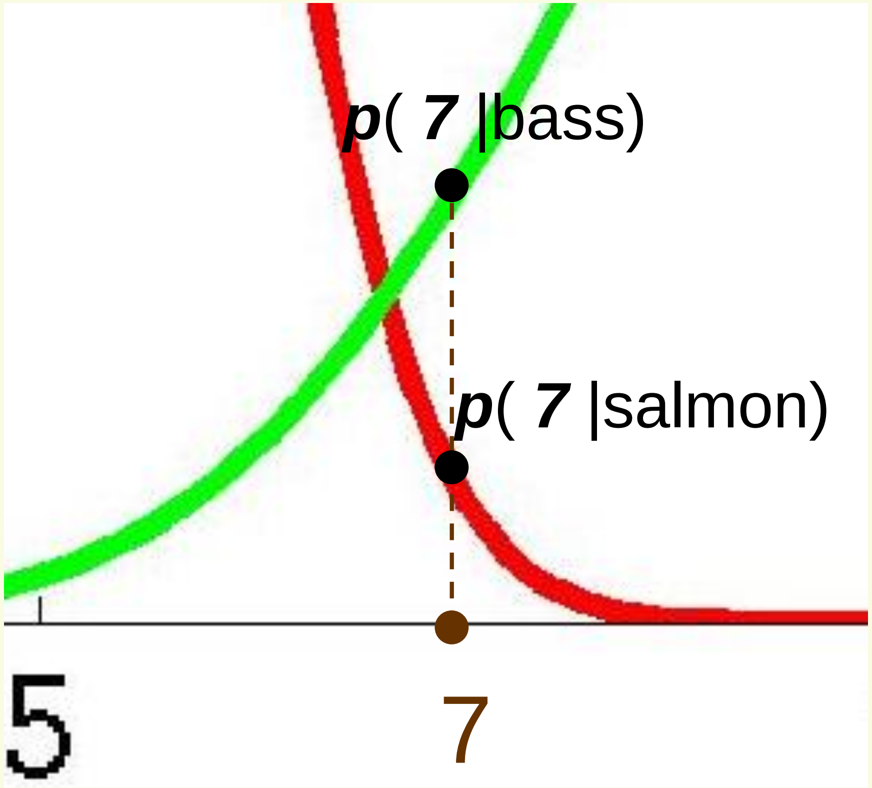
$$p(l \mid salmon) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}} \qquad\qquad p(l \mid bass) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{2*4}}$$

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

$$p(l \mid salmon) \underset{salmon}{\overset{bass}{\underset{>}{\overset{<}{?}}}} p(l \mid bass)$$

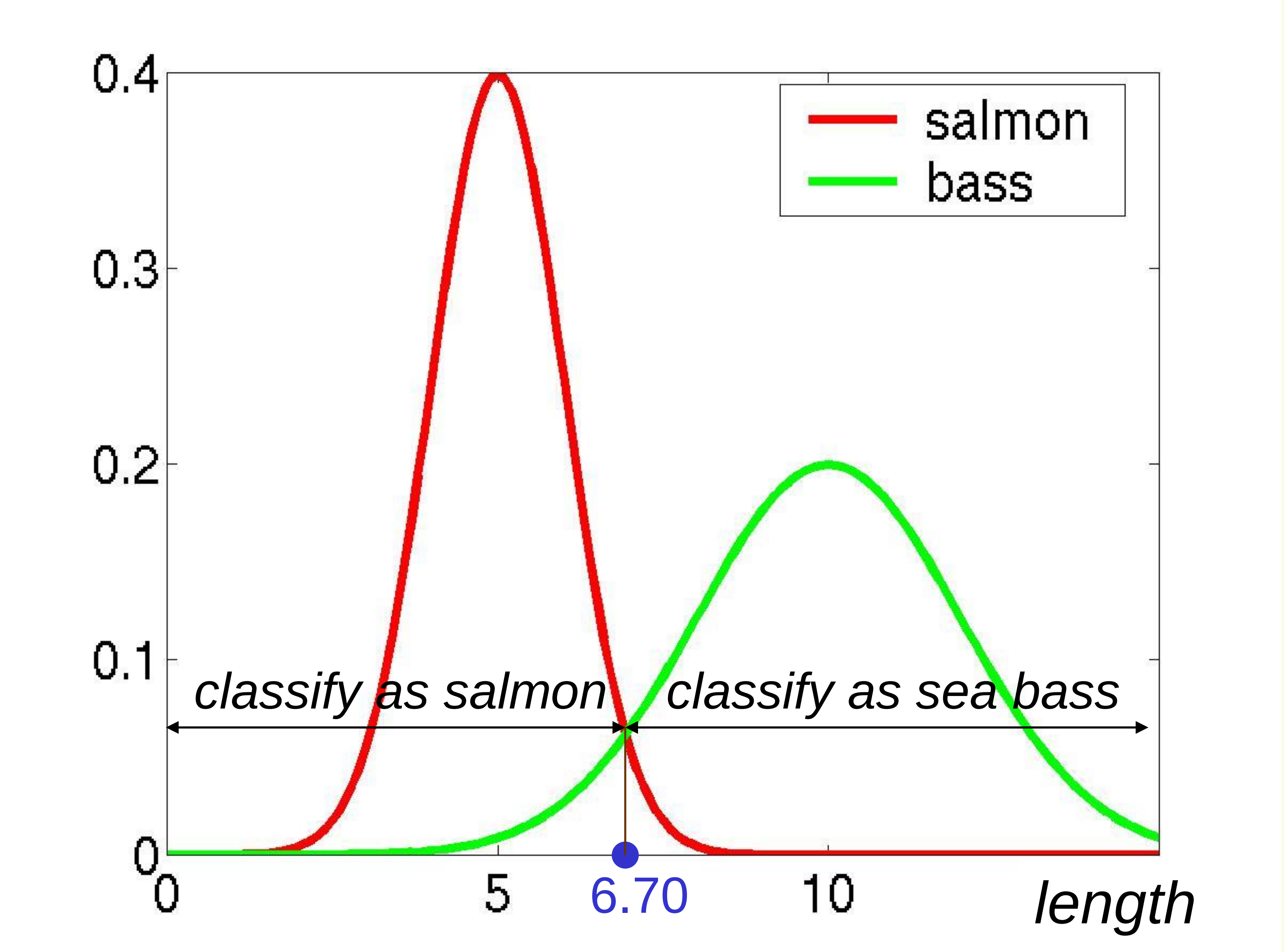
in words: if p(l | salmon) > p(l | bass), classify as salmon, else classify as bass

ML (maximum likelihood) Classifier



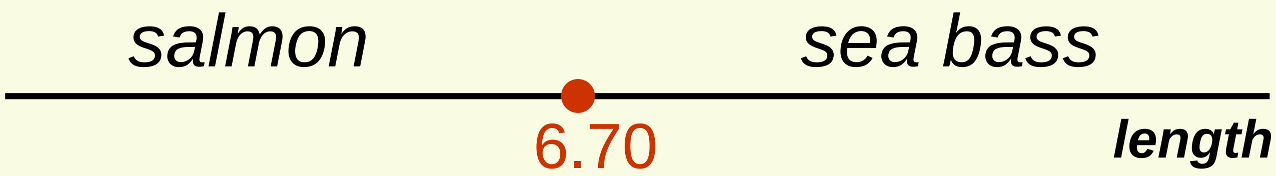
Thus we choose the class (bass) which is more likely to have given the observation

Decision Boundary

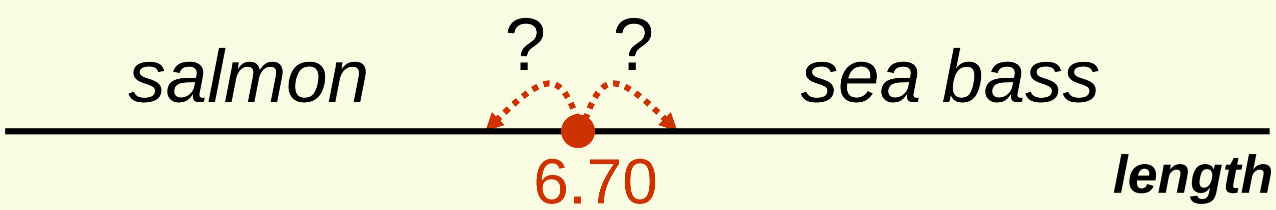


How Prior Changes Decision Boundary?

- Without priors



- How should this change with prior?
  - ***P***(salmon) = 2/3
  - ***P***(bass) = 1/3



Bayes Decision Rule

1. Have likelihood functions ***p***(length | salmon) and ***p***(length | bass)
2. Have priors ***P***(salmon) and ***P***(bass)

- Question: Having observed fish of certain length, do we classify it as salmon or bass?

- Natural Idea:
  - salmon if $P(salmon \mid length) > P(bass \mid length)$
  - bass if $P(bass \mid length) > P(salmon \mid length)$

Posterior

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

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

- Similarly:

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

MAP (maximum a posteriori) classifier

$$P(salmon \mid length) \underset{bass}{\overset{salmon}{\gtrless}} P(bass \mid length)$$

$$\frac{p(length \mid salmon)P(salmon)}{p(length)} \underset{bass}{\overset{salmon}{\gtrless}} \frac{p(length \mid bass)P(bass)}{p(length)}$$

$$p(length \mid salmon)P(salmon) \underset{bass}{\overset{salmon}{\gtrless}} p(length \mid bass)P(bass)$$

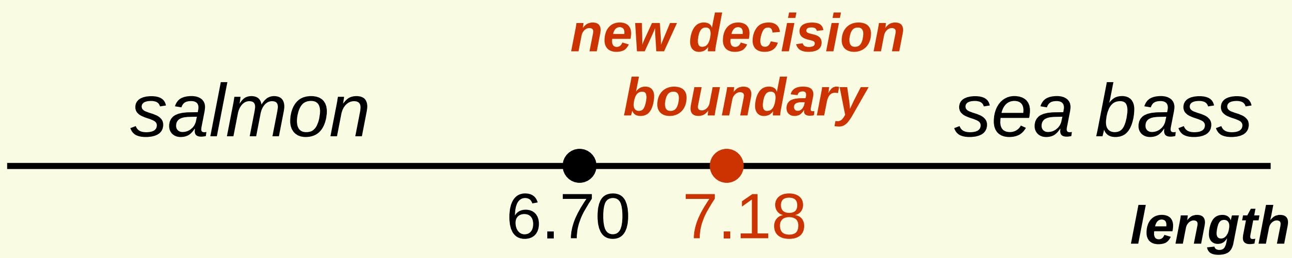
Back to Fish Sorting Example

- Likelihood

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

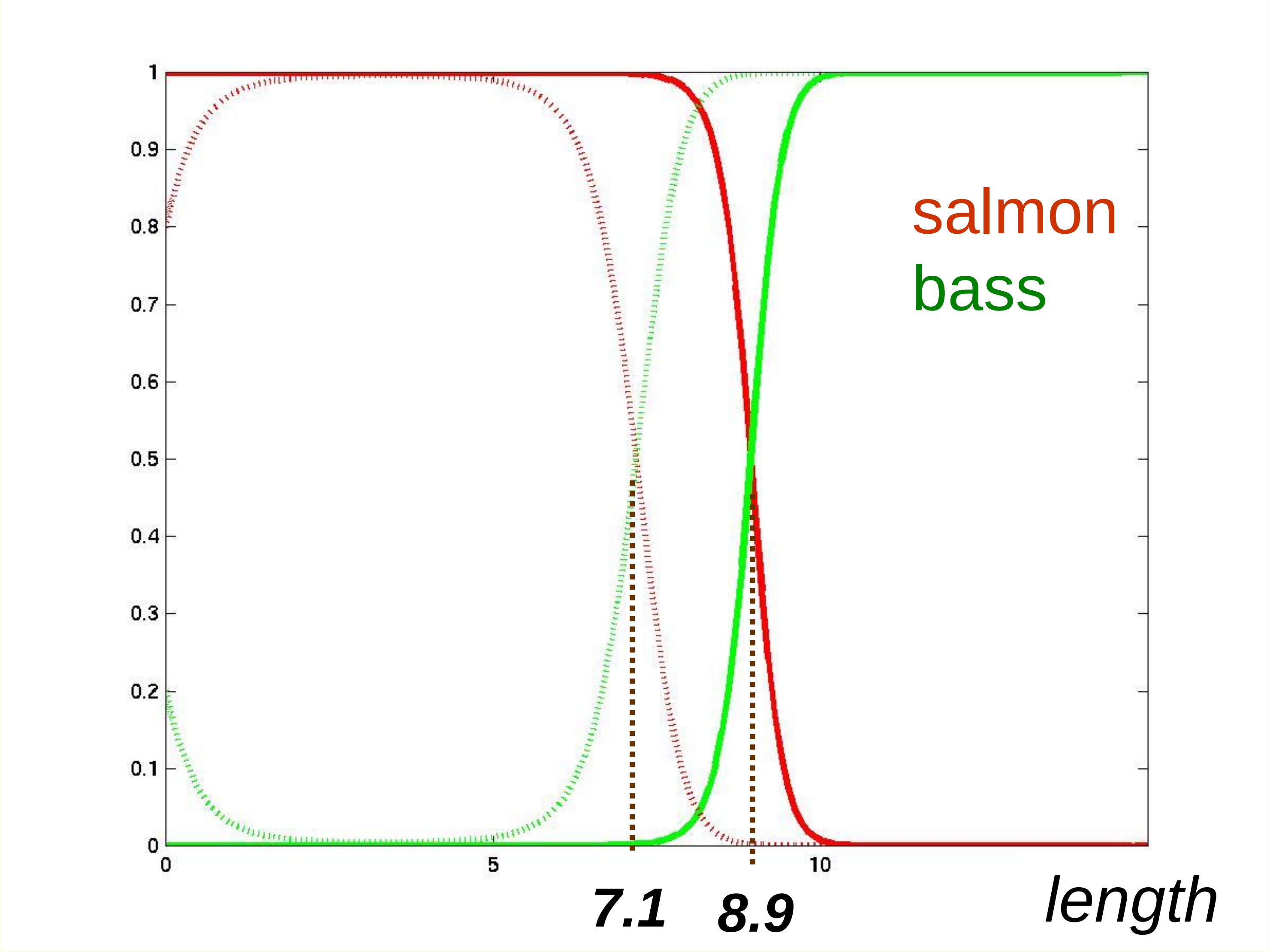
- Priors: $\boldsymbol{P}$(salmon) = 2/3, $\boldsymbol{P}$(bass) = 1/3

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

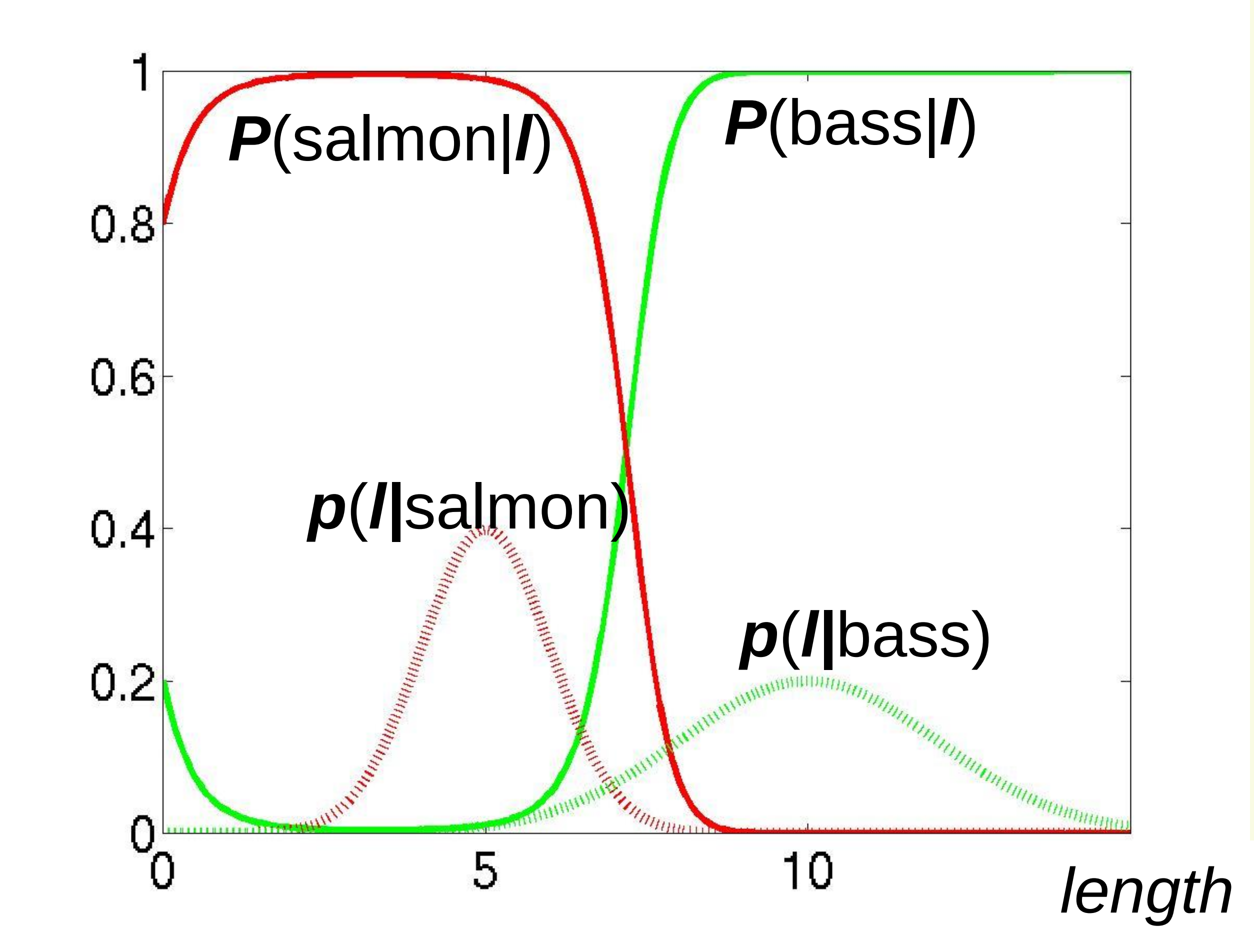


- New decision boundary makes sense since we expect to see more salmon

*Prior **P(s)**=2/3 and **P(b)**= 1/3 vs. Prior **P(s)**=0.999 and **P(b)**= 0.001*



Likelihood vs Posteriors



likelihood
***p*(*l*|*fish class*)**

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

posterior ***P*(*fish class*| *l*)**
mass function with respect to fish class, so for each ***l***, ***P***(salmon| ***l***)+***P***(bass| ***l***) = 1

More on Posterior

posterior density (our goal)

likelihood (given)

Prior (given)

$$P(c \mid l) = \frac{P(l \mid c)\, P(c)}{P(l)}$$

*normalizing factor, often do not even need it for classification since **P(l)** does not depend on class **c**. If we do need it, from the law of total probability:*

$$P(l) = p(l \mid salmon)\,p(salmon) + p(l \mid bass)\,p(bass)$$

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

$$\overset{\text{posterior}}{P(c \mid l)} = \frac{\overset{\text{likelihood}}{P(l \mid c)}\ \overset{\text{prior}}{P(c)}}{P(l)}$$

- Do not care about $P(l)$ when maximizing $P(c|l)$

$$P(c \mid l) \overset{\text{proportional}}{\propto} P(l \mid c) P(c)$$

- If P(salmon)=P(bass) (uniform prior) MAP classifier becomes ML classifier $P(c \mid l) \propto P(l \mid c)$

- If for some observation l, $P(l|\text{salmon})=P(l|\text{bass})$, then this observation is uninformative and decision is based solely on the prior $P(c \mid l) \propto P(c)$

Justification for MAP Classifier

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

$$\boldsymbol{P}(\textit{salmon} \mid \boldsymbol{l}) \underset{\textit{bass}<}{\overset{>\ \textit{salmon}}{?}} \boldsymbol{P}(\textit{bass} \mid \boldsymbol{l})$$

- For any particular ***l***, probability of error

$$\boldsymbol{Pr}[\text{error} \mid \boldsymbol{l}\,] = \begin{cases} \boldsymbol{P}(\text{bass} \mid \boldsymbol{l}) & \text{if we decide salmon} \\ \boldsymbol{P}(\text{salmon} \mid \boldsymbol{l}) & \text{if we decide bass} \end{cases}$$

Thus MAP classifier is optimal for each individual ***l*** !

Justification for MAP Classifier

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

$$Pr\left[error\right] = \int_{-\infty}^{\infty} p\left(error, l\right) dl = \int_{-\infty}^{\infty} Pr\left[error \mid l\right] p\left(l\right) dl$$

- If ***Pr***[error| ***l***]is as small as possible, the integral is small as possible

- But Bayes rule makes ***Pr***[error| ***l***] 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 $\boldsymbol{x} = [x_1, x_2, \ldots, x_d]$
  - Have more than 2 classes $\{c_1, c_2, \ldots, c_m\}$

More General Case

- As before, for each ***j*** we have
  - $p(x \mid c_j)$ is likelihood of observation ***x*** given that the true class is c_j
  - $P(c_j)$ is prior probability of class c_j
  - $P(c_j \mid x)$ is posterior probability of class c_j given that we observed data ***x***
- Evidence, or probability density for data

$$p(x) = \sum_{j=1}^{m} p(x \mid c_j) P(c_j)$$

Minimum Error Rate Classification

- Want to minimize average probability of error

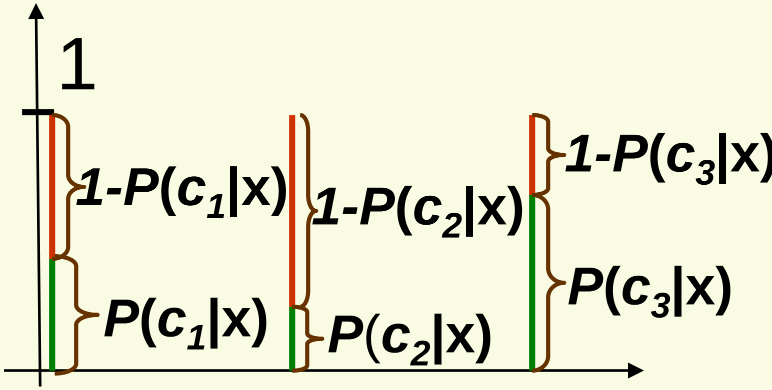
$$Pr[error] = \int p(error, x)dx = \int Pr[error \mid x]p(x)dx$$

need to make this
as small as possible

- $Pr[error \mid x] = 1 - P(c_i \mid x)$ if we decide class c_i
- $Pr[error \mid x]$ is minimized with MAP classifier
  - Decide on class c_i if

$$P(c_i \mid x) > P(c_j \mid x) \quad \forall j \neq i$$

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



General Bayesian Decision Theory

- In some cases we may want to refuse to make a decision (let human expert handle tough case)
  - allow actions $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$
- 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 $\lambda(\alpha_i \mid c_j)$ describing loss occurred when taking action α_i when the true class is c_j

Conditional Risk

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

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

Conditional Risk

sum over disjoint events (different classes)

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

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

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 c_i

$$\lambda(\alpha_i \mid c_j) = \begin{cases} 0 & \text{if } \ i = j \quad \text{(no mistake)} \\ 1 & \text{otherwise} \quad \text{(mistake)} \end{cases}$$

$$R(\alpha_i \mid x) = \sum_{j=1}^{m} \lambda(\alpha_i \mid c_j) P(c_j \mid x) = \sum_{i \neq j} P(c_j \mid x) =$$

$$= 1 - P(c_i \mid x) = Pr[\text{error if decide } c_i]$$

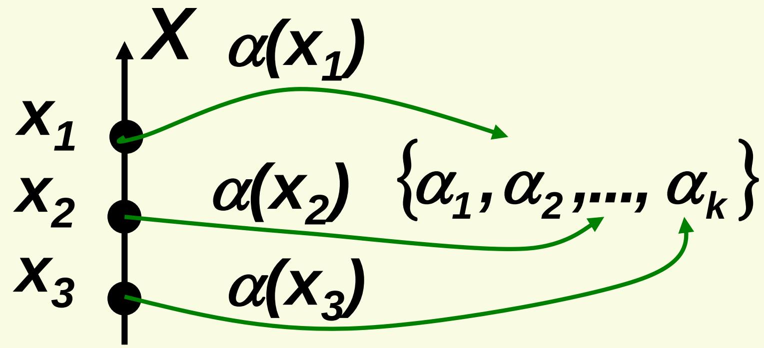
- Thus MAP classifier optimizes $R(\alpha_i|x)$

$$P(c_i \mid x) > P(c_j \mid x) \quad \forall j \neq i$$

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

Overall Risk

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



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

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

need to make this as small as possible

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

$$R(\alpha_i \mid \mathbf{x}) = \sum_{j=1}^{m} \lambda(\alpha_i \mid c_j) P(c_j \mid \mathbf{x})$$

- 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

$$\lambda_{sb} = \lambda(salmon \mid bass) = 2 \quad \text{classify bass as salmon}$$

$$\lambda_{bs} = \lambda(bass \mid salmon) = 1 \quad \text{classify salmon as bass}$$

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

- Likelihoods $p(l \mid salmon) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}}$ $p(l \mid bass) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{2*4}}$
- Priors P(salmon)= P(bass)
- Risk $R(\alpha \mid x) = \sum_{j=1}^{m} \lambda(\alpha \mid c_j) P(c_j \mid x) = \lambda_{\alpha s} P(s \mid l) + \lambda_{\alpha b} P(b \mid l)$

$$R(salmon \mid l) = \lambda_{ss} P(s \mid l) + \lambda_{sb} P(b \mid l) = \lambda_{sb} P(b \mid l)$$

$$R(bass \mid l) = \lambda_{bs} P(s \mid l) + \lambda_{bb} P(b \mid l) = \lambda_{bs} P(s \mid l)$$

Bayes Risk: Example

$$R(salmon \mid l) = \lambda_{sb} P(b \mid l) \qquad R(bass \mid l) = \lambda_{bs} P(s \mid l)$$

- Bayes decision rule (optimal for our loss function)

$$\lambda_{sb} P(b \mid l) \underset{>\ \textit{bass}}{\overset{<\ \textit{salmon}}{?}} \lambda_{bs} P(s \mid l)$$

- Need to solve $\dfrac{P(b \mid l)}{P(s \mid l)} < \dfrac{\lambda_{bs}}{\lambda_{sb}}$

- Or, equivalently, since priors are equal:

$$\frac{P(l \mid b) P(b) p(l)}{p(l) P(l \mid s) P(s)} = \frac{P(l \mid b)}{P(l \mid s)} < \frac{\lambda_{bs}}{\lambda_{sb}}$$

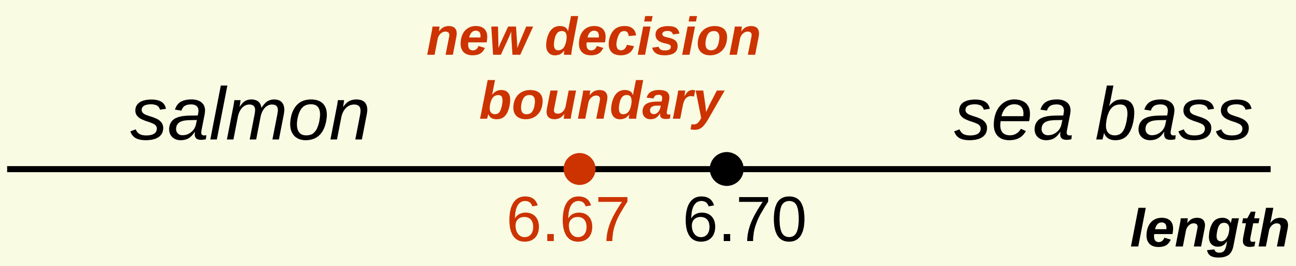
Bayes Risk: Example

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

- Substituting likelihoods and losses

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

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



Likelihood Ratio Rule

- In 2 category case, use likelihood ratio rule

$$\frac{P(x\,|\,c_1)}{P(x\,|\,c_2)} > \frac{\lambda_{12}-\lambda_{22}}{\lambda_{21}-\lambda_{11}}\frac{P(c_2)}{P(c_1)}$$

likelihood ratio

fixed number Independent of x

- If above inequality holds, decide c_1
- Otherwise decide c_2

Discriminant Functions

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

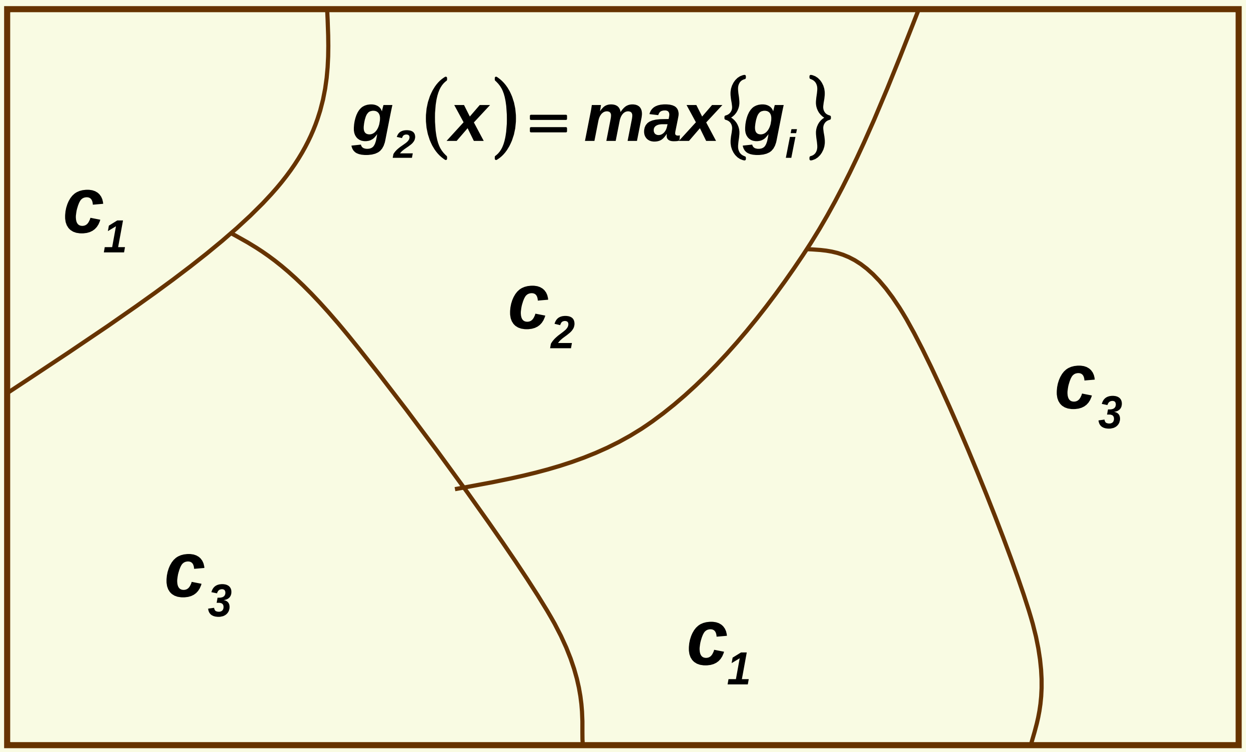
$$\boldsymbol{g}_i(\boldsymbol{x}) > \boldsymbol{g}_j(\boldsymbol{x}) \quad \forall \boldsymbol{j} \neq \boldsymbol{i}$$

discriminant function

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

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