

CS 559: Machine Learning Fundamentals and Applications

1st Set of Notes

Instructor: Philippos Mordohai
Webpage: www.cs.stevens.edu/~mordohai
E-mail: Philippos.Mordohai@stevens.edu
Office: Lieb 215

Objectives

- Hands-on experience with fundamental algorithms
 - Useful for everyday problems
- Exposure to state of the art in machine learning and pattern recognition

Logistics

- Office hours: Tuesday 5-6 and by email
- Evaluation:
 - 4 homework sets (20%)
 - Project (25%)
 - Pop-up quizzes and participation (10%)
 - Midterm (20%)
 - Final exam (25%)

Project

- Pick topic around middle of the semester
- I will suggest ideas and datasets in next lectures
- Deliverables:
 - Project proposal
 - Presentation in class
 - Poster in CS department event
 - Final report (around 8 pages)

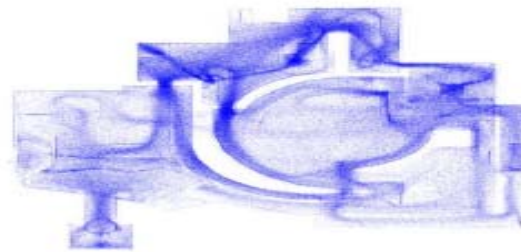
Project Examples

- Face detection using boosting

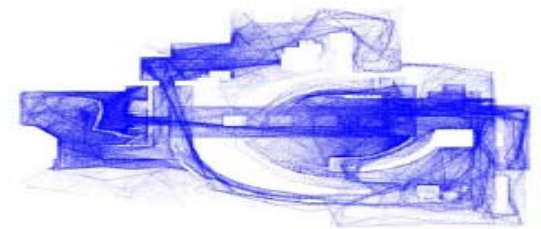


Project Examples

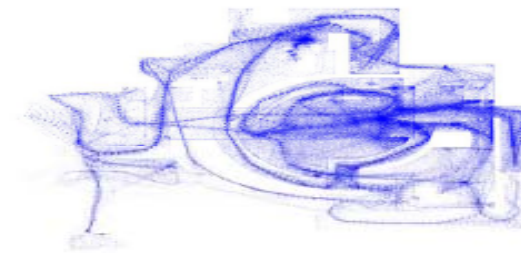
- Detecting bots in Quake 2



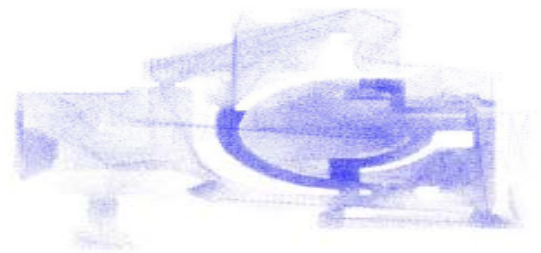
(a) Human players



(b) CR Bot



(c) Eraser Bot



(d) ICE Bot

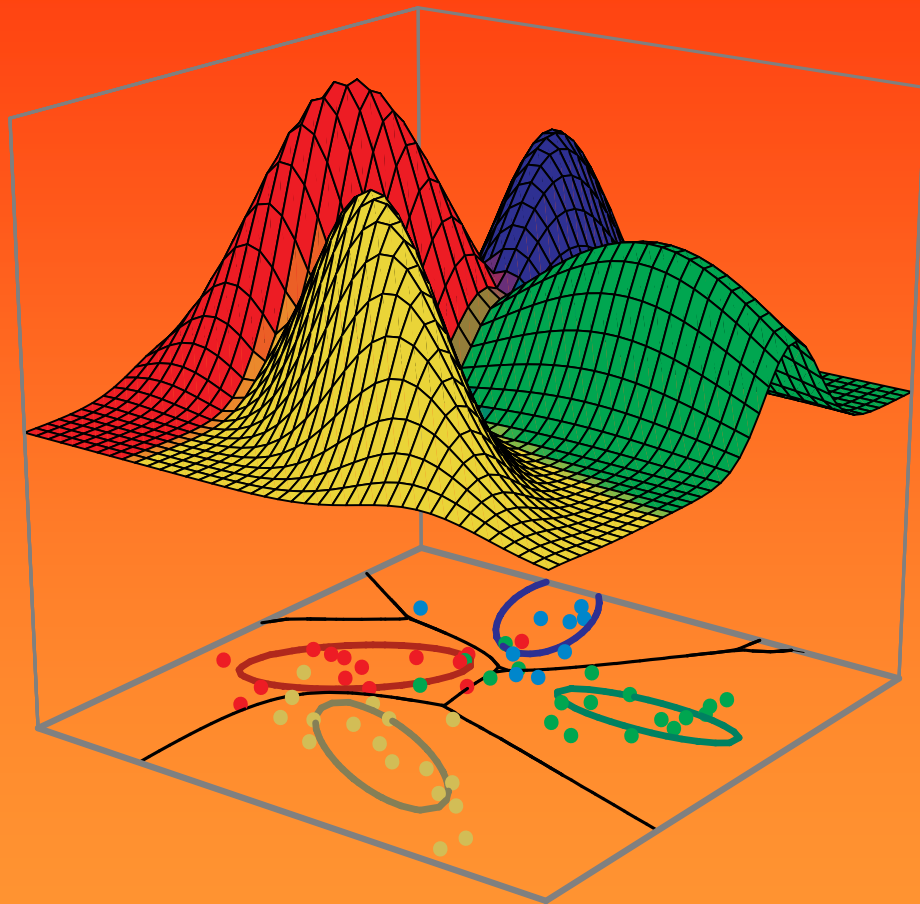
Fig. 1. Presence locations of all players

Project Examples

- Spam filtering
- Gender identification from emails
- Handwriting recognition
 - Also cool demo, but requires hardware

Prerequisites

- Probability theory
- Matlab or C/C++ programming
 - This could be “language of your choice”, but then you are responsible for debugging etc.
- Some linear algebra
 - Must not be afraid of eigenvalues
- Your grade will be affected by any weaknesses in these



Pattern Classification

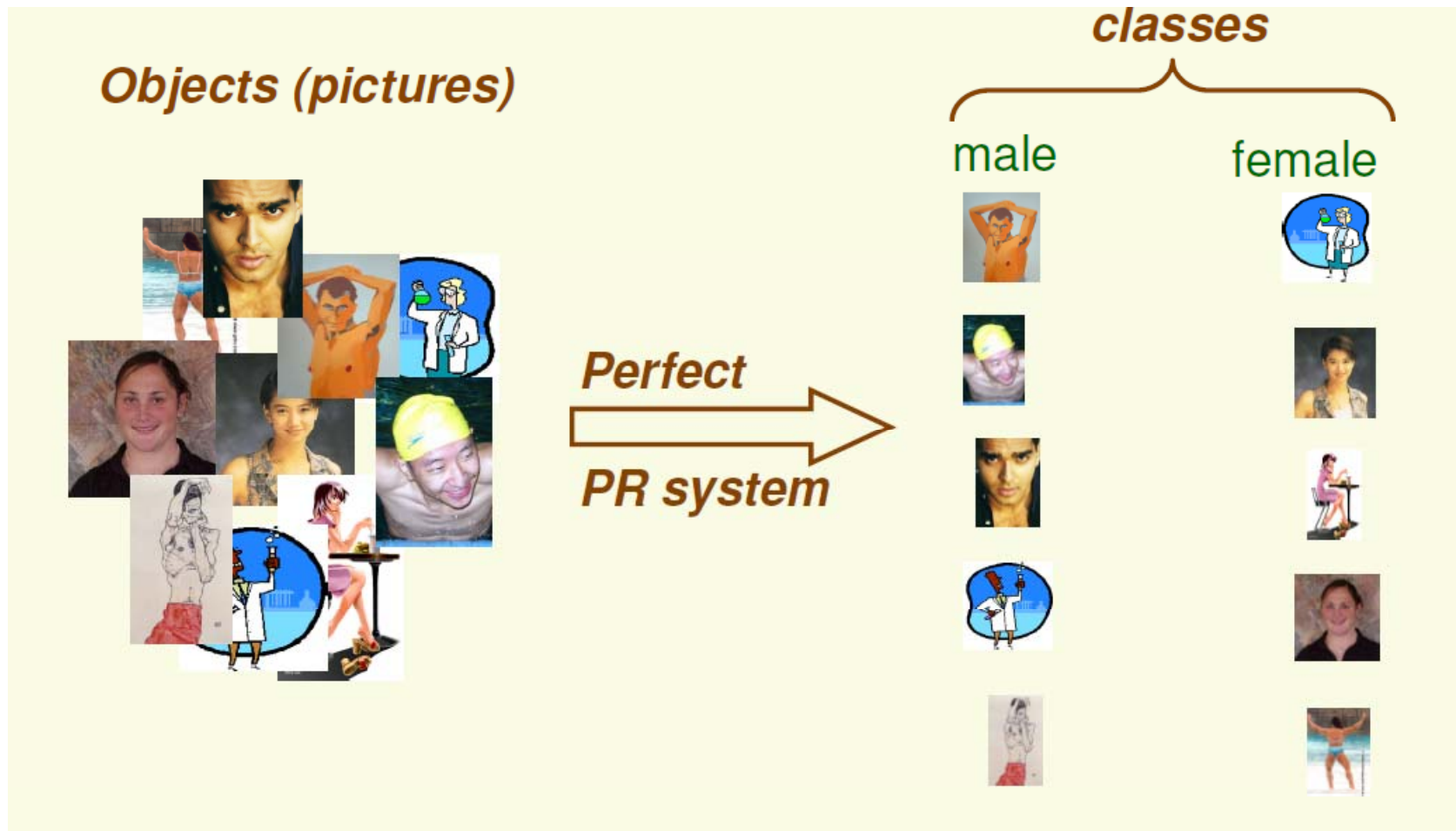
A lot of material in these slides was taken from
Pattern Classification (2nd ed) by R. O. Duda, P. E.
Hart and D. G. Stork, John Wiley & Sons, 2000
with the permission of the authors and the publisher

What is Pattern Recognition?

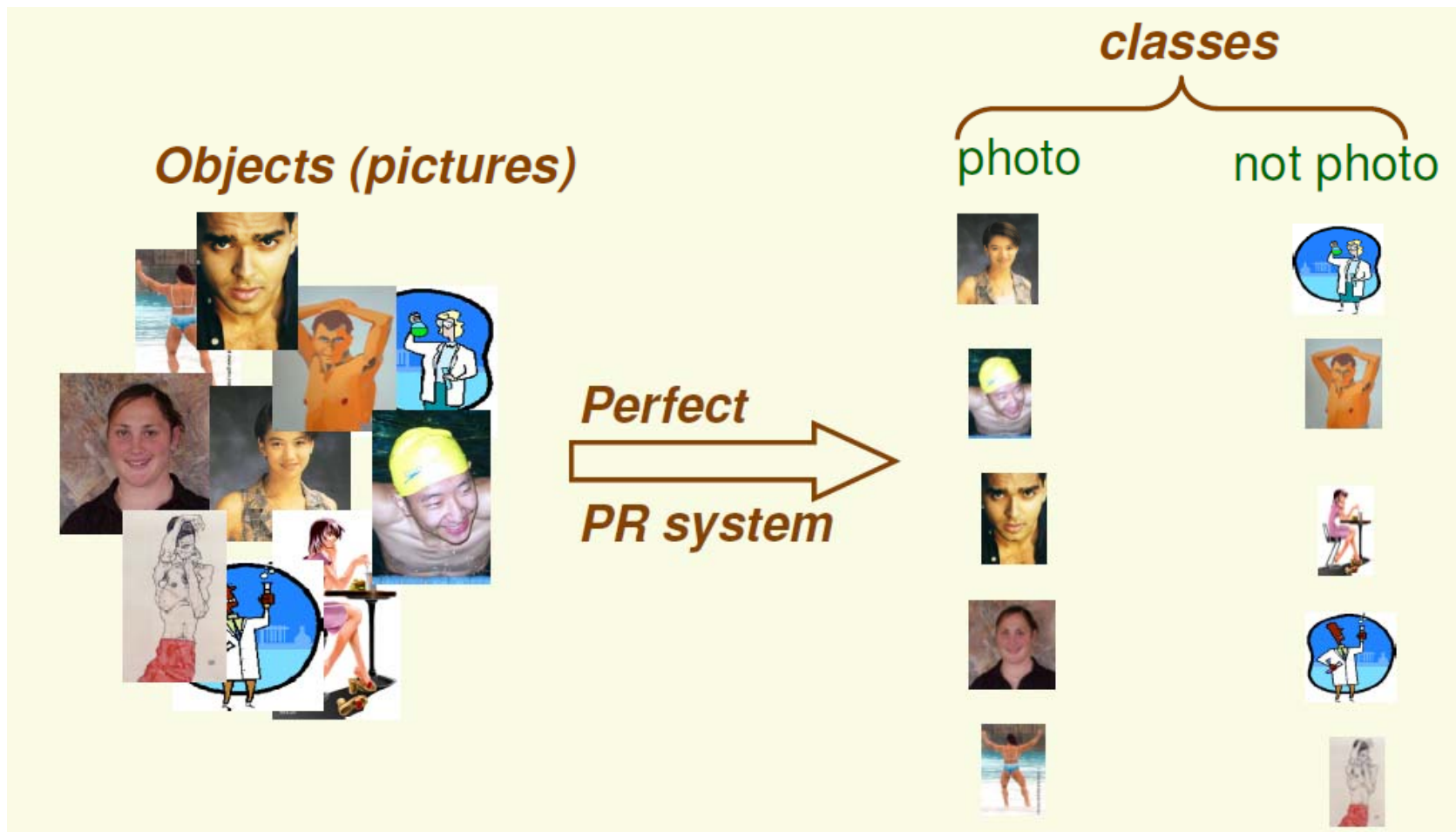
- Informally
 - Recognize **patterns** in data
- More formally
 - Assign an **object** or an **event** to one of the several pre-specified **categories** (a category is usually called a class)

Many of these slides are borrowed from Olga Veksler (U. of Western Ontario)

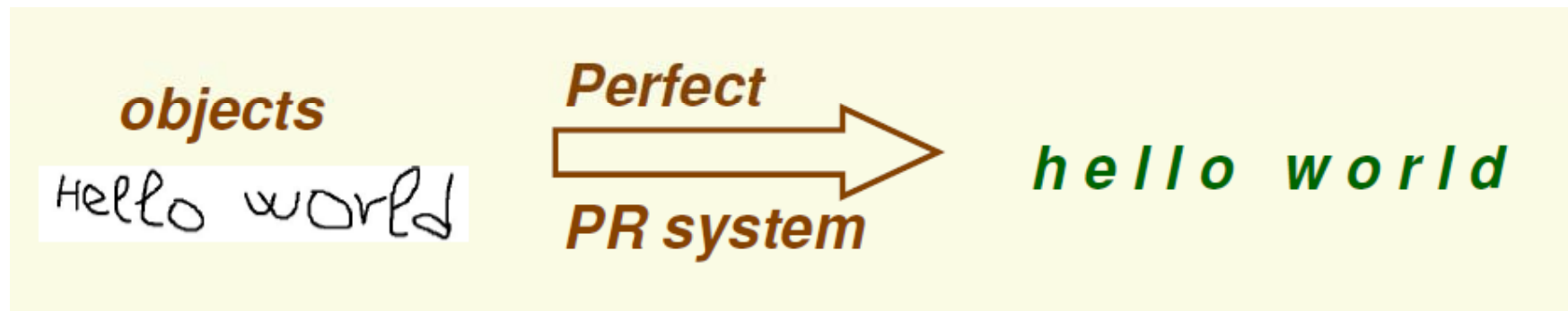
Example: Male or Female?



Example: Photograph or Not?

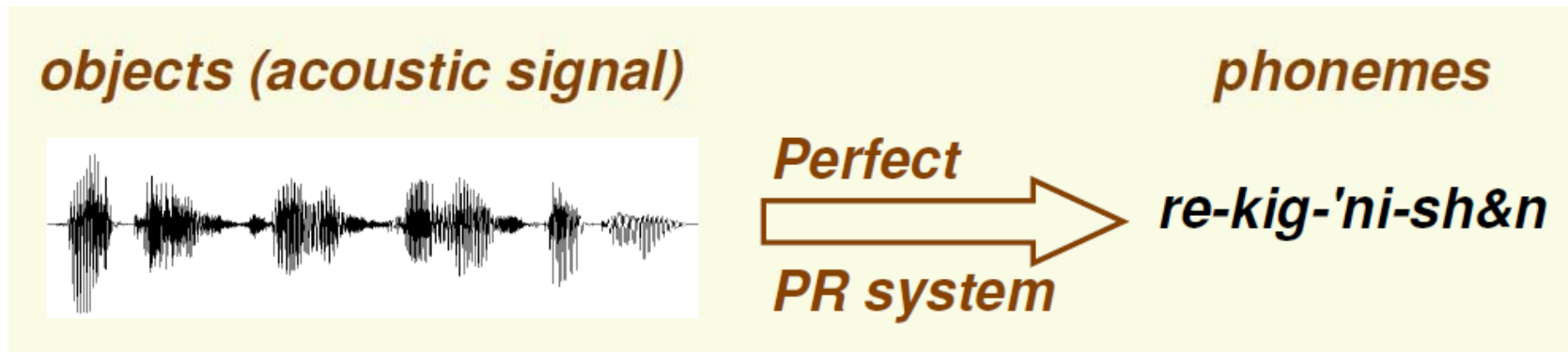


Character Recognition



- In this case the classes are all characters in the alphabet, digits etc.

Speech Understanding



- In this case the classes are all phonemes

Machine Learning Research

- Speech recognition
- Natural language processing
- Computer vision
- Medical outcomes analysis
- Robot control
- Computational biology
- Sensor networks

Real-life Applications

- Loan applications
- Recommendation systems
 - Amazon, Netflix
- Targeted advertising
 - countless examples...

Chapter 1: Introduction to Pattern Recognition (Sections 1.1-1.6)

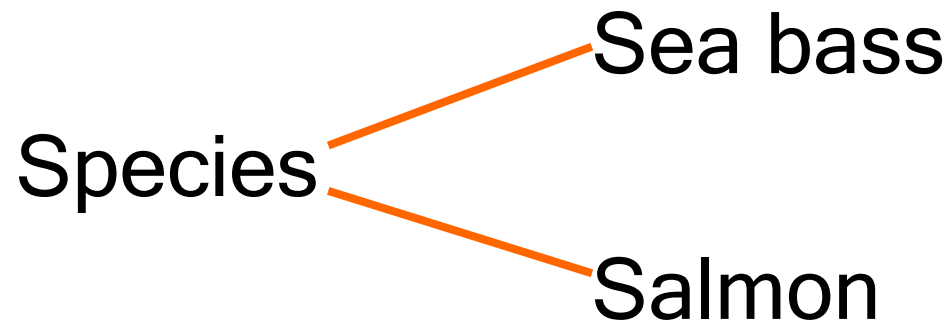
- Machine Perception
- An Example
- Pattern Recognition Systems
- The Design Cycle
- Learning and Adaptation
- Conclusion

Machine Perception

- Build a machine that can recognize patterns:
 - Speech recognition
 - Computer Vision: object recognition, face detection
 - Fingerprint identification
 - OCR (Optical Character Recognition)
 - DNA sequence identification

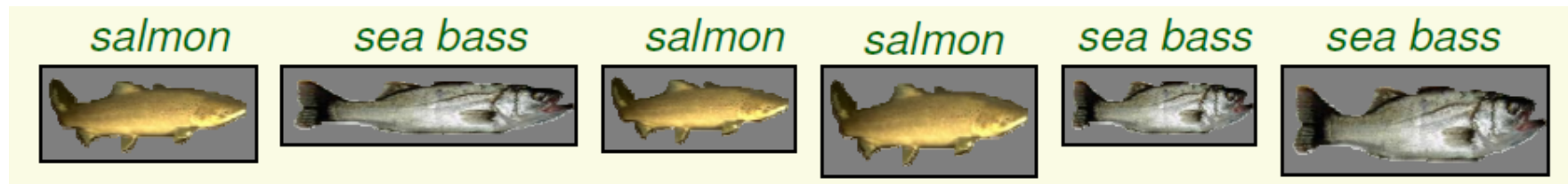
An Example

- “Sorting incoming Fish on a conveyor according to species using optical sensing”



Training

- Set up a camera and take some sample images
 - Label these images by hand



- Extract features
 - Length
 - Lightness
 - Width
 - Number and shape of fins
 - Position of the mouth, etc...

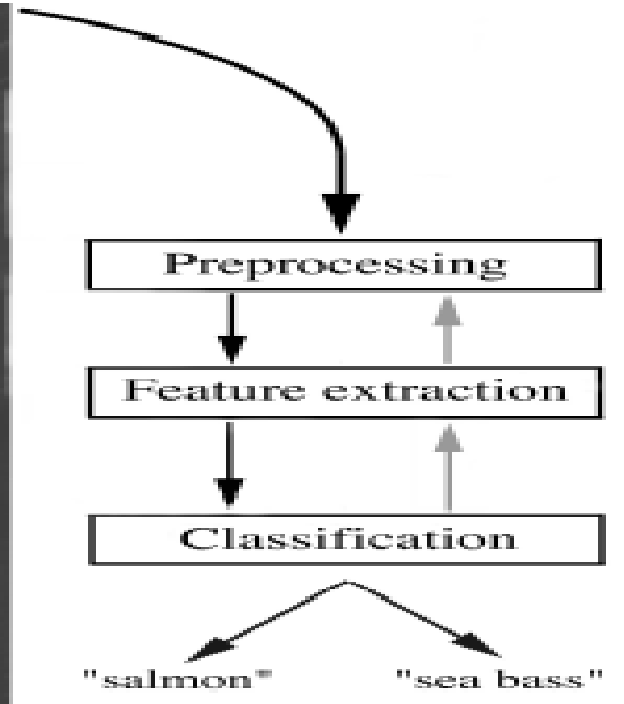
- Test whether this set of features is useful for a classifier

Preprocessing

- Use a segmentation operation to isolate fishes from one another and from the background

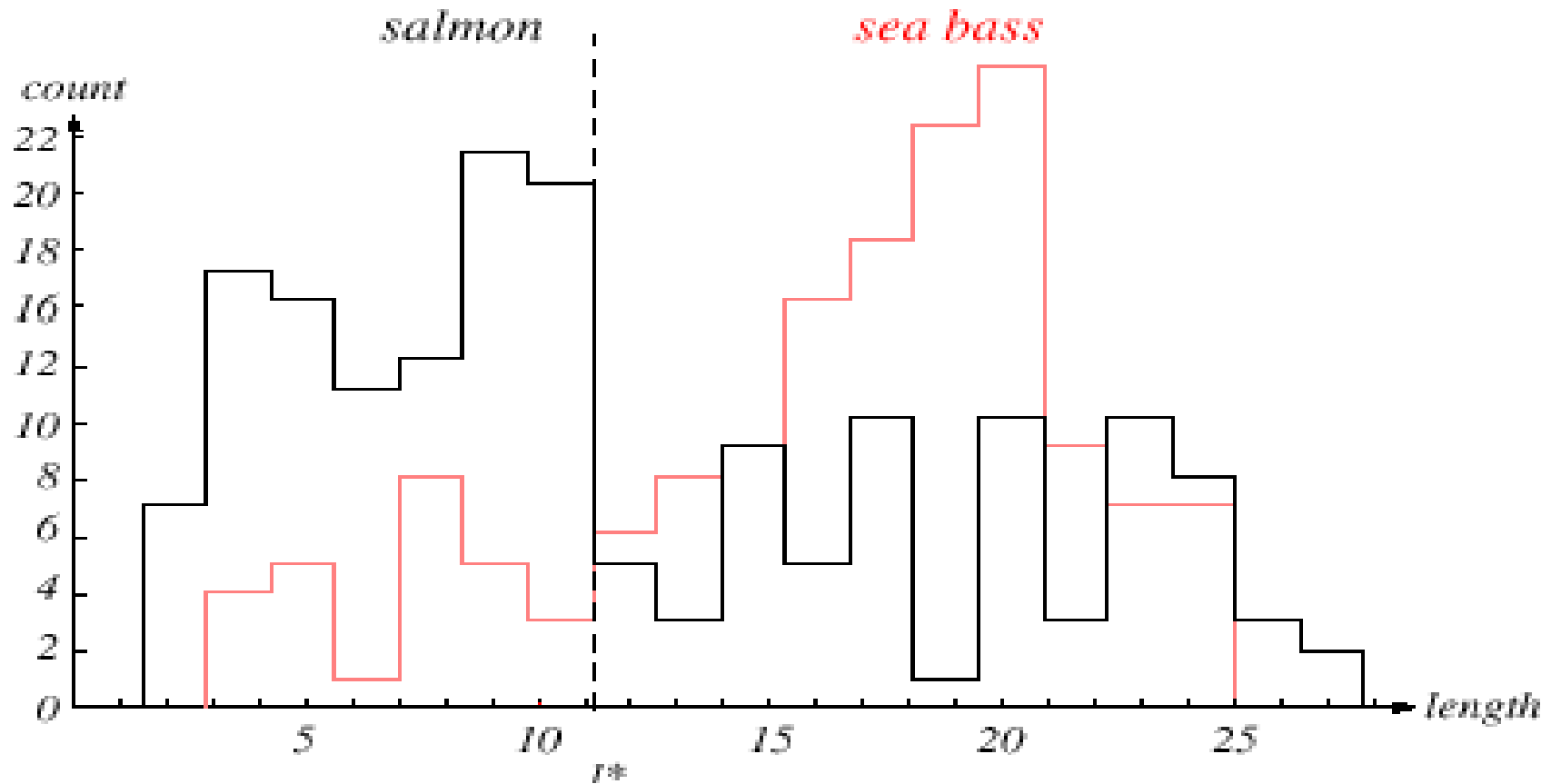


- Information from a single fish is sent to a feature extractor whose purpose is to reduce the data by measuring certain quantities
- The features are passed to a classifier



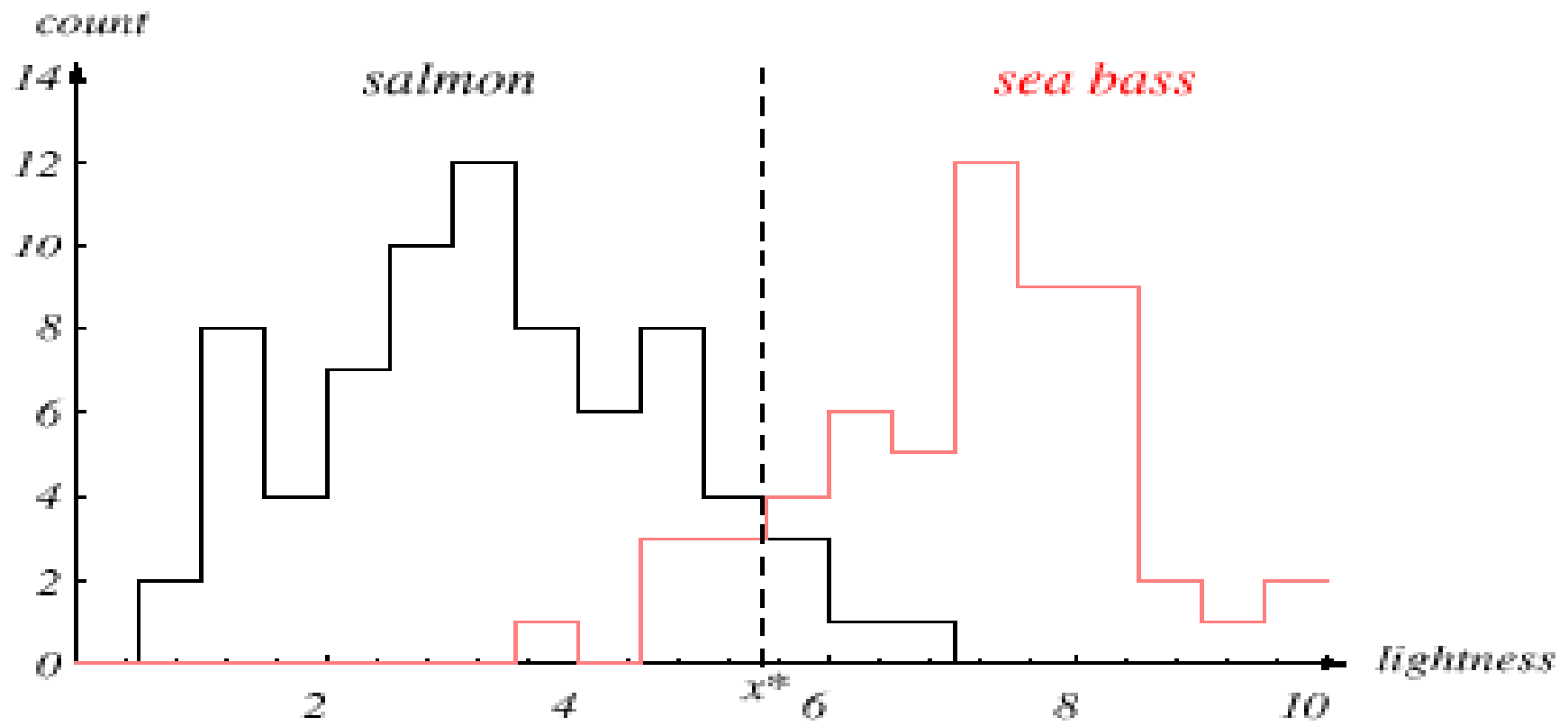
Classification

- Select the length of the fish as a possible feature for discrimination



Preliminary Results

- **Length** is a poor feature alone!
 - About 20% misclassification rate at best threshold choice
- Select **lightness** as a possible feature.

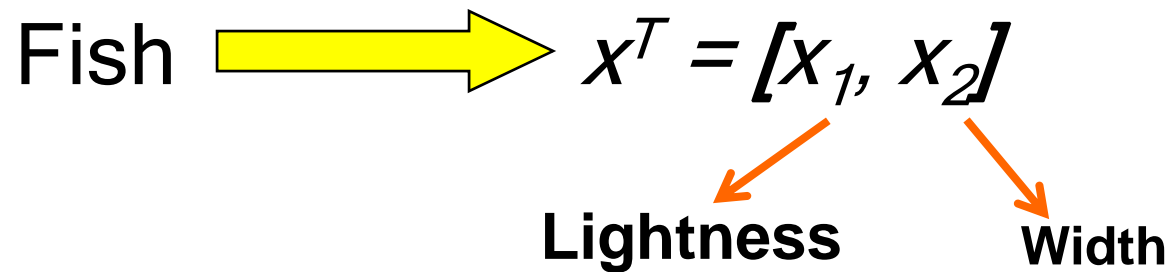


The Decision Boundary

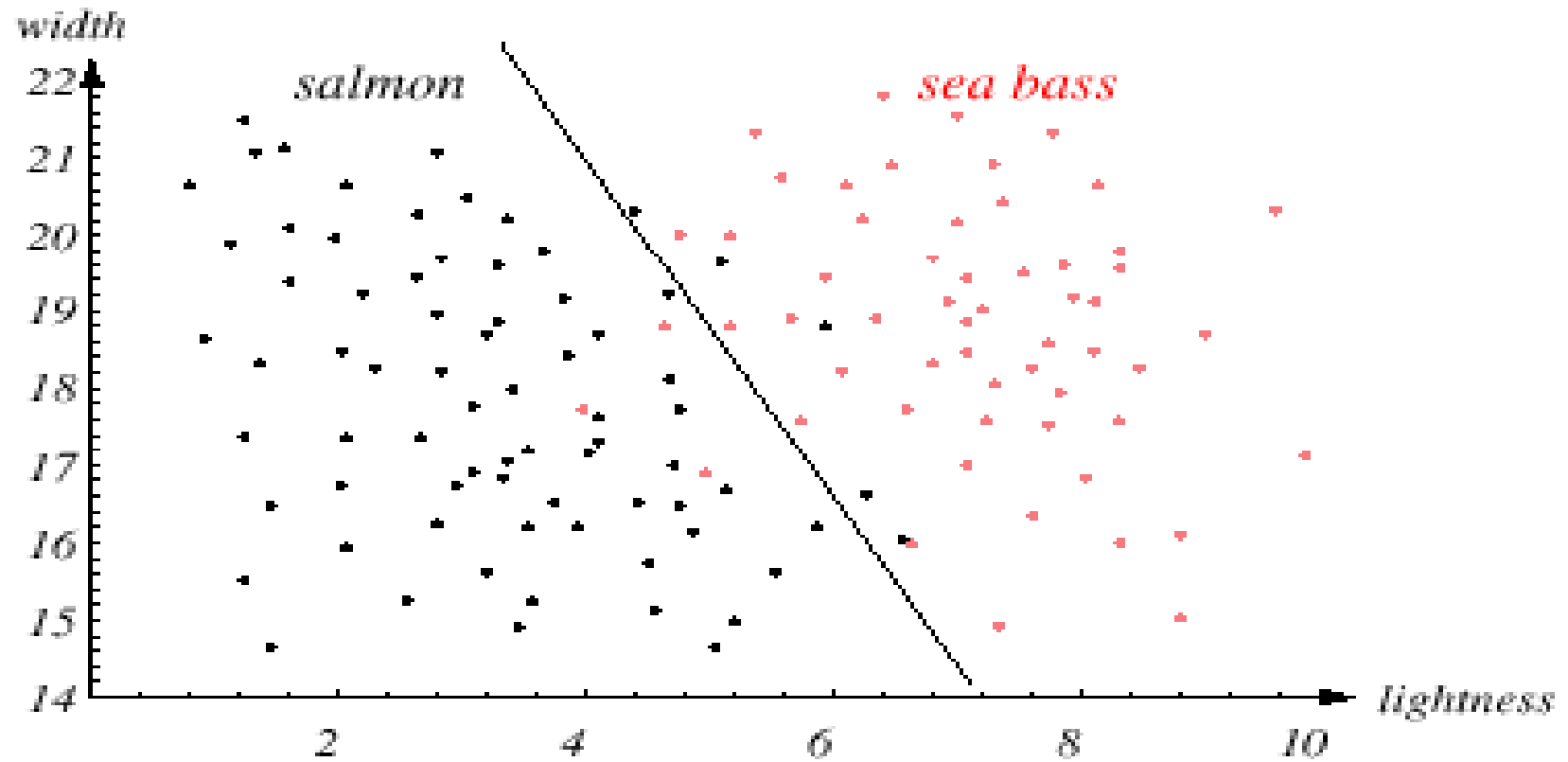
- Move the decision boundary toward smaller values of lightness in order to minimize the cost (reduce the number of misclassifications)

New Classifier

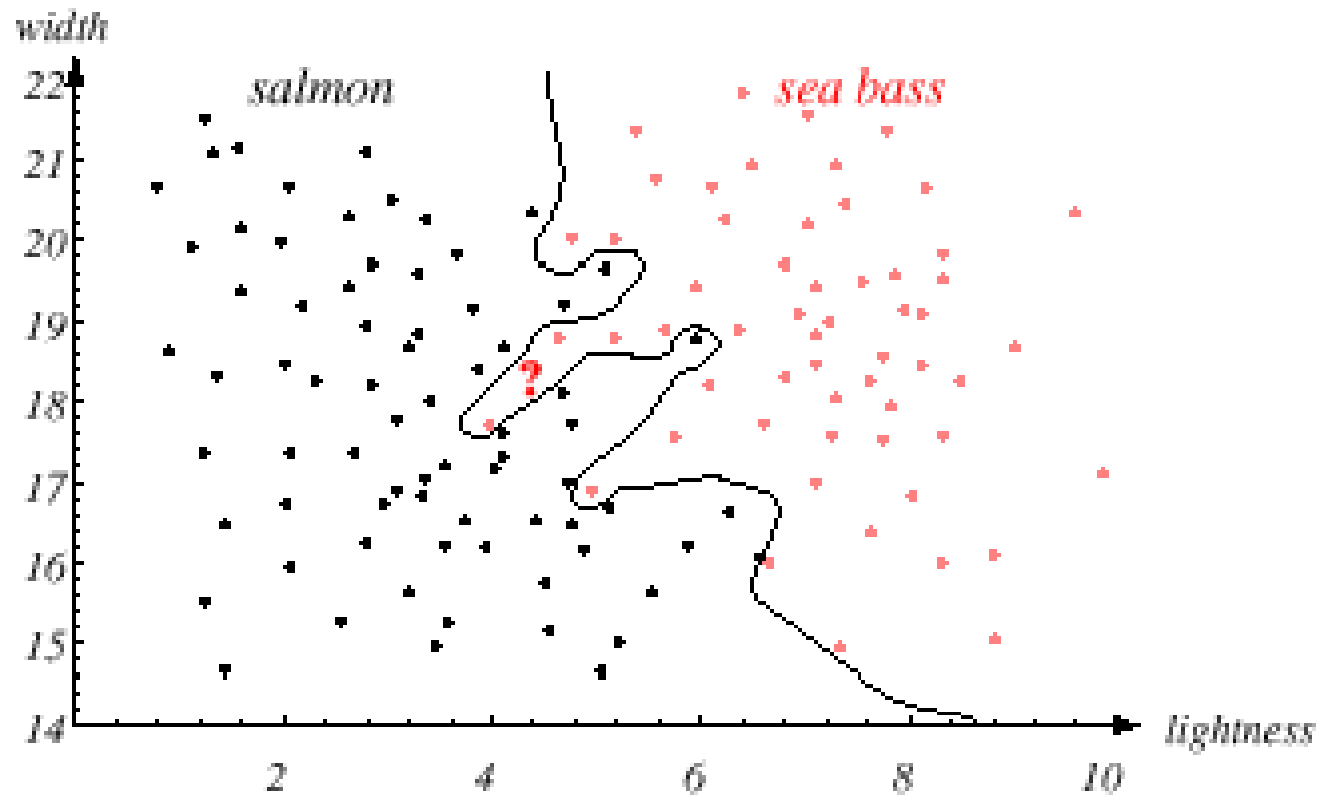
- Adopt the lightness and add the width of the fish



New Classifier

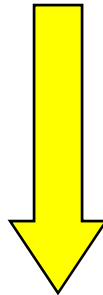


- We may add other features that are not correlated with the ones we already have
- Intuitively, the best decision boundary should be the one which provides an optimal performance such as in the following figure:



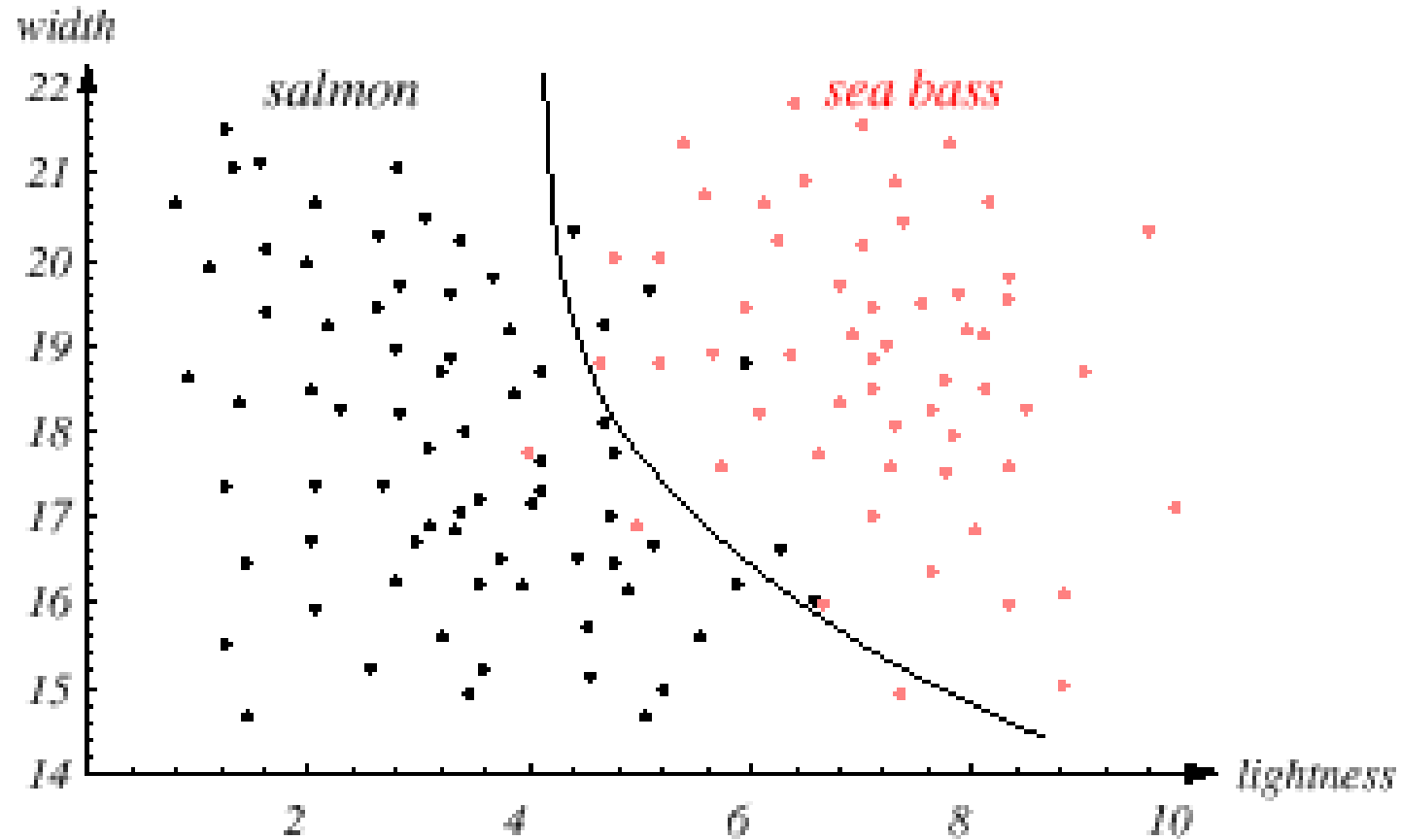
Generalization

- However, our satisfaction is premature because the central aim of designing a classifier is to correctly classify novel input



Issue of generalization!

Final Decision Boundary



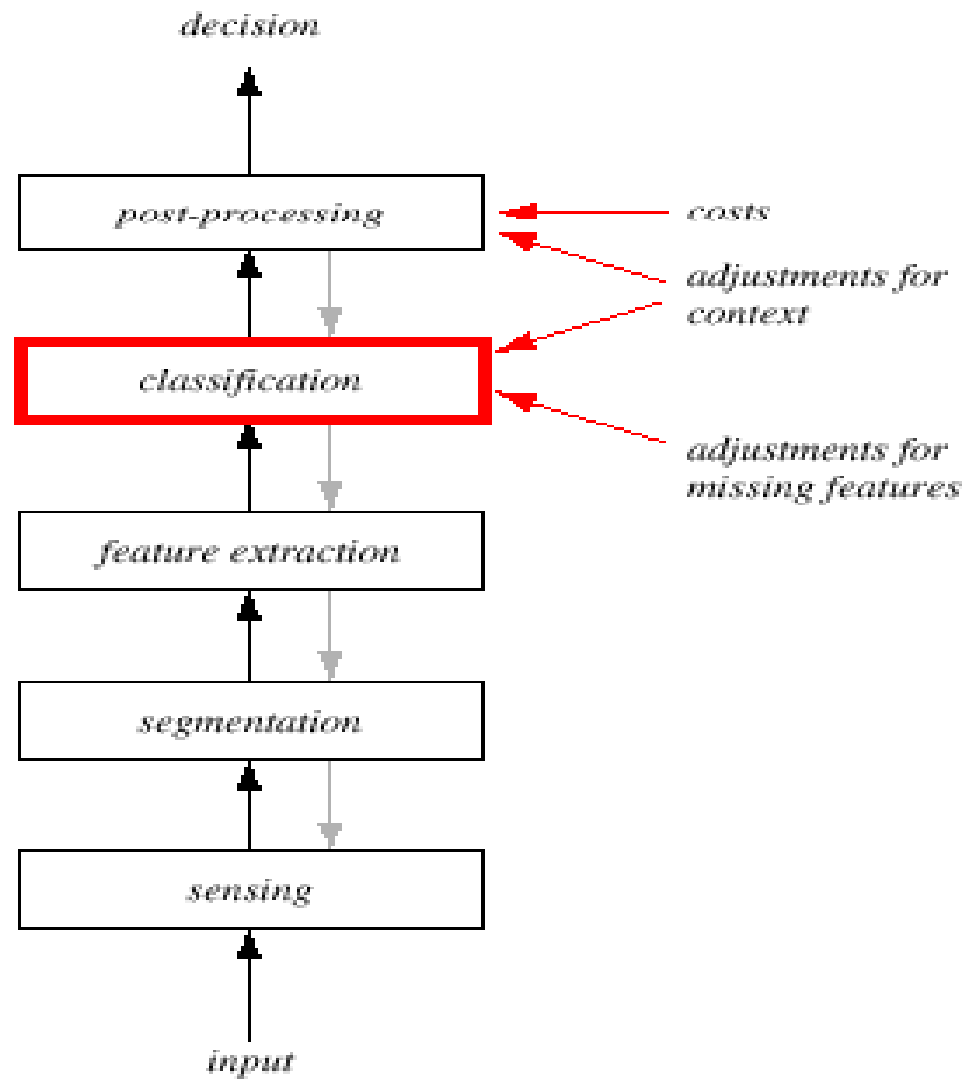
End of fish example. Back to business...

Pattern Recognition Systems

- Sensing
 - Use of a transducer (camera or microphone)
 - PR system depends of the bandwidth, the resolution sensitivity distortion of the transducer
- Segmentation and grouping
 - Patterns should be well separated and should not overlap

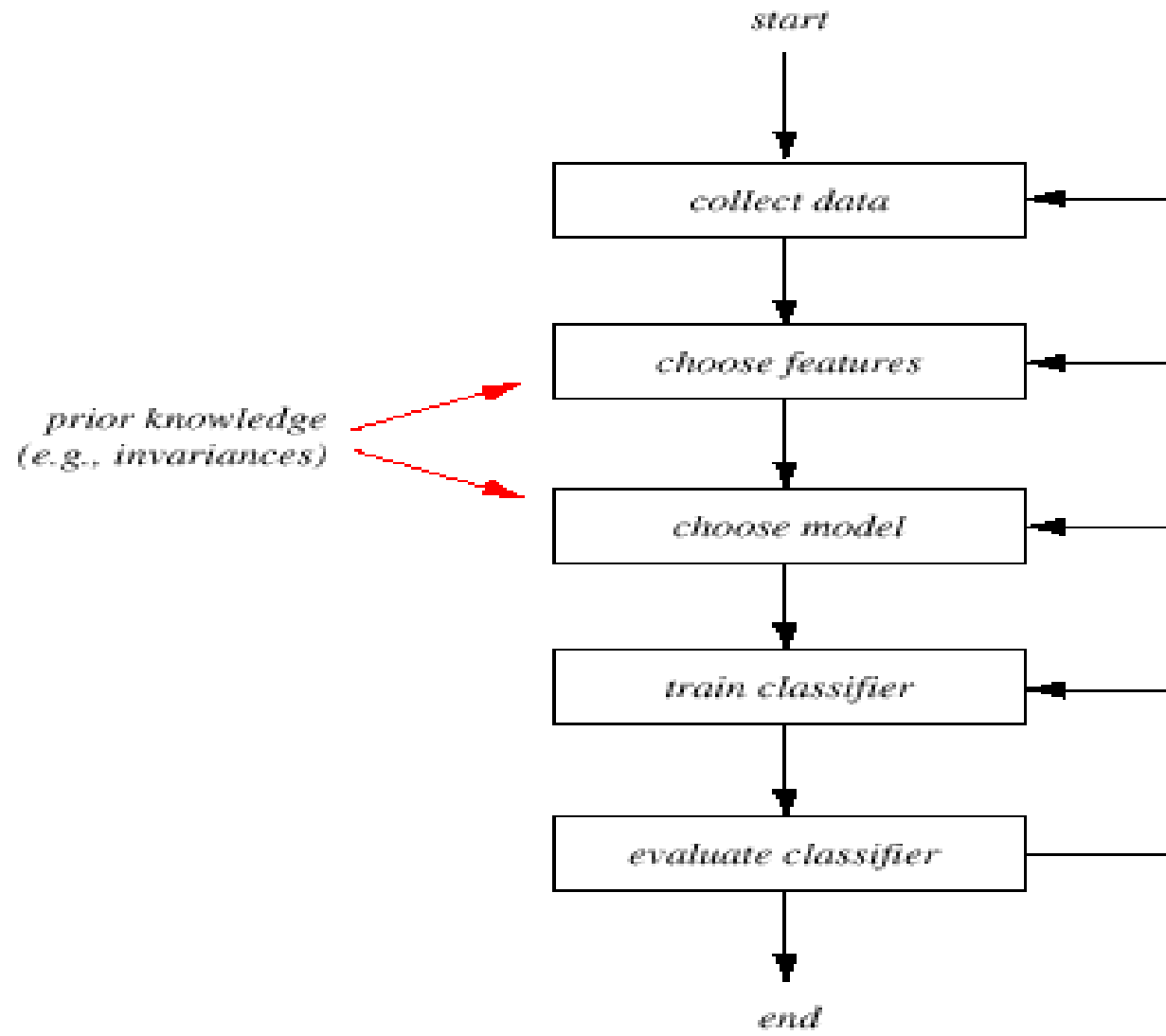
Pattern Recognition Systems

- Feature extraction
 - Discriminative features
 - Invariant features with respect to translation, rotation and scale.
- Classification
 - Use a feature vector provided by a feature extractor to assign the object to a category
- Post Processing
 - Exploit **context** other than the target pattern itself to improve performance



The Design Cycle

- Data collection
- Feature Choice
- Model Choice
- Training
- Evaluation
- Computational Complexity

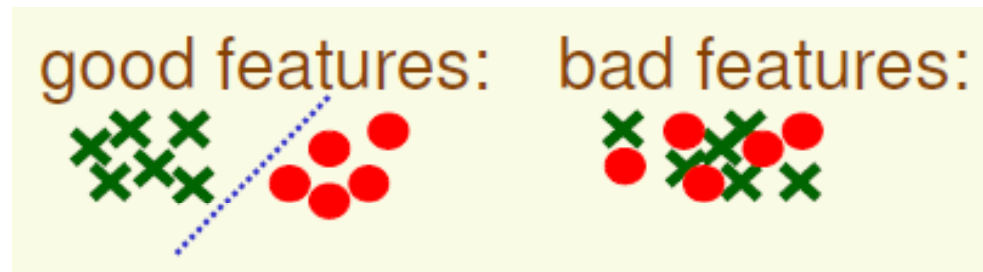


Data Collection

- How do we know when we have collected an adequately large and representative set of examples for training and testing the system?

Feature Choice

- Depends on the characteristics of the problem domain. Simple to extract, invariant to irrelevant transformations, insensitive to noise.



Model Choice

- What type of classifier to use?
- When should we try to reject one model and try another one?
- What is the best classifier for the problem?

Training

- Process of using data to determine the parameters of classifier
- Change parameters of the chosen model so that the model fits the collected data
- Many different procedures for training classifiers
- **Main scope of the course**

Evaluation

- Measure system performance (e.g. error rate)
- Identify the need for improvements in system components
- How to adjust complexity of the model to avoid overfitting? Any principled methods to do this?

Computational Complexity

- What is the trade-off between computational ease and performance?
- How does an algorithm scale as a function of the number of features, patterns or categories?

Learning and Adaptation

- Supervised learning
 - A teacher provides a category label or cost for each pattern in the training set
- Unsupervised learning
 - The system forms clusters or “natural groupings” of the input patterns

Probability Theory Review

See DHS Appendix A.4

The Axioms of Probability

- $0 \leq P(A) \leq 1$
- $P(\text{True}) = 1$
- $P(\text{False}) = 0$
- $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$

Overview

- Discrete Random Variables
- Expected Value
- Pairs of Discrete Random Variables
 - Conditional Probability
 - Bayes Rule
- Continuous Random Variables

Discrete Random Variables

- A ***Random Variable*** is a measurement on an outcome of a random experiment – denoted by r.v. x
- ***Discrete*** versus ***Continuous random variable***: an r.v. x is discrete if it can assume a finite or countably infinite number of values. An r.v. x is continuous if it can assume all values in an interval.

Examples

- Which of the following random variables are discrete and which are continuous?
- X = Number of houses sold by real estate developer per week?
- X = Number of heads in ten tosses of a coin?
- X = Weight of a child at birth?
- X = Time required to run 100 yards?

Probability Distribution Example: X is the Sum of Two Dice

Copyright Christopher Dougherty 1999–2006

red	1	2	3	4	5	6

This sequence provides an example of a discrete random variable. Suppose that you have a red die which, when thrown, takes the numbers from 1 to 6 with equal probability.

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Suppose that you also have a green die that can take the numbers from 1 to 6 with equal probability.

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

We will define a random variable X as the sum of the numbers when the dice are thrown.

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1						
2						
3						
4						
5						
6				10		

For example, if the red die is 4 and the green one is 6, X is equal to 10.

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1						
2						
3						
4						
5						
6						

Similarly, if the red die is 2 and the green one is 5, X is equal to 7.

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The table shows all the possible outcomes.

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

X
2
3
4
5
6
7
8
9
10
11
12

If you look at the table, you can see that X can be any of the numbers from 2 to 12.

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

X	f
2	
3	
4	
5	
6	
7	
8	
9	
10	
11	
12	

We will now define f , the frequencies associated with the possible values of X .

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

X	f
2	
3	
4	
5	4
6	
7	
8	
9	
10	
11	
12	

For example, there are four outcomes which make X equal to 5.

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

X	f
2	1
3	2
4	3
5	4
6	5
7	6
8	5
9	4
10	3
11	2
12	1

Similarly you can work out the frequencies for all the other values of X .

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

X	f	p
2	1	
3	2	
4	3	
5	4	
6	5	
7	6	
8	5	
9	4	
10	3	
11	2	
12	1	

Finally we will derive the probability of obtaining each value of X .

Probability Distribution Example: X is the Sum of Two Dice

red green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

X	f	p
2	1	
3	2	
4	3	
5	4	
6	5	
7	6	
8	5	
9	4	
10	3	
11	2	
12	1	

If there is $1/6$ probability of obtaining each number on the red die, and the same on the green die, each outcome in the table will occur with $1/36$ probability.

Probability Distribution Example: X is the Sum of Two Dice

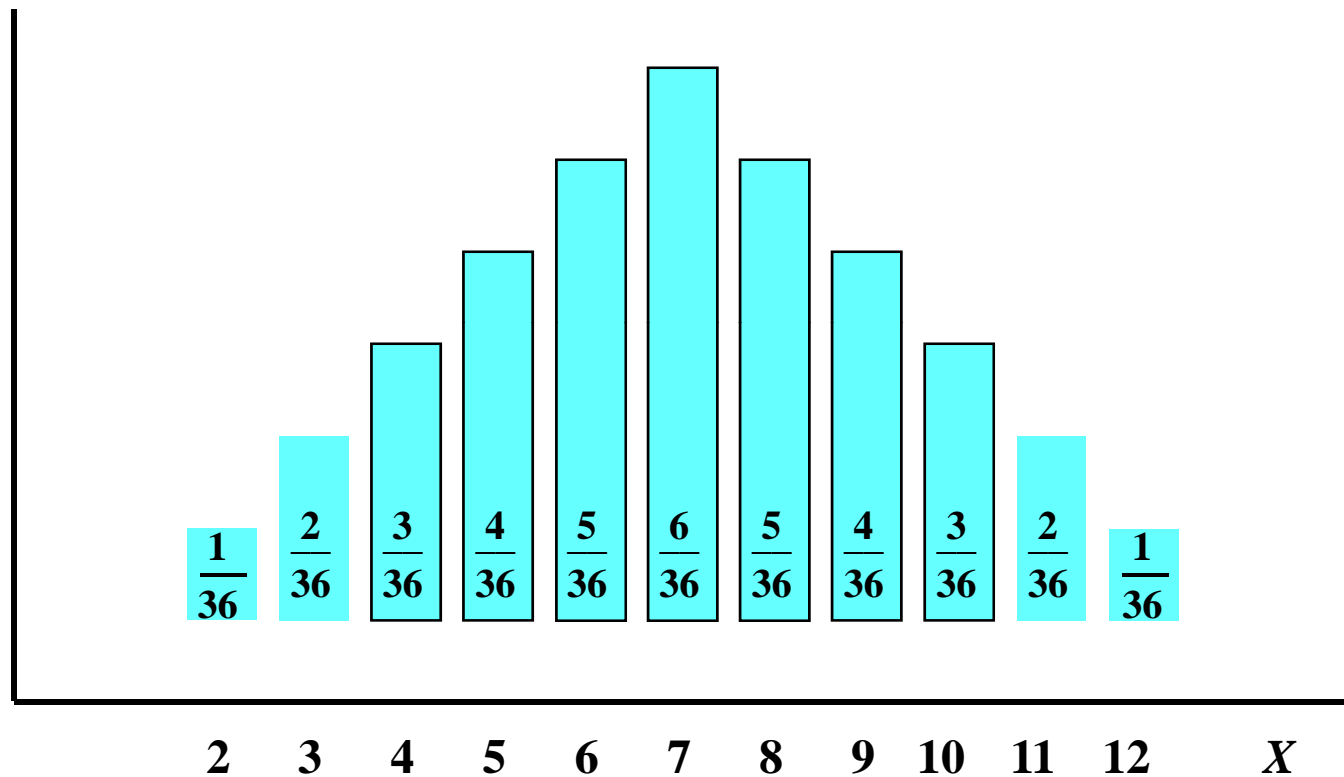
red green	1	2	3	4	5	6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

X	f	p
2	1	1/36
3	2	2/36
4	3	3/36
5	4	4/36
6	5	5/36
7	6	6/36
8	5	5/36
9	4	4/36
10	3	3/36
11	2	2/36
12	1	1/36

Hence to obtain the probabilities associated with the different values of X , we divide the frequencies by 36.

Probability Distribution Example: X is the Sum of Two Dice

probability



The distribution is shown graphically. in this example it is symmetrical, highest for X equal to 7 and declining on either side.

Overview

- Discrete Random Variables
- Expected Value
- Pairs of Discrete Random Variables
 - Conditional Probability
 - Bayes Rule
- Continuous Random Variables

Expected Value

- Definition of $E(X)$, the expected value of X :

$$E(X) = x_1 p_1 + \dots + x_n p_n = \sum_{i=1}^n x_i p_i$$

- The expected value of a random variable, also known as its population mean, is the weighted average of its possible values, the weights being the probabilities attached to the values

Expected Value Example

x_i	p_i	$x_i p_i$
x_1	p_1	$x_1 p_1$
x_2	p_2	$x_2 p_2$
x_3	p_3	$x_3 p_3$
x_4	p_4	$x_4 p_4$
x_5	p_5	$x_5 p_5$
x_6	p_6	$x_6 p_6$
x_7	p_7	$x_7 p_7$
x_8	p_8	$x_8 p_8$
x_9	p_9	$x_9 p_9$
x_{10}	p_{10}	$x_{10} p_{10}$
x_{11}	p_{11}	$x_{11} p_{11}$

$$\Sigma x_i p_i = E(X)$$

x_i	p_i	$x_i p_i$
2	1/36	2/36
3	2/36	6/36
4	3/36	12/36
5	4/36	20/36
6	5/36	30/36
7	6/36	42/36
8	5/36	40/36
9	4/36	36/36
10	3/36	30/36
11	2/36	22/36
12	1/36	12/36

$$252/36 = 7$$

Expected Value Properties

- Linear

$$E(X + Y) = E(X) + E(Y)$$

$$E(bX) = bE(X)$$

$$E(b) = b$$

$$Y = b_1 + b_2X$$

$$E(Y) = E(b_1 + b_2X)$$

$$= E(b_1) + E(b_2X)$$

$$= b_1 + b_2 E(X)$$

- Also denoted by μ

Variance

$$\text{Var}(X) = E[(X - \mu)^2] = \sum (x_i - \mu)^2 P(X = x_i)$$

$$\text{Var}(X) = \sigma^2$$

$$\text{Var}(X) = E[(X - \mu)^2] = E[X^2] - (E[X])^2$$

(Prove it.)

Overview

- Discrete Random Variables
- Expected Value
- Pairs of Discrete Random Variables
 - Conditional Probability
 - Bayes Rule
- Continuous Random Variables

Pairs of Discrete Random Variables

- Let x and y be two discrete r.v.
- For each possible pair of values, we can define a *joint probability* $p_{ij} = \Pr[x=x_i, y=y_j]$
- We can also define a *joint probability mass function* $P(x,y)$ which offers a complete characterization of the pair of r.v.

$$P_x(x) = \sum_{y \in Y} P(x, y)$$

Marginal distributions

$$P_y(y) = \sum_{x \in X} P(x, y)$$

Note that P_x and P_y are different functions

Statistical Independence

Two random variables x and y are said to be independent, if and only if

$$P(x,y)=P_x(x) P_y(y)$$

that is, when knowing the value of x does not give us additional information for the value of y .

Or, equivalently

$$E[f(x)g(y)] = E[f(x)] E[g(y)]$$

for any functions $f(x)$ and $g(y)$.

Conditional Probability

- When two r.v. are not independent, knowing one allows better estimate of the other (e.g. outside temperature, season)

$$\Pr[x = x_i | y = y_j] = \frac{\Pr[x = x_i, y = y_j]}{\Pr[y = y_j]}$$

- If independent $P(x|y)=P(x)$

Conditional Probability Example

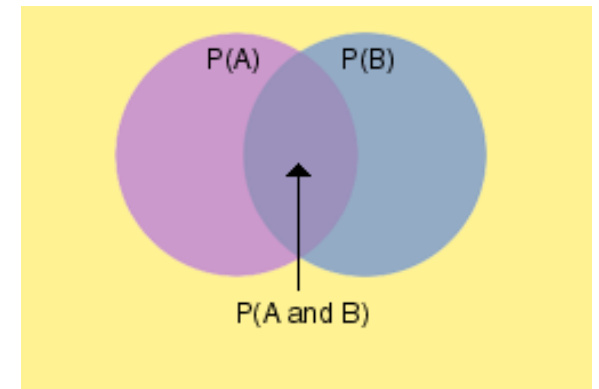
A jar contains black and white marbles. Two marbles are chosen without replacement. The probability of selecting a black marble and then a white marble is 0.34, and the probability of selecting a black marble on the first draw is 0.47. What is the probability of selecting a white marble on the second draw, given that the first marble drawn was black?

Conditional Probability Example

A jar contains black and white marbles. Two marbles are chosen without replacement. The probability of selecting a black marble and then a white marble is 0.34, and the probability of selecting a black marble on the first draw is 0.47. What is the probability of selecting a white marble on the second draw, given that the first marble drawn was black?

$$P(\text{White} | \text{Black}) = \frac{P(\text{Black} \wedge \text{White})}{P(\text{Black})} = \frac{0.34}{0.47} = 0.72$$

A is black in first draw, B is white in second draw



Law of Total Probability

- If an event A can occur in m different ways and if these m different ways are mutually exclusive, then the probability of A occurring is the sum of the probabilities of the sub-events

$$P(X = x_i) = \sum_j P(X = x_i | Y = y_j)P(Y = y_j)$$

Law of Total Probability

- This can also be written as:

$$P_x(x) = \sum_{y \in Y} P(x, y)$$

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

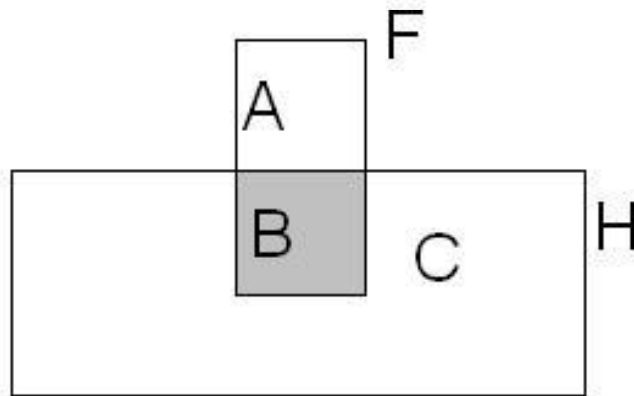
Bayes Rule

$$P(x | y) = \frac{P(x, y)}{P(y)} = \frac{P(y | x)P(x)}{\sum_{x \in X} P(x, y)}$$

$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

- x is the unknown cause
- y is the observed evidence
- Denominator often omitted (maximum a posteriori solution)
- Bayes rule shows how probability of x changes after we have observed y

Illustration



Let's say we have $P(F)$, $P(H)$, and $P(H|F)$, like in the example in class.

Areawise, $P(F) = A + B$, $P(H) = B + C$.

Also, $P(H|F) = \frac{B}{A + B}$

Thus, to get the opposite conditional probability, i.e., $P(F|H)$, we need to figure out $\frac{B}{B + C}$

Since we know $B / (A+B)$, we can get $B / (B+C)$ by multiplying by $(A+B)$ and dividing by $(B+C)$. But since we already calculated, $A+B = P(F)$, and $B + C = P(H)$, so we are actually multiplying by $P(F)$ and dividing by $P(H)$. Which is Bayes Rule:

$$P(F|H) = P(H|F) * \frac{P(F)}{P(H)}$$

Overview

- Discrete Random Variables
- Expected Value
- Pairs of Discrete Random Variables
 - Conditional Probability
 - Bayes Rule
- Continuous Random Variables

Continuous Random Variables

- Examples: room temperature, time to run 100m, weight of child at birth...
- Cannot talk about probability of that x has a particular value
- Instead, probability that x falls in an interval => **probability density function**

$$\Pr[x \in (a, b)] = \int_a^b p(x) dx$$

$$p(x) \geq 0 \text{ and } \int_{-\infty}^{\infty} p(x) dx = 1$$

Expected Value

$$E[x] = \mu = \int_{-\infty}^{\infty} xp(x)dx$$

$$E[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)dx$$

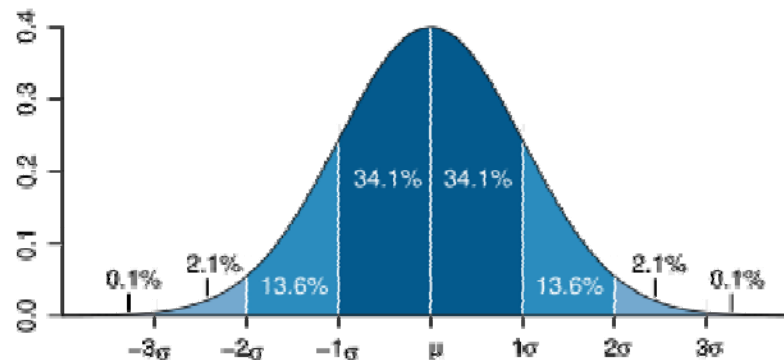
$$Var[x] = \sigma^2 = \int_{-\infty}^{\infty} (x - \mu)^2 p(x)dx$$

- **Bayes rule**
$$p(x | y) = \frac{p(y | x)p(x)}{\int_{-\infty}^{\infty} p(y | x)p(x)dx}$$
$$\text{posterior} = \frac{\text{likelihood} * \text{prior}}{\text{evidence}}$$

Normal (Gaussian) Distribution

- Central Limit Theorem: under various conditions, the distribution of the sum of d independent random variables approaches a limiting form known as **the normal distribution**

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



Normal (Gaussian) Distribution

