Activity Recognition 3 Classification

Mobile Computing

Minho Shin

2012. 9

Entropy: Formula

- The entropy (in bits) of a discrete random variable M:

$$
\begin{aligned}
H(M) &= -\sum_m Pr(M=m)\log_2 Pr(M=m) \\
&= -\sum_m p_m \log_2 p_m \\
&= \sum_m p_m \log_2 \frac{1}{p_m}
\end{aligned}
$$

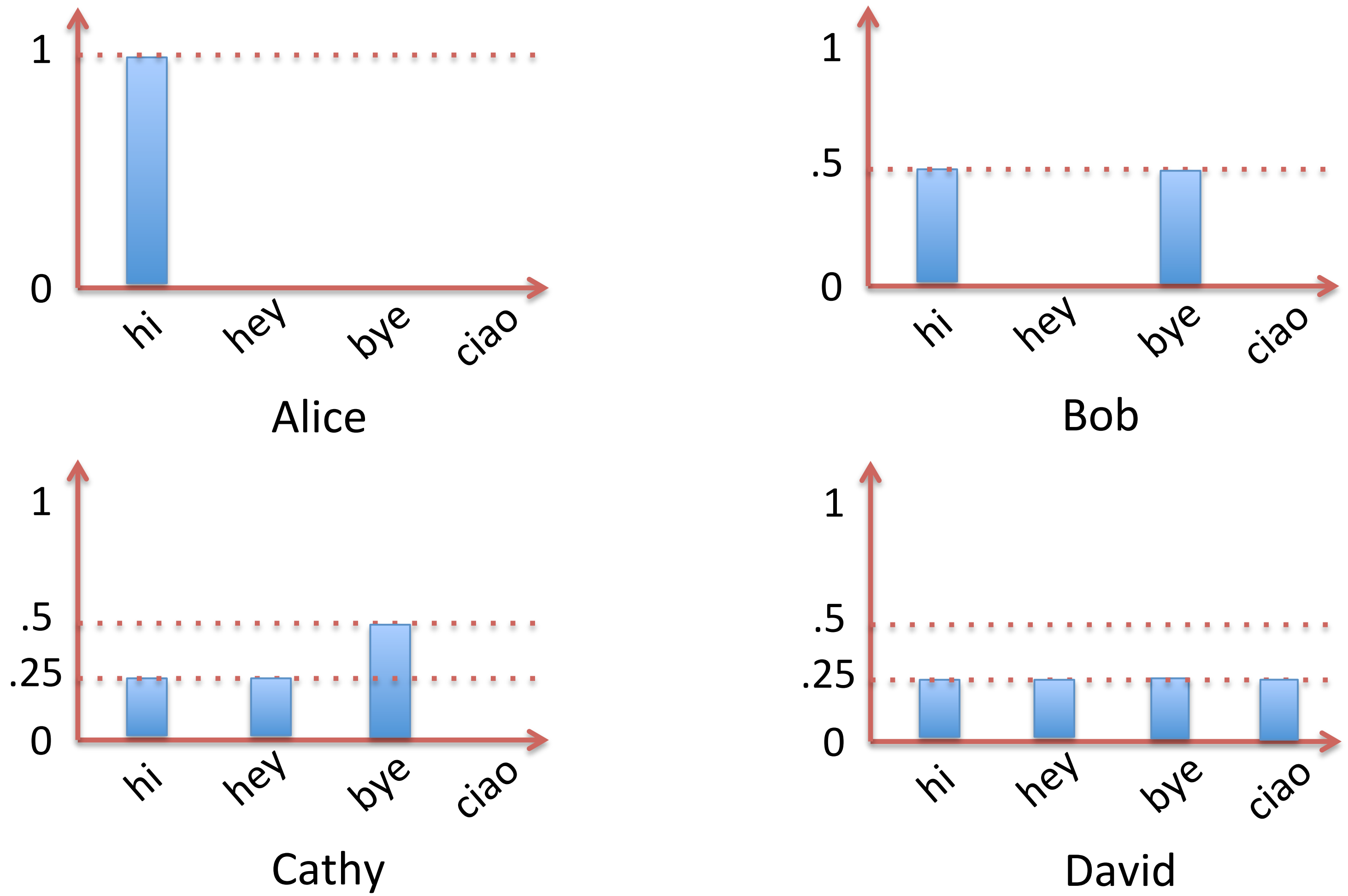
- Interpretation
  - Average # of bits to express each message
- Maximized when uniform
  - p_m is the same for all messages

Message Distribution

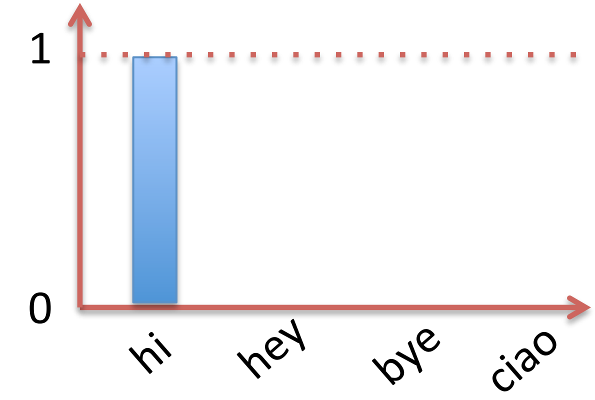
- Speakers have different message distribution
  - 4 messages possible: “hi”, “hey”, “bye”, “ciao”

  - Alice always says “hi”
  - Bob says only “hi” or “bye” with same probability
  - Cathy says “bye” half the time, and “hi” and “hey” half the time with equal probability
  - David says “hi”, “hey”, “bye”, “ciao” with equal probability

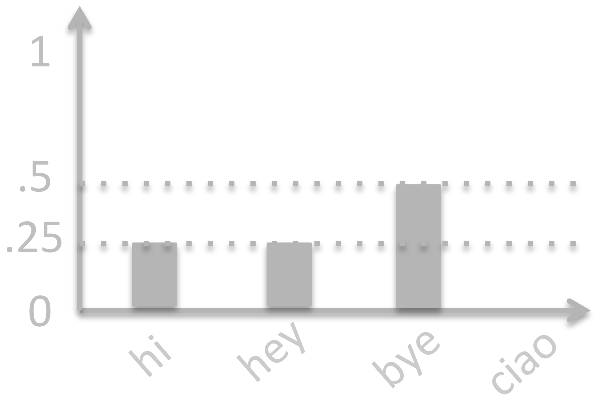
Message Distribution



Message Distribution



Alice



Cathy

$$p_{hi} = 1 \quad p_{hey} = 0$$
$$p_{bye} = 0 \quad p_{ciao} = 0$$

$$\begin{aligned} H(M) &= p_{hi} \log \frac{1}{p_{hi}} + p_{hey} \log \frac{1}{p_{hi}} \\ &+ p_{bye} \log \frac{1}{p_{bye}} + p_{ciao} \log \frac{1}{p_{ciao}} \end{aligned}$$

$$\begin{aligned} H(M) &= 1 \cdot \log \frac{1}{1} + 0 + 0 + 0 \\ &= 0 \end{aligned}$$

David

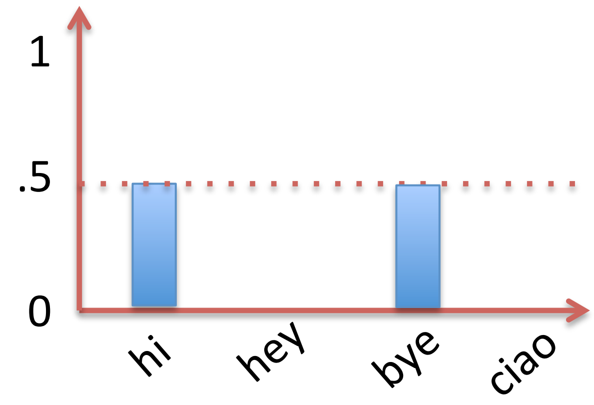
Message Distribution

$$p_{hi} = 0.5 \quad p_{hey} = 0$$
$$p_{bye} = 0.5 \quad p_{ciao} = 0$$

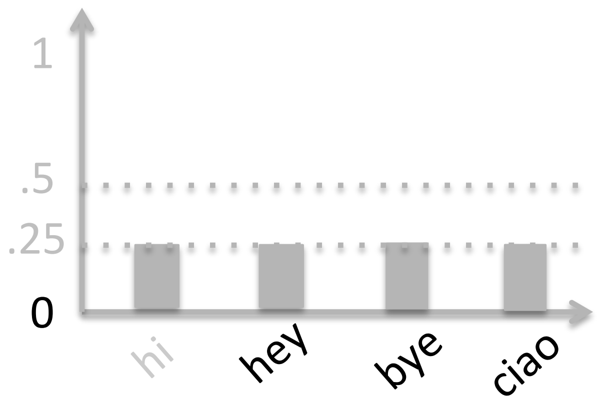
$$\begin{aligned} H(M) &= p_{hi} \log \frac{1}{p_{hi}} + p_{hey} \log \frac{1}{p_{hi}} \\ &+ p_{bye} \log \frac{1}{p_{bye}} + p_{ciao} \log \frac{1}{p_{ciao}} \end{aligned}$$

$$\begin{aligned} H(M) &= \frac{1}{2} \cdot \log \frac{1}{\frac{1}{2}} + 0 \\ &+ \frac{1}{2} \cdot \log \frac{1}{\frac{1}{2}} + 0 \\ &= \frac{1}{2} + \frac{1}{2} = 1 \end{aligned}$$

Cathy

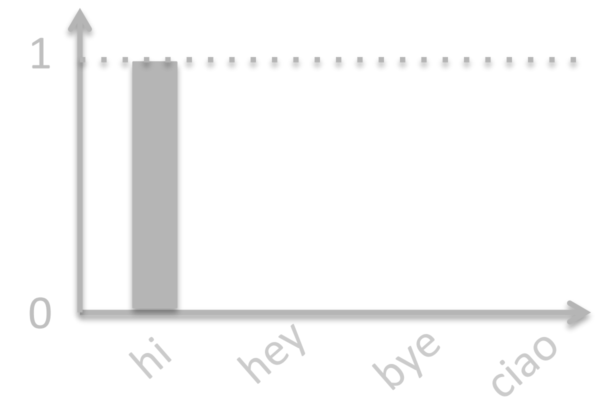


Bob

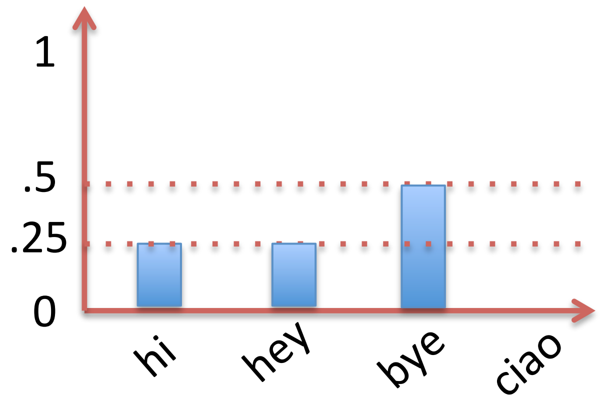


David

Message Distribution



Alice



Cathy

$$p_{hi} = 0.25 \quad p_{hey} = 0.25$$
$$p_{bye} = 0.5 \quad p_{ciao} = 0$$

$$\begin{aligned} H(M) &= p_{hi}\log\frac{1}{p_{hi}} + p_{hey}\log\frac{1}{p_{hi}} \\ &+ p_{bye}\log\frac{1}{p_{bye}} + p_{ciao}\log\frac{1}{p_{ciao}} \end{aligned}$$

$$\begin{aligned} H(M) &= \frac{1}{4}\cdot\log\frac{1}{\frac{1}{4}} + \frac{1}{4}\cdot\log\frac{1}{\frac{1}{4}} \\ &+ \frac{1}{2}\cdot\log\frac{1}{\frac{1}{2}} + 0 \\ &= \frac{1}{2} + \frac{1}{2} + \frac{1}{2} = 1.5 \end{aligned}$$

David

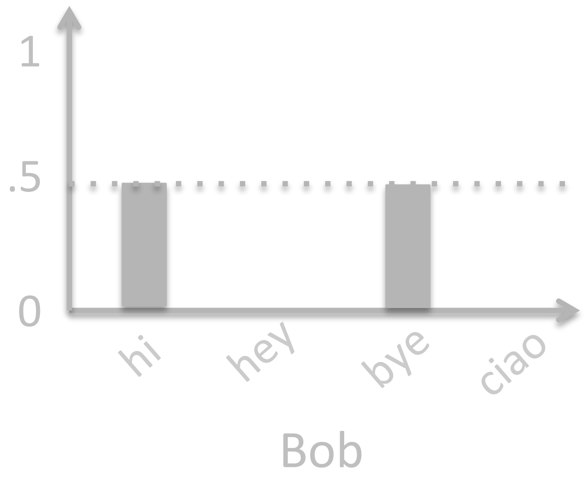
Message Distribution

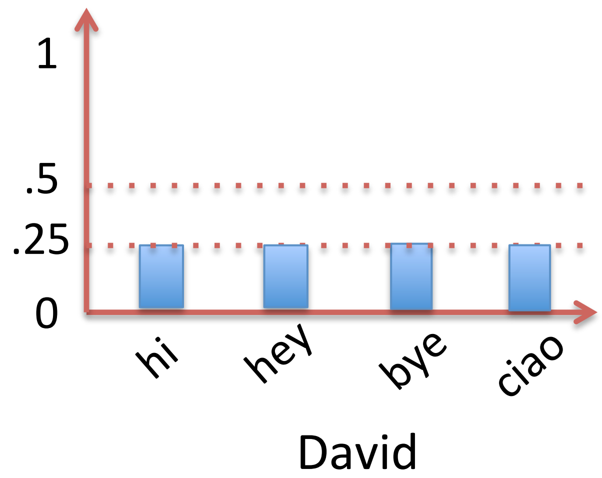
$$
\begin{aligned}
p_{hi} &= 0.25 \quad p_{hey} = 0.25 \\
p_{bye} &= 0.25 \quad p_{ciao} = 0.25
\end{aligned}
$$

$$
\begin{aligned}
H(M) &= p_{hi} \log \frac{1}{p_{hi}} + p_{hey} \log \frac{1}{p_{hi}} \\
&+ p_{bye} \log \frac{1}{p_{bye}} + p_{ciao} \log \frac{1}{p_{ciao}}
\end{aligned}
$$

$$
\begin{aligned}
H(M) &= 4 \times \frac{1}{4} \cdot \log \frac{1}{\frac{1}{4}} \\
&= 2
\end{aligned}
$$

Cathy



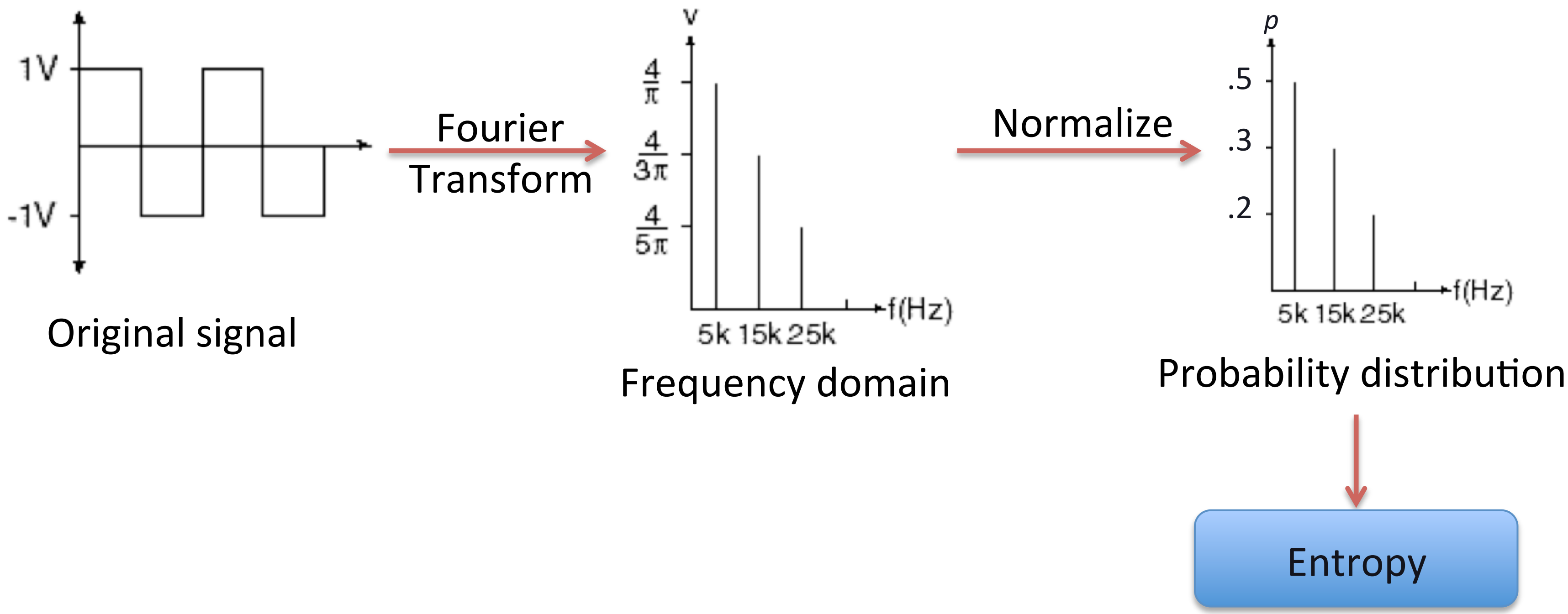


Features (revisited)

- Frequency-domain entropy
  - Differentiate between walking and cycling
- What is *frequency domain*?
- What is *Entropy*?
- What is *frequency-domain entropy*?

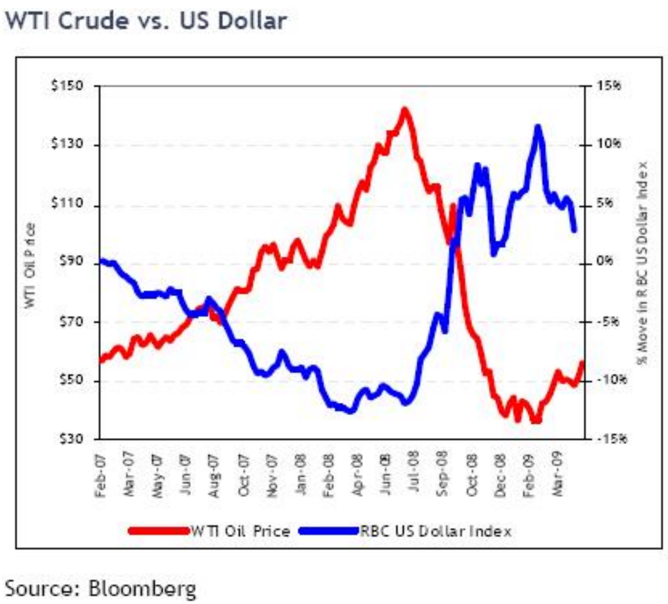
Frequency-domain Entropy

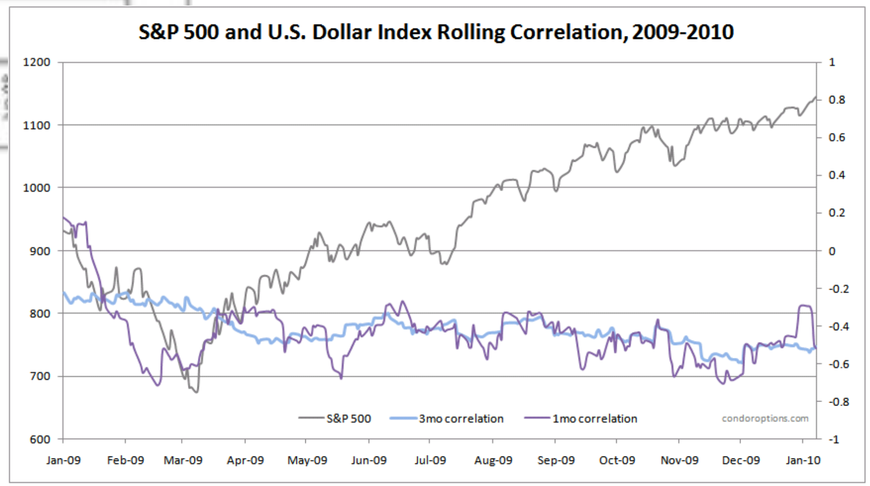
- Given a signal in time-domain, convert to frequency-domain, normalize it, then compute entropy



Correlation

- Degree of dependency between two signals





Correlation Coefficient

- Given two random variables X, Y, corr-coef is

$$corr(X,Y) = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y} \in [-1,1]$$

- Given two series of *n* measurements x_i *and* y_i

$$corr(X,Y) = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2}\sqrt{\sum_i (x_i - \bar{x})^2}}$$

- Interpretation
  - *+1: perfect dependency*
  - *0: no dependency*
  - *-1: opposite dependency*

$$\bar{x} = \frac{\sum_i x_i}{n} \quad \bar{y} = \frac{\sum_i y_i}{n}$$

Feature Set

- X average
- Y average
- Z average
- X variance
- Y variance
- Z variance
- X energy
- Y energy
- Z energy
- X entropy
- Y entropy
- Z entropy
- X, Y correlation
- Y, Z correlation
- Z, X correlation

Feature selection/extraction

- Curse of dimensionality
  - If dimension of features is high, classification becomes difficult
- Solution
  - Reduce the number of features
  - Feature selection: remove less important features
  - Feature extraction: generate smaller # of features

Feature selection

- Given a feature set *F*, get a subset *G* of *F*
- Discarding features that are little helpful for classification
- But, finding the subset is exponentially expensive
  - For example, if *F={f1, f2, …, fd}* (*d* is *F*'s dimension), for *m=1, 2, .., d*, the we have to check all subsets of *F* of size *m*

Feature selection: Suboptimal Algorithms

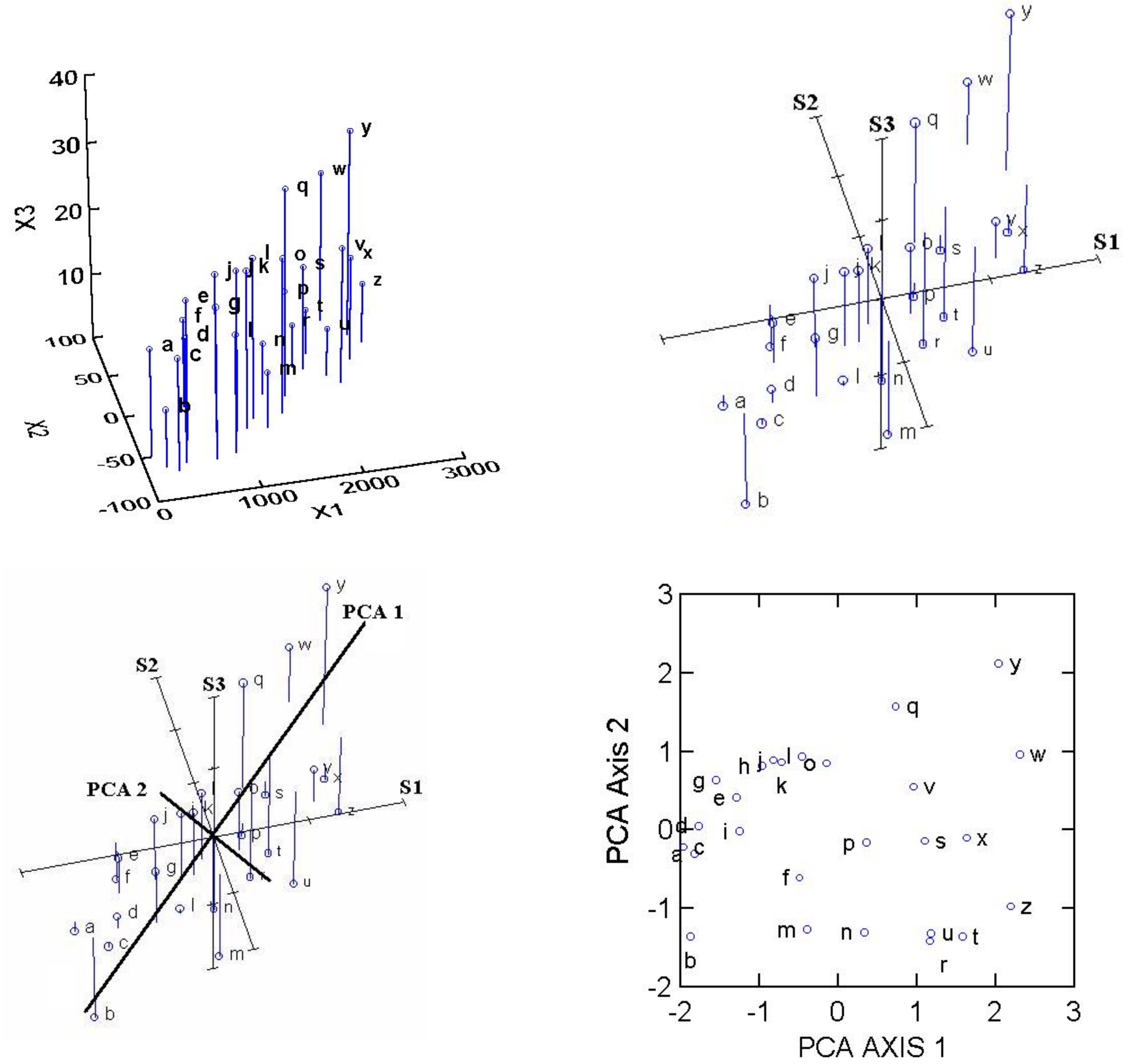
- Use sub-optimal search algorithm
  - Branch-and-bound search
  - Sequential forward/backward search (SFS-SBS)
  - Sequential forward/backward floating search (SFFS-SBFS)
- Sequential search algorithms
  - Iterative procedure
  - Add or remove some features at each step so that the new set leads to a better classification performance, measured by
    - Inter-class distance / intra-class distance
    - Analyze classifier output

Feature Extraction

- Idea: another data representation can be constructed in a subspace (less dimension) while keeping discriminative capability
- Lose physical meaning
- Example algorithms
  - PCA (Principle Component Analysis): transform features into small number of uncorrelated variables
  - ICA (Independent Component Analysis)
- Feature selection and extractions can be used together

PCA

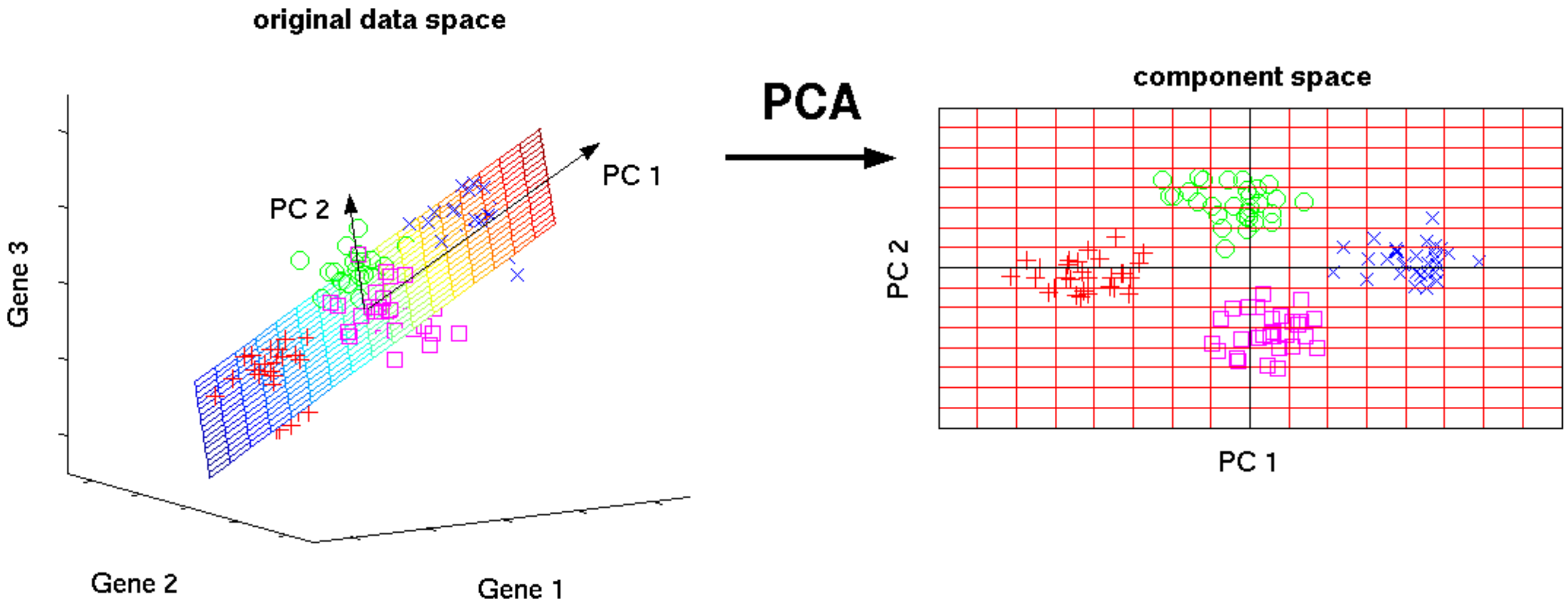
- How can we reduce dimension?



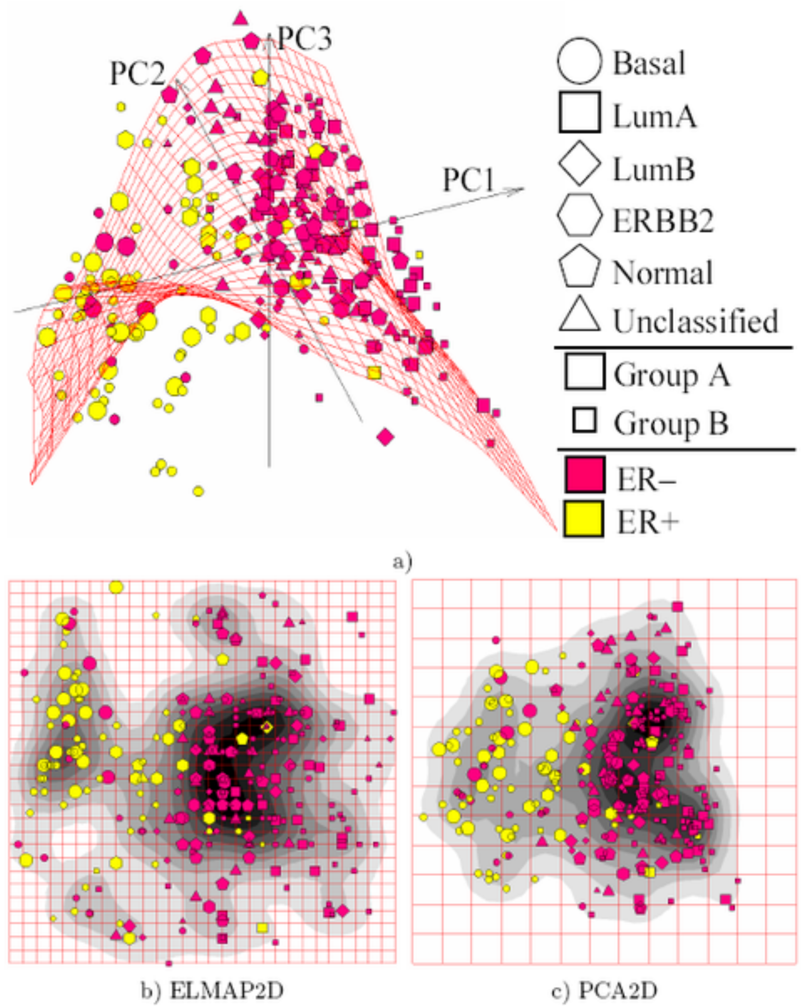
X3
X2
X1
S1
S2
S3
PCA 1
PCA 2
PCA Axis 2
PCA AXIS 1

PCA

- Identify orthogonal axes with maximum variance of data
  - discard the axis with low variance

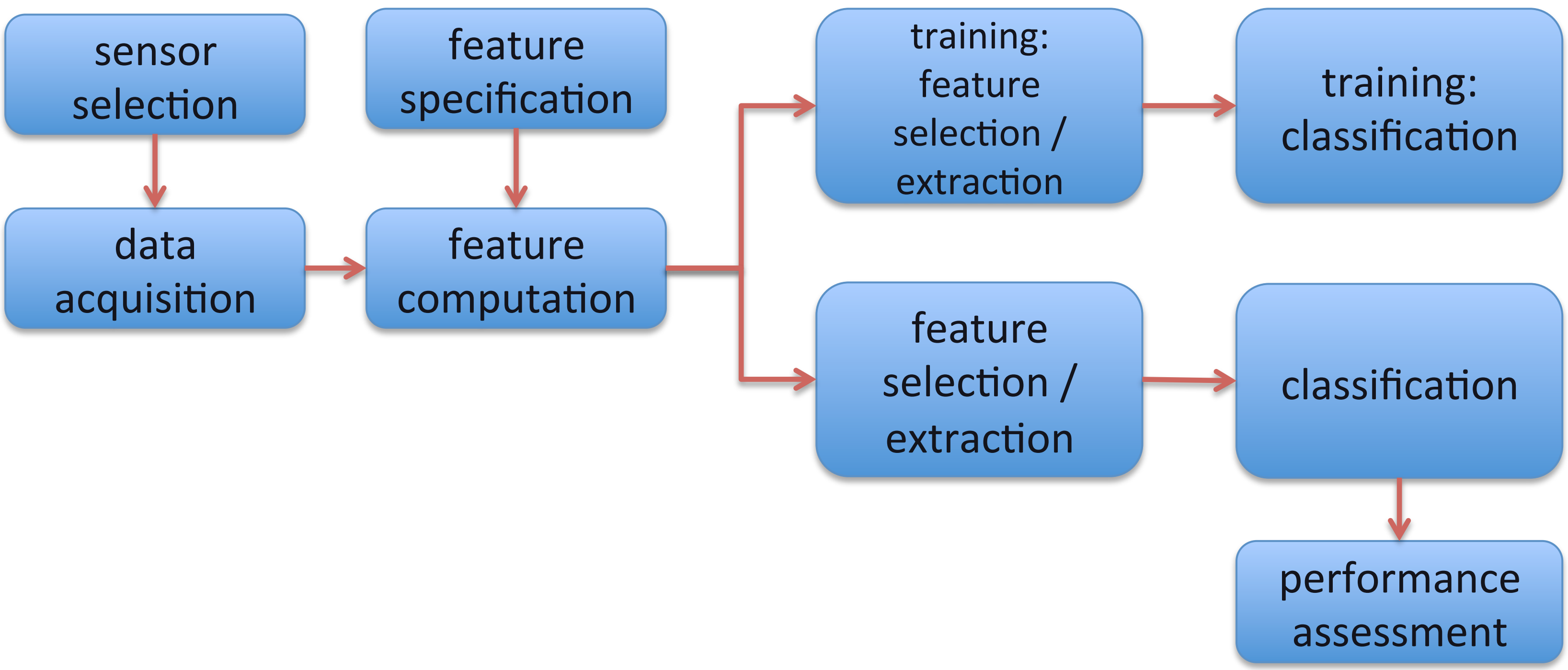


Non-linear PCA



Classification with supervised learning

- Classification: determine the type of activity



Type of Classifiers

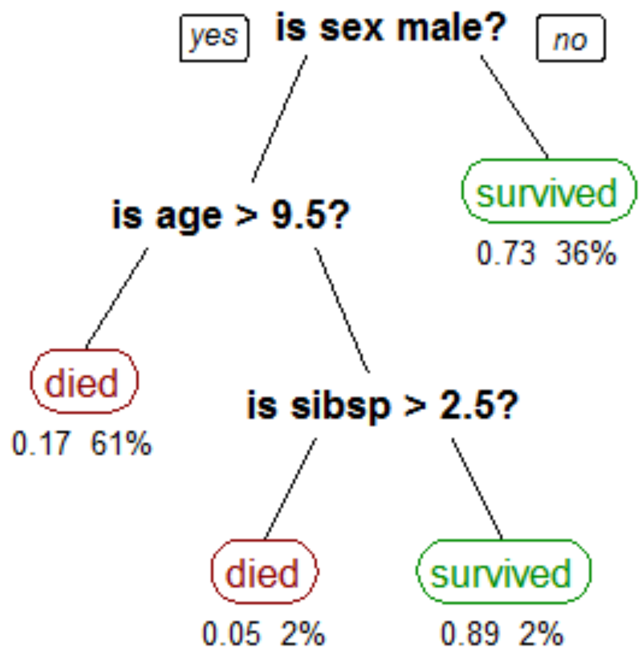
- Supervised & Unsupervised
  - Supervised: class membership of each feature vector is known
  - Unsupervised: Only the number of classes is known
- Single-frame & Sequential
  - Single-frame: Each frame is classified regardless of previous frames
  - Sequential: Each frame is classified in consideration of previous frames

Type of Classifiers

- Probabilistic & Geometric & Template matching
  - Template matching: Based on similarity between data and templates obtained by training or defined by the designer
  - Binary classifier: Descend a binary decision tree from the root to leaves as refining the classification

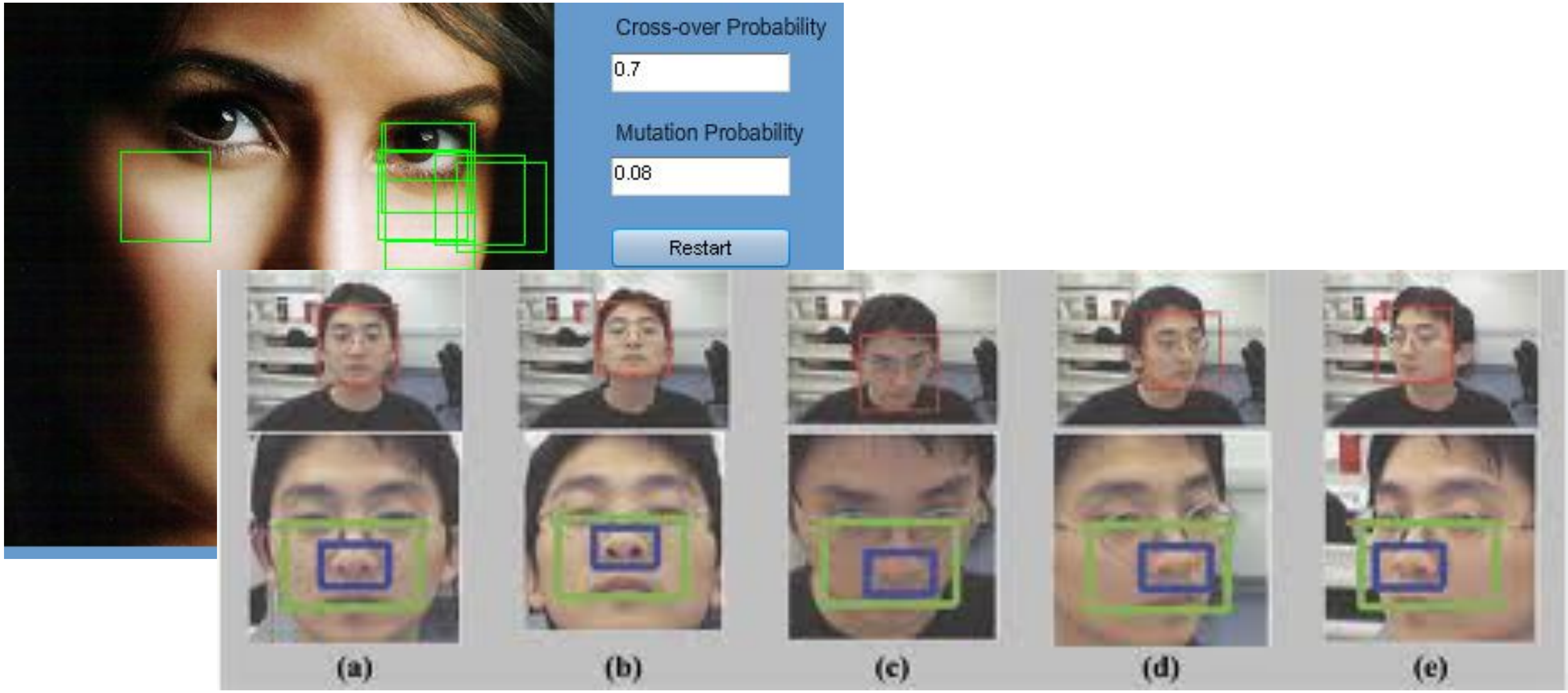
Decision Tree

- *"20 Questions"*
- Build a tree by dividing the data into two sets recursively, until remains only one class



Template Matching

- Compare test data with well-prepared template, mainly used for image processing



Type of Classifiers

- Probabilistic & Geometric & Template matching
  - Geometric: Construct decision boundaries that divide feature space into classes
    - k-NN/ Nearest Mean (NM): geometrical distance between feature vectors of from different classes
    - Support Vector Machine (SVM): construct boundaries maximizing the margins between nearest features relative to two distinct classes

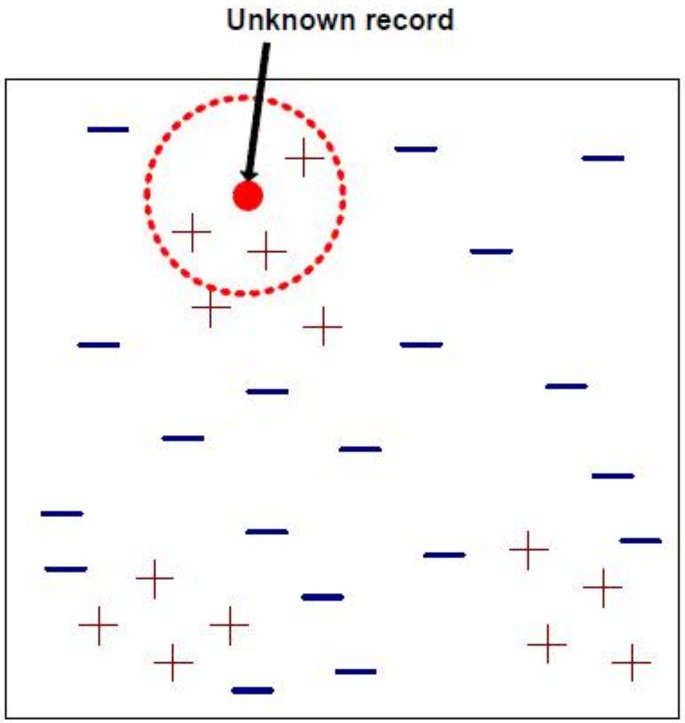
k-NN

- k Nearest Neighbor
- The simplest learning algorithm
- Lazy Learning Classifier
  - Given training data, it does nothing (don’t model) until test data is given
  - cf. Eager Learning Classifier: decision tree, rule-based classifier

k-NN

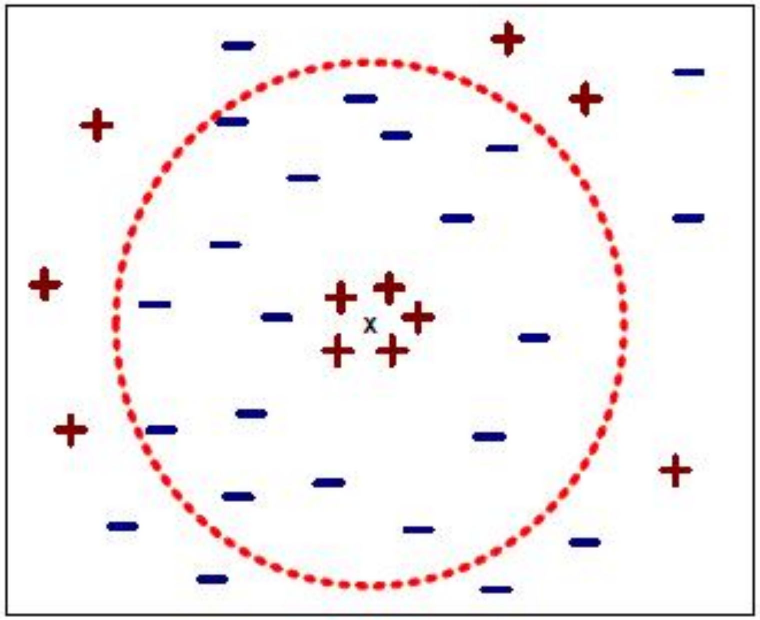
- *"If it walks like a duck, quacks like a duck, and looks like a duck, then it's probably a duck"*

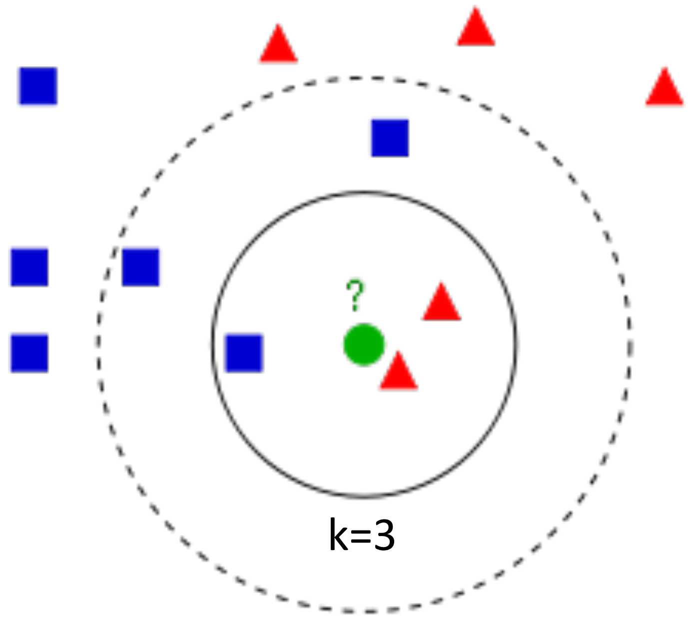
  - Plot each training data in space
  - Given test data, compute the k nearest training data
  - Test data is classified to the same class of majority of k-NN



kNN: Choosing K

- Not too small, not too large

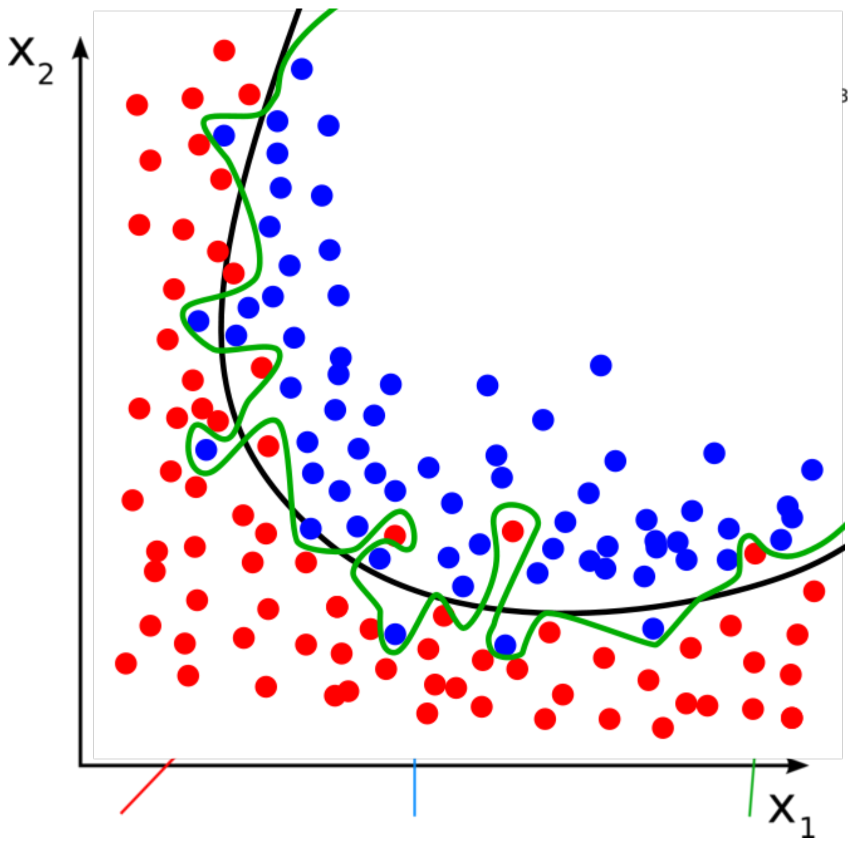




SVM(Support Vector Machine)

- Binary classification

- Find a hyperplane that separates two classes
- (evenly) maximize the distance to the nearest training data of any class (margin)



Type of Classifiers

- Probabilistic & Geometric & Template matching
  - Probabilistic: feature vector $\boldsymbol{x}$ is classified to class C_{i*} if class-conditional PDF $p(\boldsymbol{x}|C_i)$ is maximized for $i=1, \ldots, C$
    - Optimal Bayesian classifier
    - Since class-conditional pdf is unknown, use suboptimal
      - naïve Bayesian, Logistic, Parzen, Gaussian Mixture Model (GMM)

Sequential Classifiers

- So far, single-frame classifiers
- Now, Sequential classifier
  - Exploit decisions made in the past
  - Composite activity is a chain of primitive activities
- Model a composite activity as a first-order Markov chain