CS434a/541a: Pattern Recognition
Prof. Olga Veksler

Lecture 3

Announcements

- Link to error data in the book
- Reading assignment
- Assignment 1 handed out, due Oct. 4
- Please send me an email with your name and course number in the subject, and in the body:
  - if you are graduate or undergraduate
  - your department
  - the year you are in if undergraduate

Announcements

Information Session on
"HOW TO APPLY FOR EXTERNAL AWARDS"
Location: Social Science Centre, Room 2036

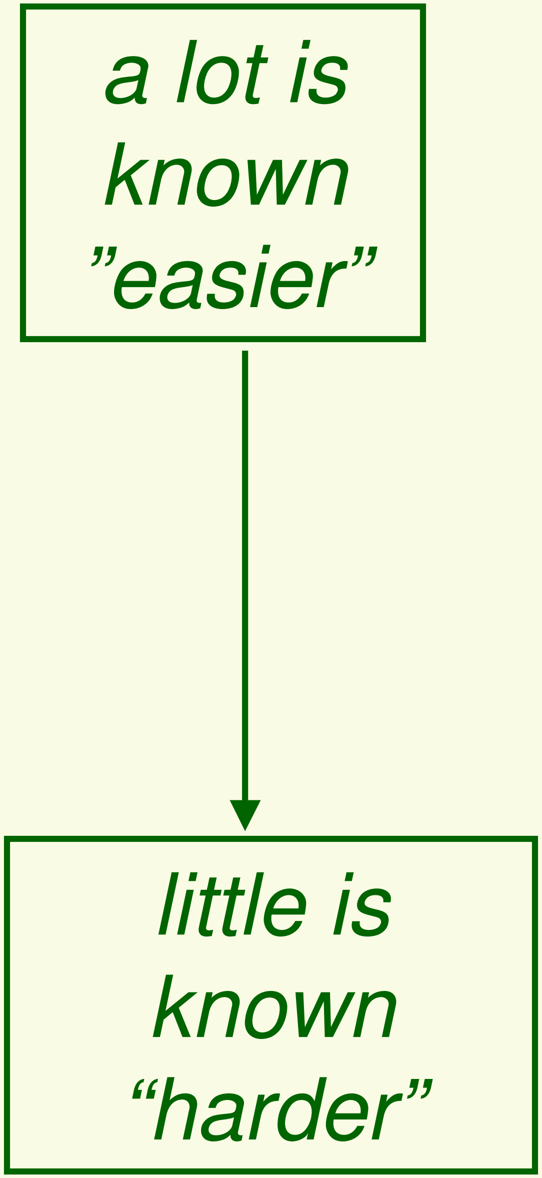
Tuesday, September 21, 2004 from 4 pm to 5pm

Hanan Lutfiyya is
holding an information session on graduate school next Monday (Sept. 27) at 6:00 in MC 320

Today

- Finish Matlab Introduction
- Course Roadmap
- Change in Notation (for consistency with textbook)
- Conditional distributions (forgot to review)
- Bayesian Decision Theory
  - Two category classification
  - Multiple category classification
  - Discriminant Functions

Course Road Map



a lot is known ”easier”

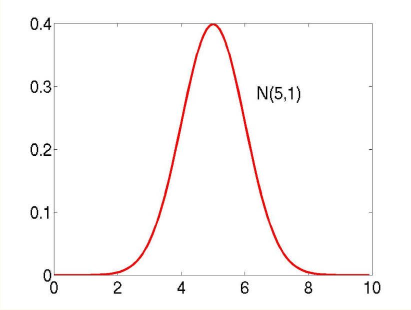
little is known “harder”

Bayesian Decision theory

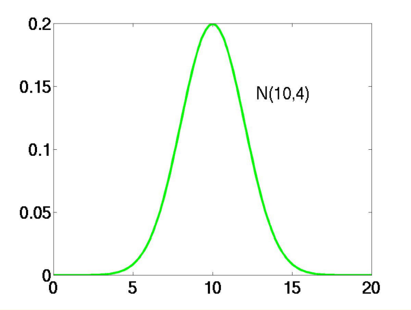
- Know probability distribution of the categories
  - never happens in real world
- Do not even need training data
- Can design optimal classifier

Example

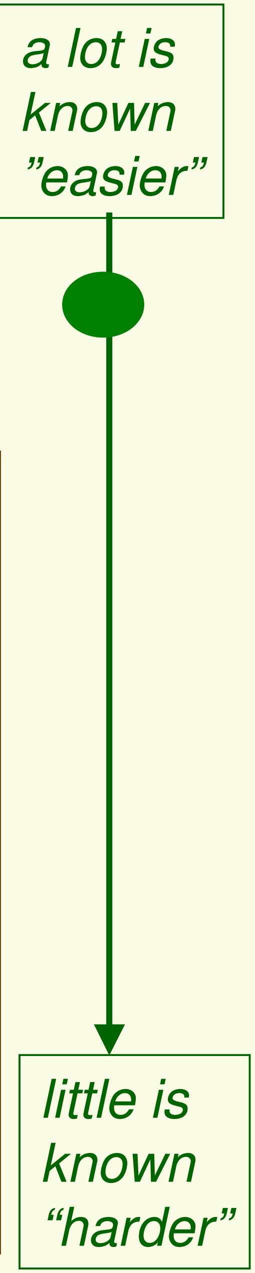
respected fish expert says that salmon’s length has distribution ***N***(5,1) and sea bass’s length has distribution ***N***(10,4)



salmon



sea bass



a lot is known ”easier”

little is known “harder”

ML and Bayesian parameter estimation

- Shape of probability distribution is known
  - Happens sometimes
- Labeled training data

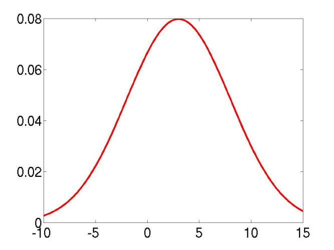
- Need to estimate parameters of probability distribution from the training data

a lot is known ”easier”

little is known “harder”

Example

respected fish expert says salmon’s length has distribution $N(\mu_1, \sigma_1^2)$ and sea bass’s length has distribution $N(\mu_2, \sigma_2^2)$

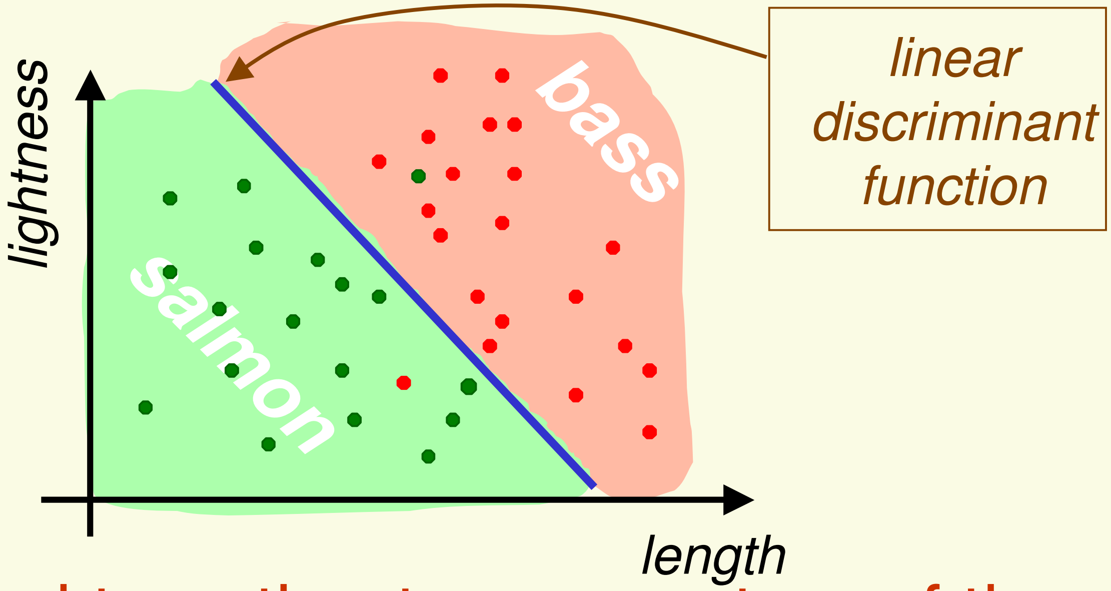


- Need to estimate parameters $\mu_1, \sigma_1^2, \mu_2, \sigma_2^2$
- Then can use the methods from the bayesian decision theory

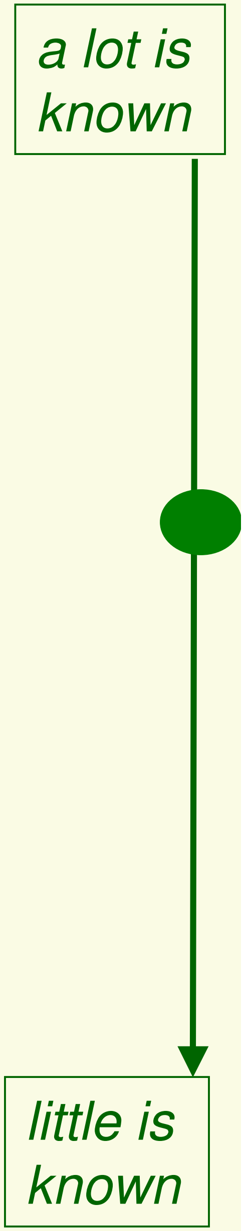
Linear discriminant functions and Neural Nets

- No probability distribution (no shape or parameters are known)
- Labeled data

- The shape of discriminant functions is known



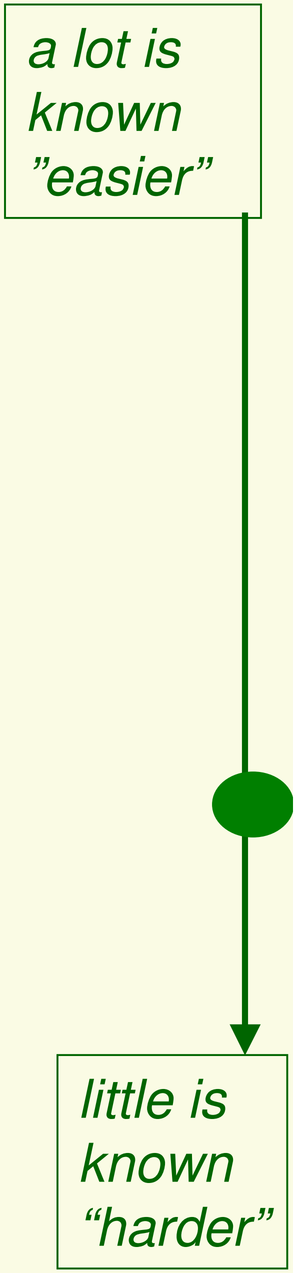
- Need to estimate parameters of the discriminant function (parameters of the line in case of linear discriminant)



Non-Parametric Methods

- Neither probability distribution nor discriminant function is known
  - Happens quite often
- All we have is labeled data

- Estimate the probability distribution from the labeled data

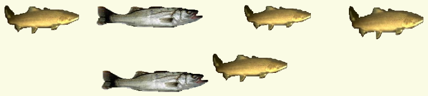


a lot is known "easier"

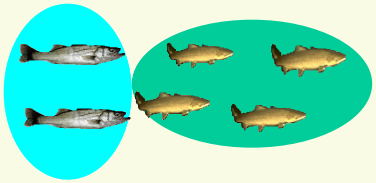
little is known "harder"

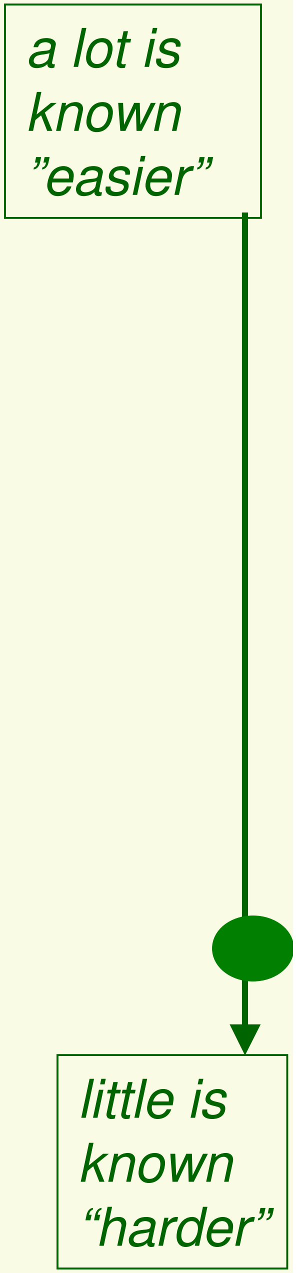
Unsupervised Learning and Clustering

- Data is *not labeled*
  - Happens quite often



1. Estimate the probability distribution from the *unlabeled* data
2. Cluster the data





a lot is known ”easier”

little is known “harder”

Course Road Map

a lot is known

1. Bayesian Decision theory (rare case)
   - Know probability distribution of the categories
   - Do not even need training data
   - Can design optimal classifier
2. ML and Bayesian parameter estimation
   - Need to estimate Parameters of probability dist.
   - Need training data
3. Non-Parametric Methods
   - No probability distribution, labeled data
4. Linear discriminant functions and Neural Nets
   - The shape of discriminant functions is known
   - Need to estimate parameters of discriminant functions
5. Unsupervised Learning and Clustering
   - No probability distribution and unlabeled data

little is known

Notation Change (for consistency with textbook)

$\boldsymbol{Pr}\,[\boldsymbol{A}]$	probability of event ***A***
$\boldsymbol{P}(\boldsymbol{x})$	probability mass function of discrete r.v. ***x***
$\boldsymbol{p}(x)$	probability density function of continuous r.v. ***x***
$\boldsymbol{p}(\boldsymbol{x},\boldsymbol{y})$	joint probability density of r.v. ***x*** and ***y***
$\boldsymbol{p}(\boldsymbol{x} \mid \boldsymbol{y})$	conditional density of ***x*** given ***y***
$\boldsymbol{P}(\boldsymbol{x} \mid \boldsymbol{y})$	conditional mass of ***x*** given *y*

More on Probability

- For events A and B, we have defined

conditional probability

$$Pr(A|B)=\frac{Pr(A\cap B)}{Pr(B)}$$

law of total probability

$$Pr(A)=\sum_{k=1}^{n}Pr(A\,|\,B_k)Pr(B_k)$$

Bayes' rule

$$Pr(B_i\,|\,A)=\frac{Pr(A\,|\,B_i)Pr(B_i)}{\sum_{k=1}^{n}Pr(A\,|\,B_k)Pr(B_k)}$$

- Usually model with random variables not events. Need equivalents of these laws for mass and density functions (could go from random variables back to events, but time consuming)

Conditional Mass Function: Discrete RV

- For discrete RV nothing new because mass function is really a probability law
- Define conditional mass function of $\boldsymbol{X}$ given $\boldsymbol{Y=y}$ by

$$P(x\,/\,y)=\frac{P(x,y)}{P(y)}$$

y is fixed

- This is a probability mass function because:

$$\sum_{\forall x} P(x\,/\,y)=\frac{\sum_{\forall x} P(x,y)}{P(y)}=\frac{P(y)}{P(y)}=1$$

- This is really nothing new because:

$$P(x\mid y)=\frac{P(x,y)}{P(y)}=\frac{\Pr\left[X=x\cap Y=y\right]}{\Pr\left[Y=y\right]}=\Pr\left[X=x\mid Y=y\right]$$

Conditional Mass Function: Bayes Rule

- The law of Total Probability:

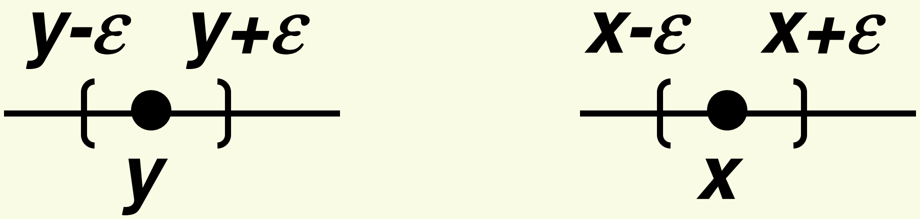
$$P(x)=\sum_{\forall y}P(x,y)=\sum_{\forall y}P(x/y)P(y)$$

- The Bayes Rule:

$$P(y/x)=\frac{P(y,x)}{P(x)}=\frac{P(x/y)P(y)}{\sum_{\forall y}P(x/y)P(y)}$$

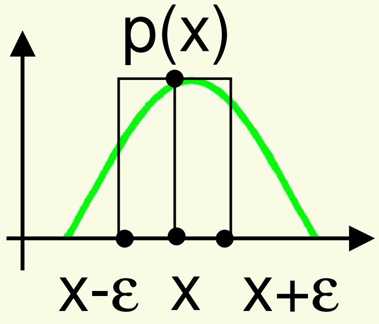
Conditional Density Function: Continuous RV

- Does it make sense to talk about conditional density p($\boldsymbol{x}$|$\boldsymbol{y}$) if $\boldsymbol{Y}$ is a continuous random variable? After all, $\boldsymbol{Pr}[\boldsymbol{Y}=\boldsymbol{y}]=0$, so we will never see $\boldsymbol{Y}=\boldsymbol{y}$ in practice
- Measurements have limited accuracy. Can interpret observation $\boldsymbol{y}$ as observation in interval [$\boldsymbol{y}-\varepsilon$, $\boldsymbol{y}+\varepsilon$], and observation x as observation in interval [$\boldsymbol{x}-\varepsilon$, $\boldsymbol{x}+\varepsilon$]



Conditional Density Function: Continuous RV

- Let B(x) denote interval $[x\text{-}\varepsilon, x+\varepsilon]$



$$Pr[X \in B(x)] = \int_{x-\varepsilon}^{x+\varepsilon} p(x)dx \approx 2\varepsilon\ p(x)$$

- Similarly $Pr[Y \in B(y)] \approx 2\varepsilon\ p(y)$

$$Pr[X \in B(x) \cap Y \in B(y)] \approx 4\varepsilon^2\ p(x,y)$$

- Thus we should have $p(x/y) \approx \dfrac{Pr[X \in B(x) / Y \in B(y)]}{2\varepsilon}$

- Which can be simplified to:

$$p(x/y) \approx \frac{Pr[X \in B(x) \cap Y \in B(y)]}{2\varepsilon\ Pr[Y \in B(y)]} \approx \frac{p(x,y)}{p(y)}$$

Conditional Density Function: Continuous RV

- Define conditional density function of ***X*** given ***Y*=*y*** by

$$p(x / y) = \frac{p(x,y)}{p(y)}$$

y is fixed

- This is a probability density function because:

$$\int_{-\infty}^{\infty} p(x / y)dx = \int_{-\infty}^{\infty} \frac{p(x,y)}{p(y)}dx = \frac{\int_{-\infty}^{\infty} p(x,y)dx}{p(y)} = \frac{p(y)}{p(y)} = 1$$

- The law of Total Probability:

$$p(x) = \int_{-\infty}^{\infty} p(x,y)\,dy = \int_{-\infty}^{\infty} p(x \mid y)p(y)\,dy$$

Conditional Density Function: Bayes Rule

- The Bayes Rule:

$$p(y \mid x) = \frac{p(y, x)}{p(x)} = \frac{p(x \mid y)p(y)}{\int_{-\infty}^{\infty} p(x \mid y)p(y)dy}$$

Mixed Discrete and Continuous

- X discrete, Y continuous
  - Bayes rule

$$P(x \mid y) = \frac{p(y \mid x)P(x)}{p(y)}$$

- X continuous, Y discrete
  - Bayes rule

$$p(x \mid y) = \frac{P(y \mid x)p(x)}{P(y)}$$

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

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
    - ***Pr***[large ears | dog] = 0.9 > 0.2= ***Pr***[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 $\boldsymbol{N}(5,1)$
  - Sea bass’s length has distribution $\boldsymbol{N}(10,4)$
- Recall if r.v. is $N(\mu,\sigma^2)$ then it’s density is

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

- Thus *class conditional* densities are

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

Likelihood function

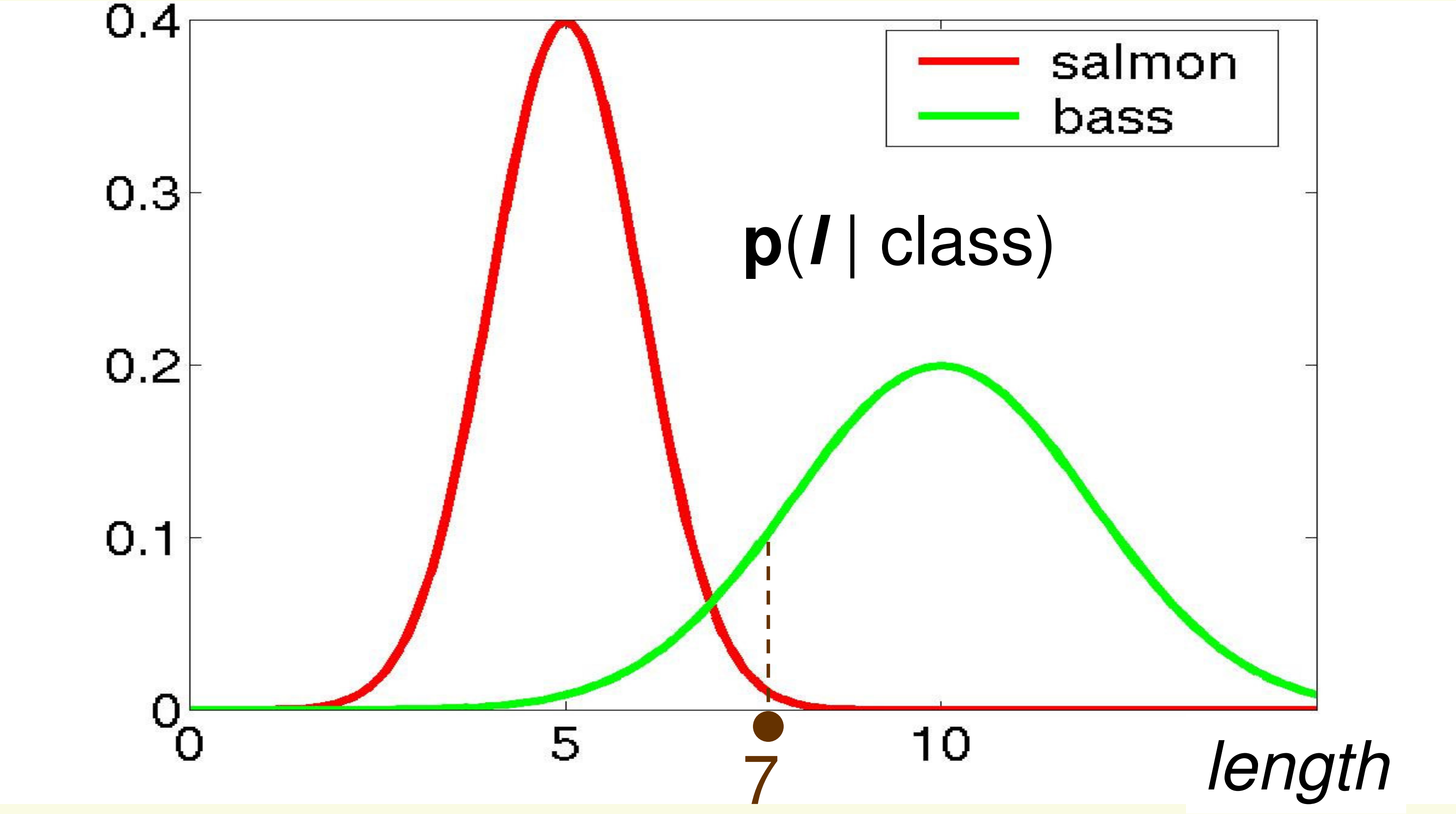
- Thus *class conditional densities* are

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

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

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

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

$$Pr[length = 7 \,|\, salmon] > Pr[length = 7 \,|\, bass]$$

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

$$Pr[length = 7 \,|\, salmon] = Pr[length = 7 \,|\, bass] = 0$$

- Instead, we choose class which maximizes likelihood

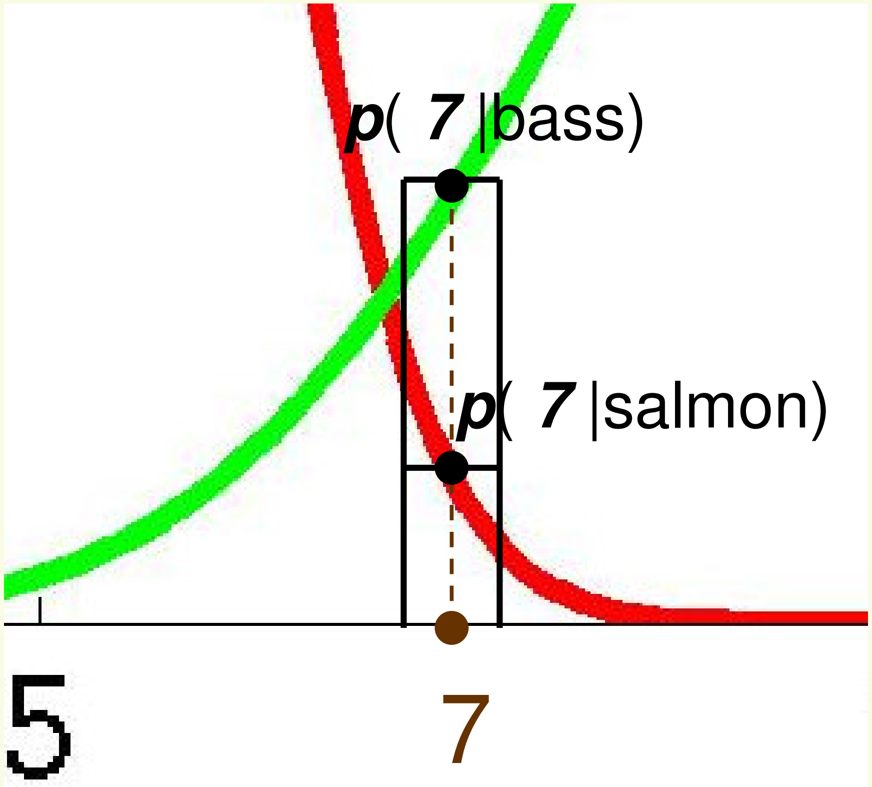
$$p(l \,|\, salmon) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}} \qquad p(l \,|\, bass) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{2*4}}$$

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

$$p(l \,|\, salmon) \underset{> \;\; salmon}{\overset{bass \;\; <}{?}} p(l \,|\, bass)$$

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

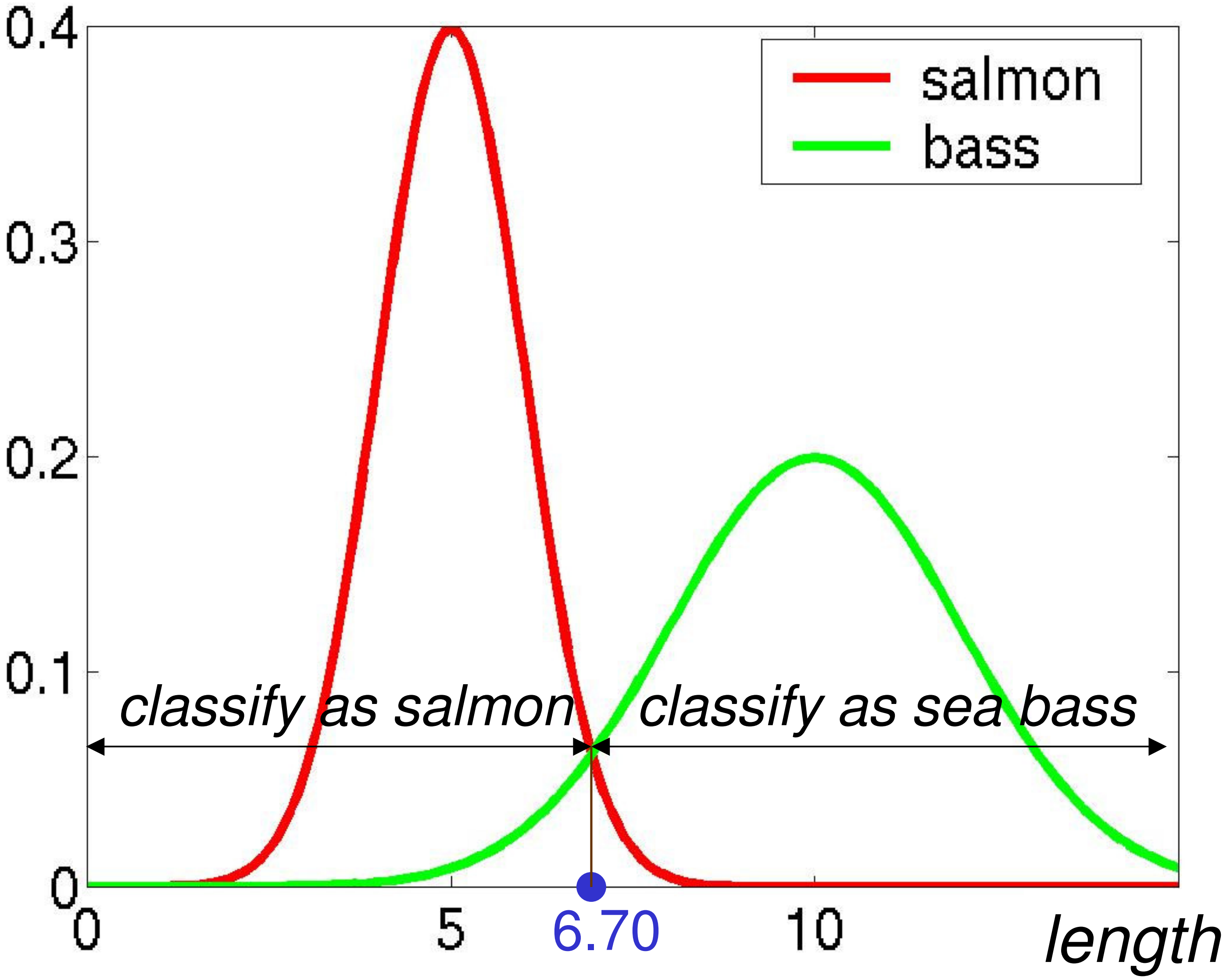
Interval Justification



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

$$\begin{array}{ccc} Pr[I \in B(7) \,/\, bass] \approx 2\varepsilon & & p(7 \,/\, bass) \\ \vee & \Leftarrow & \vee \\ Pr[I \in B(7) \,/\, salmon] \approx 2\varepsilon & & p(7 \,/\, salmon) \end{array}$$

Decision Boundary

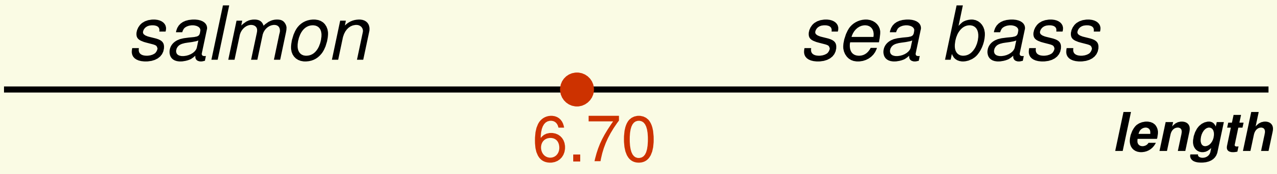


Priors

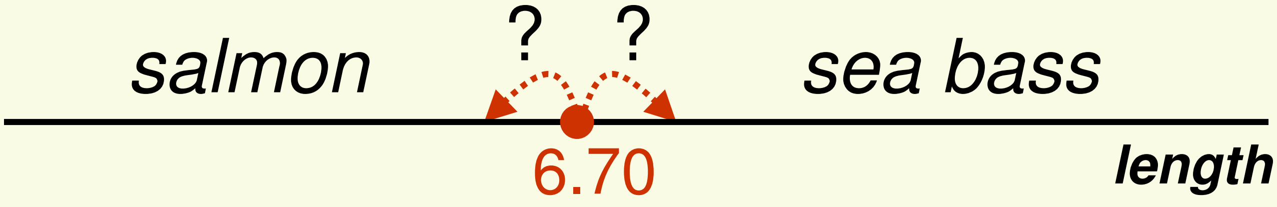
- Prior comes from prior knowledge, no data has been seen yet
- Suppose a fish expert says: in the fall, there are twice as many salmon as sea bass
- Prior for our fish sorting problem
  - ***P***(salmon) = 2/3
  - ***P***(bass) = 1/3
- With the addition of prior to our model, how should we classify a fish of length 7?

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

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

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

- Similarly:

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

MAP (maximum a posteriori) classifier

$$P(salmon \mid length) \underset{bass\, <}{\overset{>\, salmon}{?}} P(bass \mid length)$$

$$\frac{p(length \mid salmon)P(salmon)}{p(length)} \underset{bass\, <}{\overset{salmon\; >}{?}} \frac{p(length \mid bass)P(bass)}{p(length)}$$

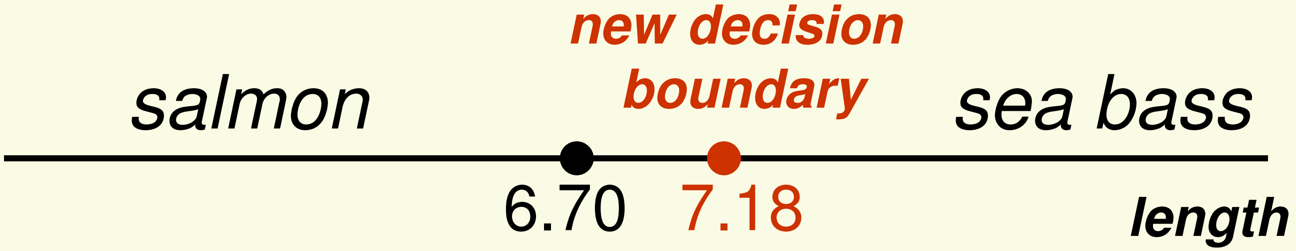
$$p(length \mid salmon)P(salmon) \underset{bass\, <}{\overset{>\, salmon}{?}} p(length \mid bass)P(bass)$$

Back to Fish Sorting Example

- likelihood

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

- Priors: P(salmon) = 2/3, 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

salmon

sea bass

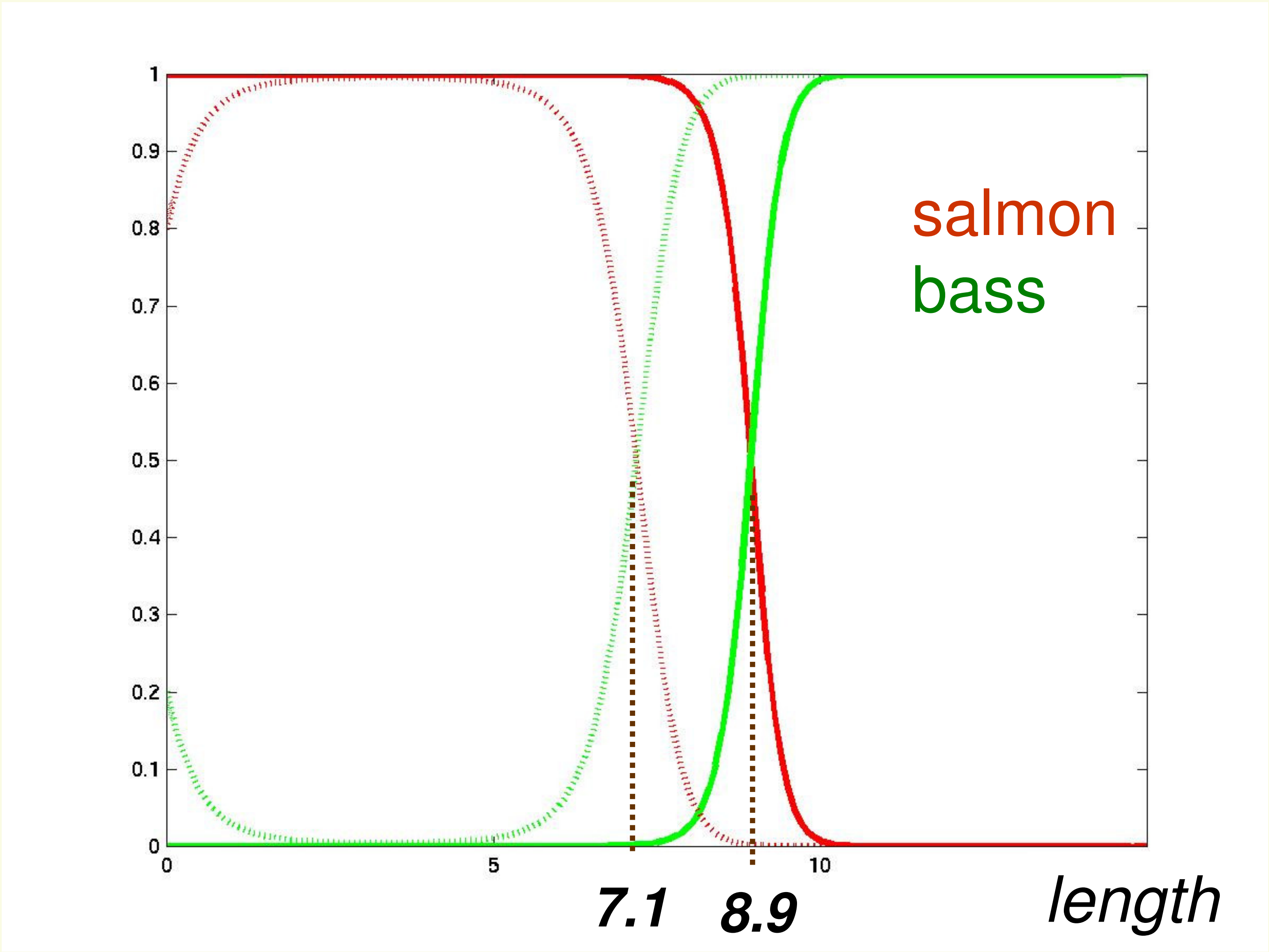
6.70

7.18

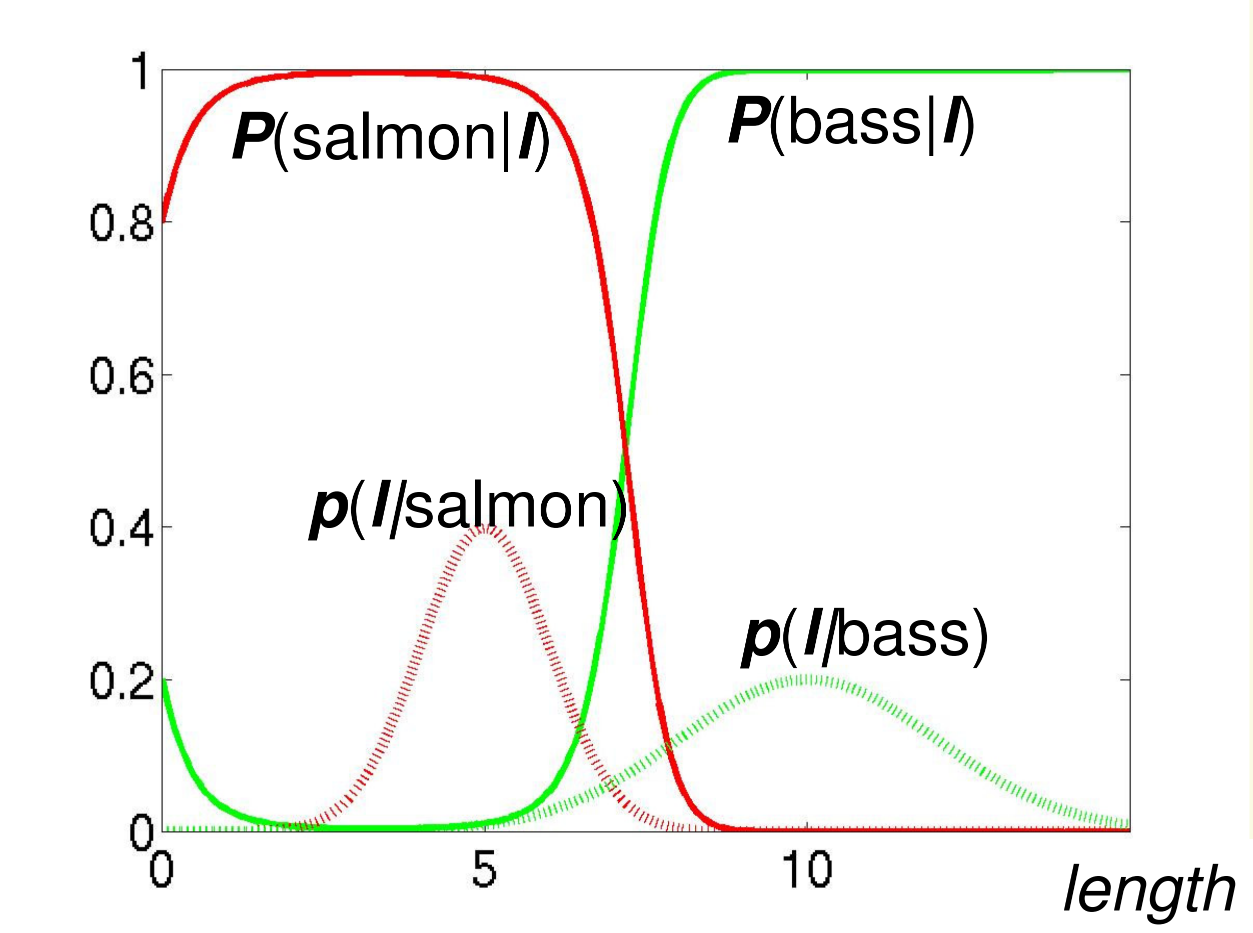
length

- 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 Posterior

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

cause (class) $\boldsymbol{c} \Longrightarrow \boldsymbol{l}$ *effect (length)*

- If cause ***c*** is present, it easy to determine the probability of effect ***l*** with likelihood ***P***(***l***|***c***)
- Usually observe the effect ***l*** without knowing cause ***c***. Hard to determine cause ***c*** because there may be several causes which could produce same effect ***l***
- Bayes rule makes ***l*** easy to determine posterior ***P***(***c***|***l***), if we know likelihood ***P***(***l***|***c***) and prior ***P***(***c***)

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
- However prior alone is not enough, we still need likelihood
  - ***P***(salmon)=2/3, ***P***(sea bass)=1/3
  - If I don't let you see the data, but ask you to guess, will you choose salmon or sea bass?

More on Map Classifier

$$\overset{\text{posterior}}{P(c/I)} = \frac{\overset{\text{likelihood}}{P(I/c)}\;\;\overset{\text{prior}}{P(c)}}{P(I)}$$

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

$$P(c/I) \overset{\text{proportional}}{\propto} P(I/c)P(c)$$

- If ***P***(salmon)=***P***(bass) (uniform prior) MAP classifier becomes ML classifier $P(c/I) \propto P(I/c)$

- 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 $P(c/I) \propto P(c)$

Justification for MAP Classifier

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

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

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

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

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

Justification for MAP Classifier

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

$$\boldsymbol{Pr[error] = \int_{-\infty}^{\infty} p(error, l)\,dl = \int_{-\infty}^{\infty} Pr[error \,|\, l]\,p(l)\,dl}$$

- 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!

Today

- Bayesian Decision theory
  - Multiple Classes
  - General loss functions
- Multivariate Normal Random Variable
  - Classifiers
  - Discriminant Functions

More General Case

- Let’s generalize a little bit
  - Have more than one feature $\boldsymbol{x} = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_d}]$
  - Have more than 2 classes $\{\boldsymbol{c_1}, \boldsymbol{c_2}, \ldots, \boldsymbol{c_m}\}$

More General Case

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

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

Minimum Error Rate Classification

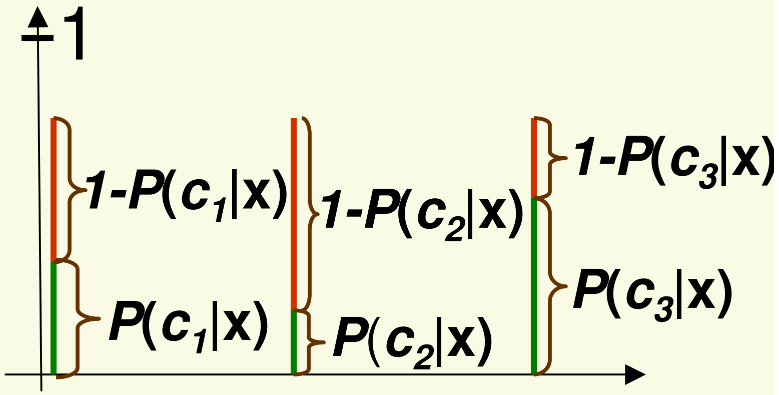
- Want to minimize average probability of error

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

need to make this as small as possible

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

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



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 $\{\alpha_1, \alpha_2, \ldots, \alpha_k\}$
- 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 $\lambda(\alpha_i / \boldsymbol{c}_j)$ describing loss occurred when taking action α_i when the true class is $\boldsymbol{c}_j$

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 $\lambda(\alpha_i / \boldsymbol{c}_j)$
- Probability that the true class is $\boldsymbol{c}_j$ after observing $\boldsymbol{x}$ is $\boldsymbol{P}(\boldsymbol{c}_j / \boldsymbol{x})$
- The expected loss associated with taking action α_i is called ***conditional risk*** and it is:

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

Conditional Risk

$$\underbrace{\mathbf{R}(\alpha_i \mid \mathbf{x})}_{\substack{\text{penalty for}\\ \text{taking action } \alpha_i\\ \text{if observe } x}} = \overset{\substack{\text{sum over disjoint events}\\ \text{(different classes)}\\ \downarrow}}{\sum_{j=1}^{m}} \underbrace{\lambda(\alpha_i \mid \mathbf{c}_j)\,\overset{\substack{\text{probability of}\\ \text{class } \mathbf{c}_j \text{ given}\\ \text{observation } x\\ \downarrow}}{\mathbf{P}(\mathbf{c}_j \mid \mathbf{x})}}_{\substack{\text{part of overall penalty}\\ \text{which comes from event}\\ \text{that true class is } \mathbf{c}_j}}$$

$\mathbf{c}_1$	$\lambda(\alpha_i \vert \mathbf{c}_1)$
$\mathbf{c}_2$	$\lambda(\alpha_i \vert \mathbf{c}_2)$
$\mathbf{c}_3$	$\lambda(\alpha_i \vert \mathbf{c}_3)$
$\mathbf{c}_4$	$\lambda(\alpha_i \vert \mathbf{c}_4)$

Example: Zero-One loss function

- action α_i is decision that true class is c_i

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

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

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

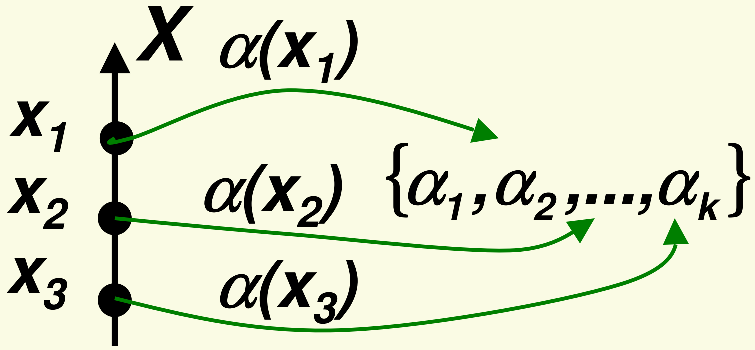
- Thus MAP classifier optimizes $R(\alpha_i|x)$

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

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

Overall Risk

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



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

$$R(\alpha)=\int R(\alpha(x)/x)p(x)dx$$

need to make this as small as possible

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

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

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

Bayes Risk: Example

- Salmon is more tasty and expensive than sea bass

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

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

$\lambda_{ss} = \lambda_{bb} = 0$ *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

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

- Bayes decision rule (optimal for our loss function)

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

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

- Or, equivalently, since priors are equal:

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

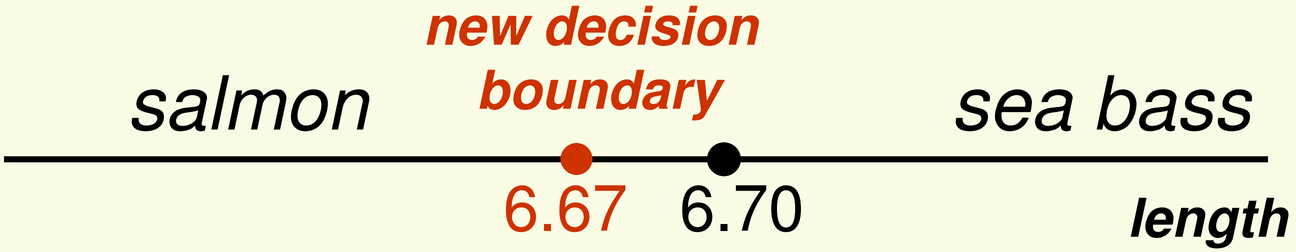
Bayes Risk: Example

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

- Substituting likelihoods and losses

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

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



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 $\boldsymbol{c}_1$
- Otherwise decide $\boldsymbol{c}_2$

Discriminant Functions

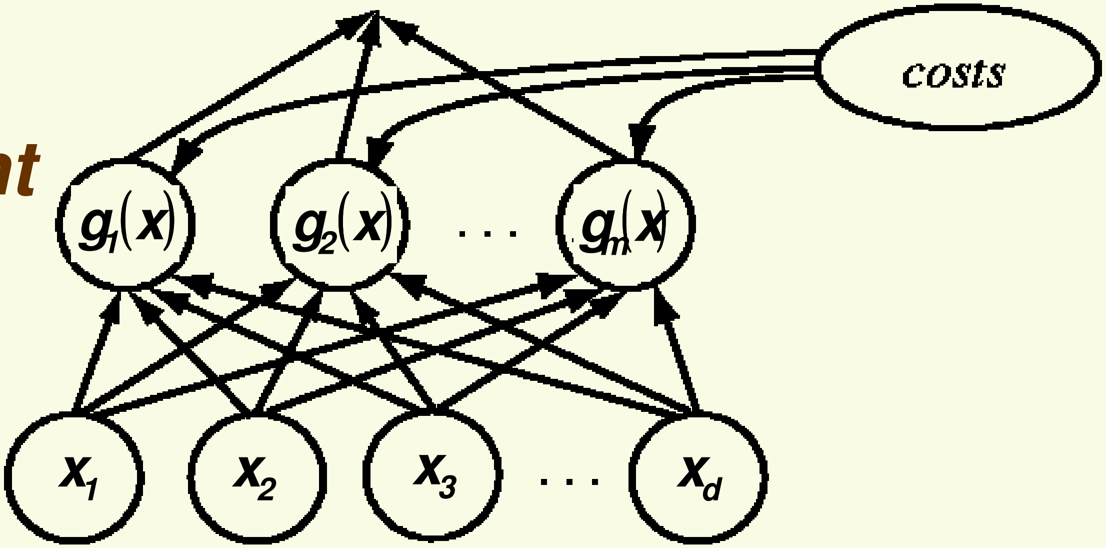
- All decision rules have the same structure: at observation $\boldsymbol{x}$ choose class $\boldsymbol{c}_i$ s.t.

$$\underbrace{\boldsymbol{g}_i(\boldsymbol{x})}_{\text{discriminant function}} > \boldsymbol{g}_j(\boldsymbol{x}) \quad \forall \boldsymbol{j} \neq \boldsymbol{i}$$

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

Discriminant Functions

- Classifier can be viewed as network which computes ***m*** discriminant functions and selects category corresponding to the largest discriminant



select class giving maximim

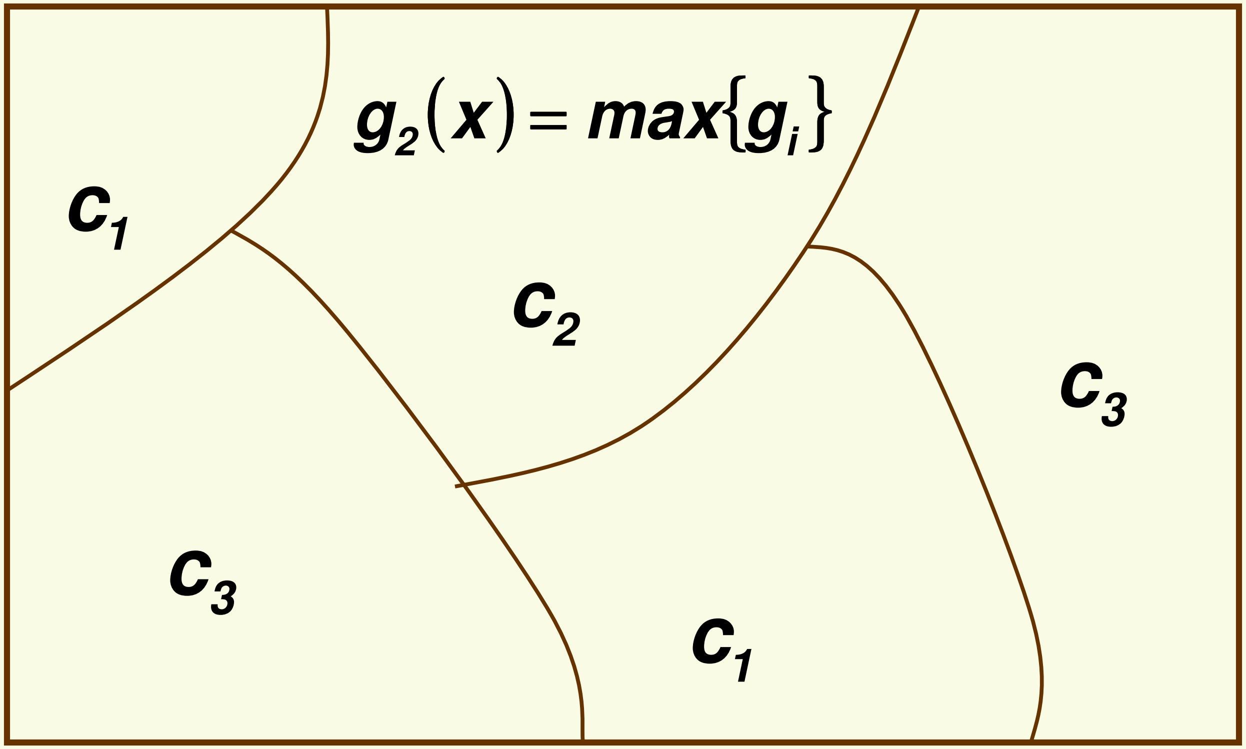
discriminant functions

features

- $g_i(x)$ can be replaced with any monotonically increasing function, the results will be unchanged

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