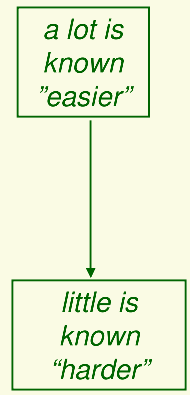
## CS434b/654b: Pattern Recognition
## Prof. Olga Veksler

Lecture 3



## Course Road Map



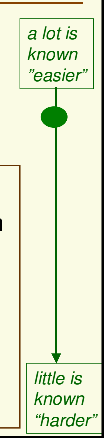

## Today

- Finish Matlab Introduction
- Course Roadmap
- Probability Topic: Conditional distributions
- Bayesian Decision Theory
  - Two category classification
  - Multiple category classification
  - Discriminant Functions



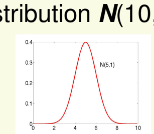
## Bayesian Decision theory

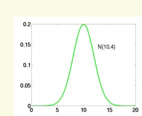
- Know probability distribution of the categories
  - never happens in real world
- Do not even need training data
- Can design optimal classifier



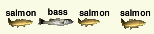

Example

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

## ML and Bayesian parameter estimation

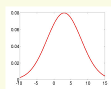
- Shape of probability distribution is known
  - Happens sometimes
- Labeled training data 

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

*a lot is known "easier"*

Example

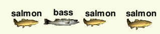
respected fish expert says salmon's length has distribution $N(\mu_1, \sigma_1^2)$ and sea bass's length has distribution $N(\mu_2, \sigma_2^2)$



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

*little is known "harder"*

## Non-Parametric Methods

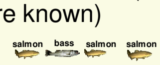
- Neither probability distribution nor discriminant function is known
  - Happens quite often
- All we have is labeled data 

- Estimate the probability distribution from the labeled data

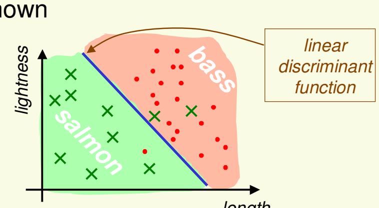
*a lot is known "easier"*

*little is known "harder"*

## Linear discriminant functions and Neural Nets

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

- The shape of discriminant functions is known



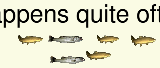

- Need to estimate parameters of the discriminant function (parameters of the line in case of linear discriminant)

*a lot is known*

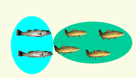
*little is known*

## 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

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

*a lot is known* → *little is known*

## Conditional Mass Function: Discrete RV

- For discrete RV nothing new because mass function is really a probability law
- Define conditional mass function of ***X*** given ***Y=y*** by

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

*y is fixed*

- This is a probability mass function because:

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

- This is really nothing new because:

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

## 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: Bayes Rule

- The law of Total Probability:

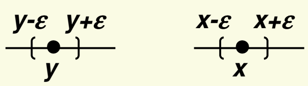
$$P(x) = \sum_{\forall y} P(x,y) = \sum_{\forall y} P(x / y) P(y)$$

- 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(***x***|***y***) if ***Y*** is a continuous random variable? After all, ***Pr***[***Y***=***y***]=0, so we will never see ***Y***=***y*** in practice
- Measurements have limited accuracy. Can interpret observation ***y*** as observation in interval [$y-\varepsilon$, $y+\varepsilon$], and observation x as observation in interval [$x-\varepsilon$, $x+\varepsilon$]




## 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:

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

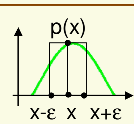
- The law of Total Probability:

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

## Conditional Density Function: Continuous RV

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

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




- 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: 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

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

## 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

## 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

18

## 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 $N(\mu,\sigma^2)$ 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

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

20

## Likelihood function

- Thus *class conditional densities* are

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

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

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

21

## ML (maximum likelihood) Classifier

- We would like to choose salmon if

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

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

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

- Instead, we choose class which maximizes likelihood

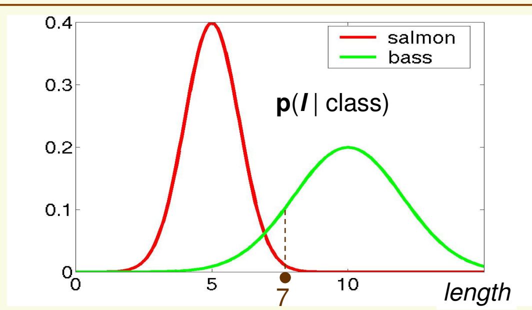
$$p(l \mid salmon) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(l-5)^2}{2}} \qquad p(l \mid bass) = \frac{1}{2\sqrt{2\pi}} e^{-\frac{(l-10)^2}{2*4}}$$

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

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

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

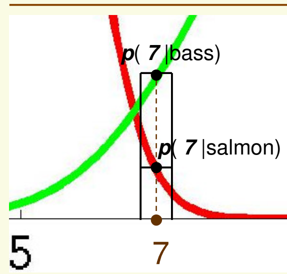
## Likelihood vs. Class Conditional Density




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

22

## Interval Justification

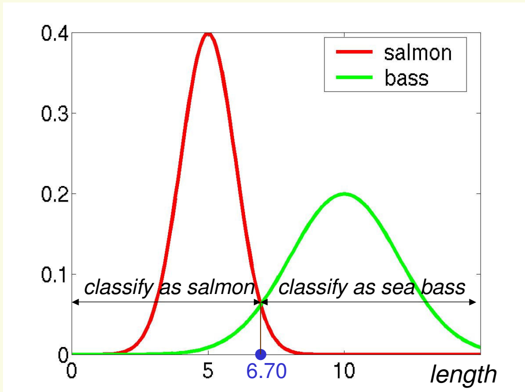



Thus we choose the class (bass) which is more likely to have given the observation

$$\begin{matrix} Pr[l \in B(7) \mid bass] \approx 2\varepsilon\, p(7 \mid bass) \\ \vee \qquad \Leftarrow \qquad \vee \\ Pr[l \in B(7) \mid salmon] \approx 2\varepsilon\, p(7 \mid salmon) \end{matrix}$$

24

## Decision Boundary



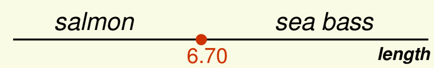

## Priors

- 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?

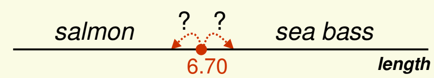
26

## How Prior Changes Decision Boundary?

- Without priors




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




27

## 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 $P(bass \mid length) > P(salmon \mid length)$

28

## 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

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

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

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



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

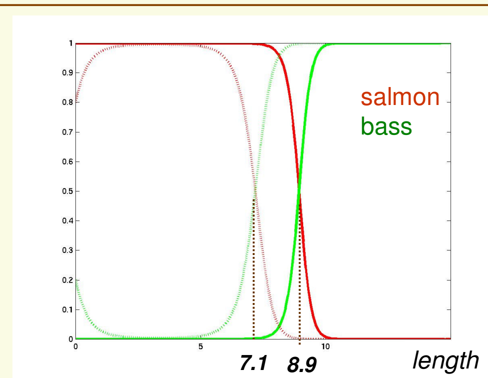
salmon — new decision boundary — sea bass

6.70 7.18 length

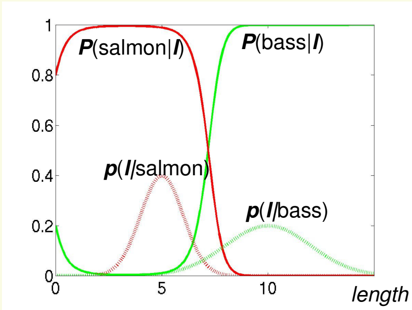
- New decision boundary makes sense since we expect to see more salmon



## 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

$$\underbrace{P(c \mid l)}_{\text{posterior}} = \frac{\overbrace{P(l \mid c)}^{\text{likelihood}}\ \overbrace{P(c)}^{\text{prior}}}{P(l)}$$

*cause (class)* ***c*** ⟹ ***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 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
- 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?

36

## More on Map Classifier

$$\overset{posterior}{P(c\,|\,l)} = \frac{\overset{likelihood}{P(l\,|\,c)}\ \overset{prior}{P(c)}}{P(l)}$$

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

$$P(c\,|\,l)\ \overset{proportional}{\propto}\ P(l\,|\,c)P(c)$$

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

## 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***

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

- 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!

## Justification for MAP Classifier

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

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

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

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

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

38

## More General Case

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

## More General Case

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

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

41

## 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 $\lambda(\alpha_i / c_j)$ describing loss occurred when taking action $\alpha_i$ when the true class is $c_j$

43

## Minimum Error Rate Classification

- Want to minimize average probability of error

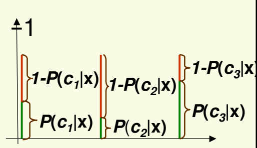
$$Pr[error] = \int p(error, x)dx = \int Pr[error / x] p(x)dx$$

*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*




## Conditional Risk

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

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

## Conditional Risk

sum over disjoint events (different classes)

probability of class $c_j$ given observation x

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

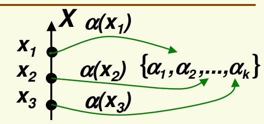
penalty for taking action $\alpha_i$ if observe x

part of overall penalty which comes from event that true class is $c_j$

| | |
|---|---|
| $c_1$ | $\lambda(\alpha_i \mid c_1)$ |
| $c_2$ | $\lambda(\alpha_i \mid c_2)$ |
| $c_3$ | $\lambda(\alpha_i \mid c_3)$ |
| $c_4$ | $\lambda(\alpha_i \mid c_4)$ |

## Overall Risk

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




- The average risk for $\alpha(x)$

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

**need to make this as small as possible**

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

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

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

## Example: Zero-One loss function

- action $\alpha_i$ 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 $R(\alpha_i \mid x)$

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

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

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

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

- Need to solve $\frac{P(b \mid l)}{P(s \mid l)} < \frac{\lambda_{bs}}{\lambda_{sb}}$
- Or, equivalently, since priors are equal:

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

## Likelihood Ratio Rule

- In 2 category case, use likelihood ratio rule

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

likelihood ratio — fixed number Independent of x

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

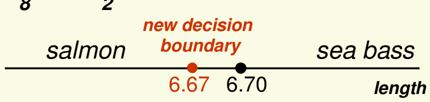
51

## Bayes Risk: Example

- Need to solve $\frac{P(l \mid b)}{P(l \mid s)} < \frac{\lambda_{bs}}{\lambda_{sb}}$
- Substituting likelihoods and losses

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

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




## Discriminant Functions

- All decision rules have the same structure: at observation $x$ choose class $c_i$ s.t.

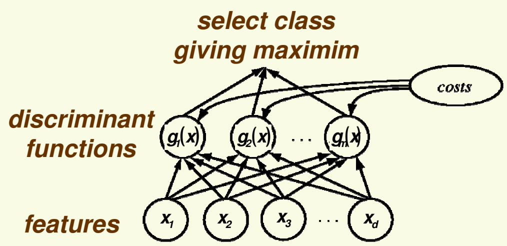
$$g_i(x) > g_j(x) \quad \forall j \neq i$$

discriminant function

- ML decision rule: $g_i(x) = P(x \mid c_i)$
- MAP decision rule: $g_i(x) = P(c_i \mid x)$
- Bayes decision rule: $g_i(x) = -R(c_i \mid x)$

## Discriminant Functions

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




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

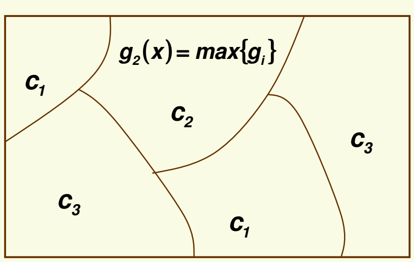
## Important Points

- 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

55

## Decision Regions

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




54