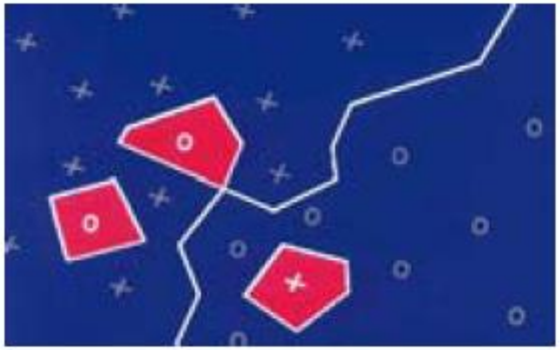
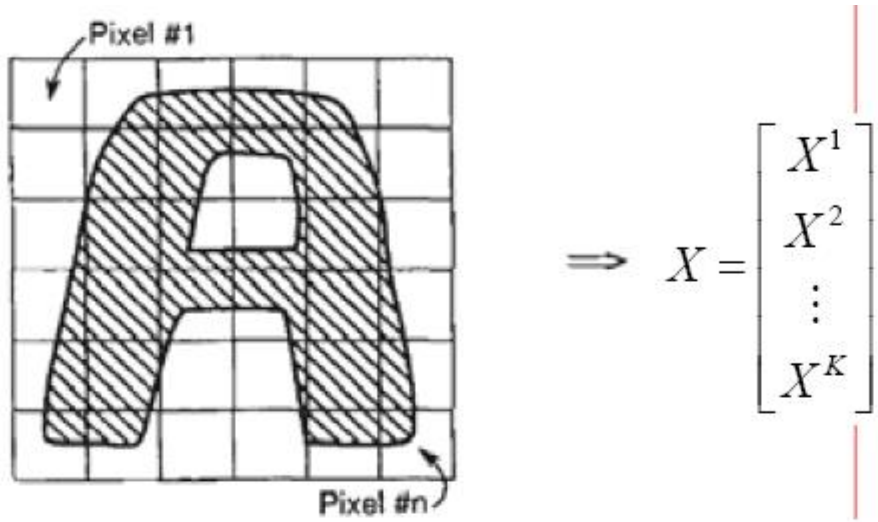
Machine Learning

Theory of Classification
and
Nonparametric Classifier

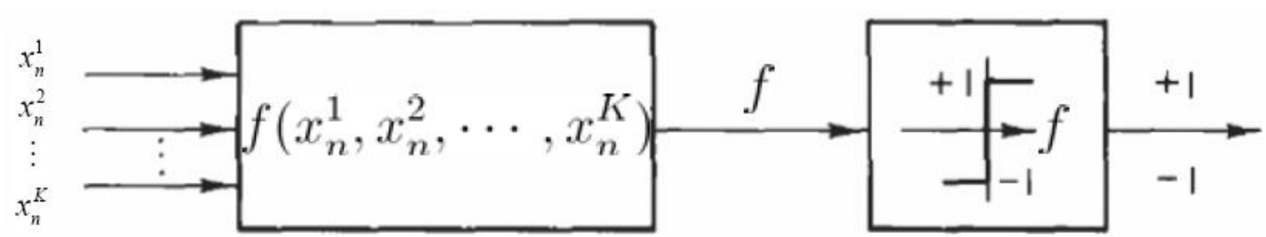


Classification

- Representing data:



- Hypothesis (classifier)



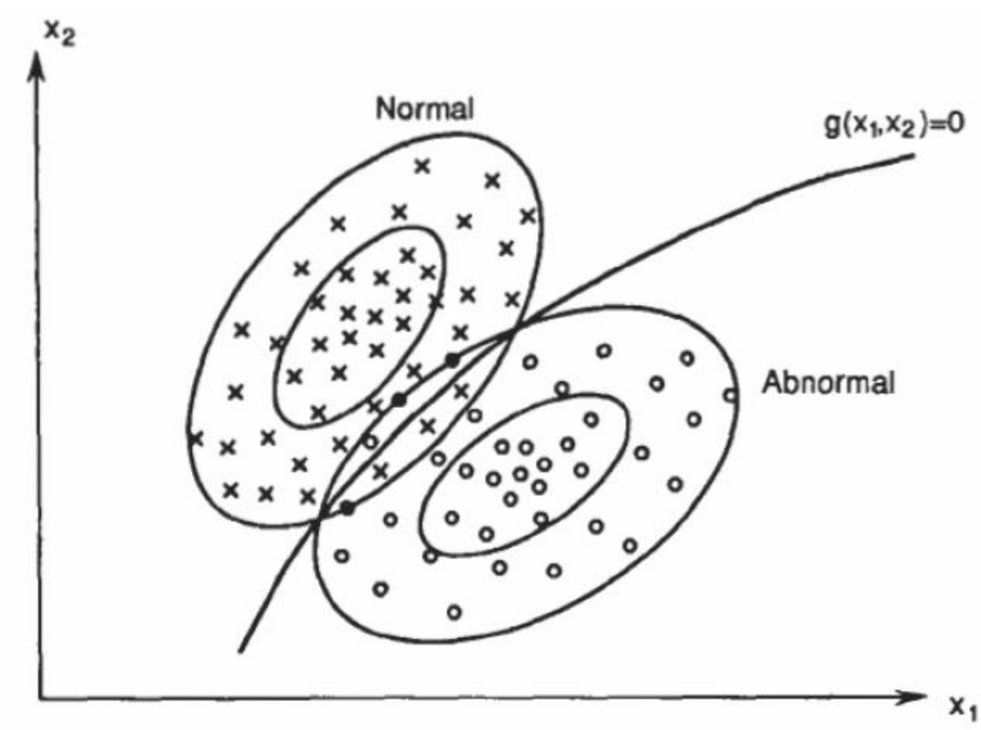
Outline

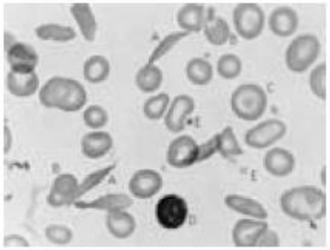
- What is theoretically the best classifier
  - Probabilistic theory of classification
  - Discrete density estimation and Bayesian theorem
  - Bayesian decision rule for Minimum Error

- Nonparametric Classifier (Instance-based learning)
  - Nonparametric density estimation
  - K-nearest-neighbor classifier(KNN)
  - Optimality of kNN
  - Problem of kNN

Decision-making as dividing a high-dimensional space

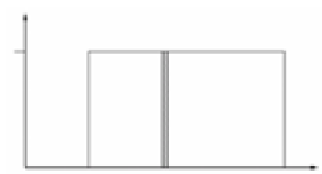
- Distributions of samples from normal and abnormal machine





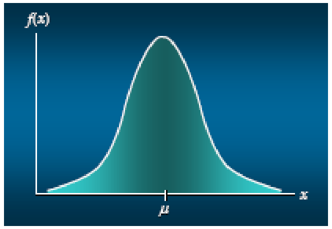
Continuous Distributions

- Uniform Probability Density Function



$$p(x) = 1/(b-a) \quad \text{for } a \le x \le b$$
$$= 0 \quad \text{elsewhere}$$

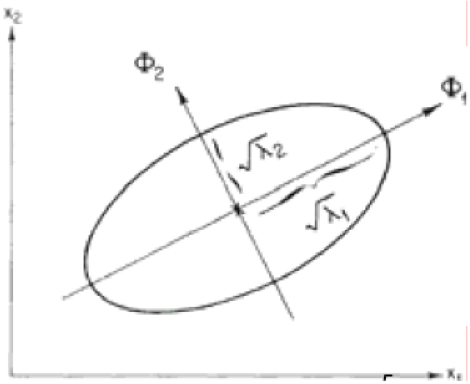
- Normal (Gaussian) Probability Density Function



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

  - The distribution is symmetric, and is often illustrated as a bell-shaped curve.
  - Two parameters, μ (mean) and σ (standard deviation), determine the location and shape of the distribution.
  - The highest point on the normal curve is at the mean, which is also the median and mode.
  - The mean can be any numerical value: negative, zero, or positive.

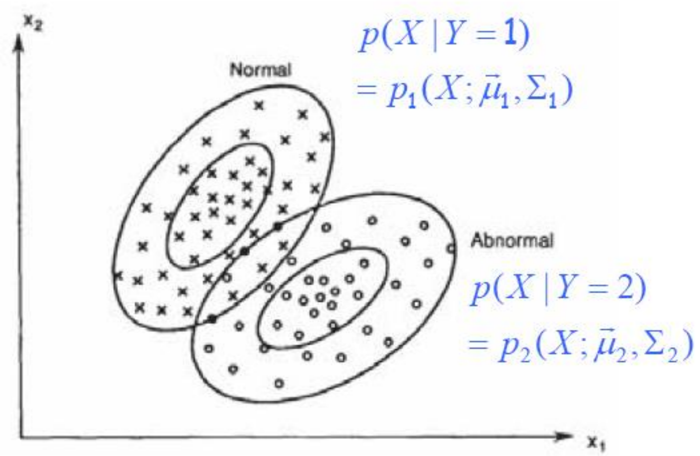
- Multivariate Gaussian

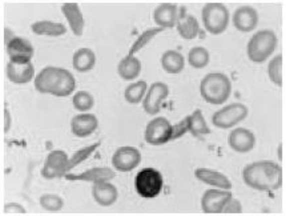


$$p(X;\vec{\mu},\Sigma) = \frac{1}{\left(\sqrt{2\pi}\right)^{n/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(X-\vec{\mu})^T \Sigma^{-1}(X-\vec{\mu})\right\}$$

Class-Conditional Probability

- Classification-specific Dist.: P(X|Y)





- Class prior (i.e., "weight"): P(Y)

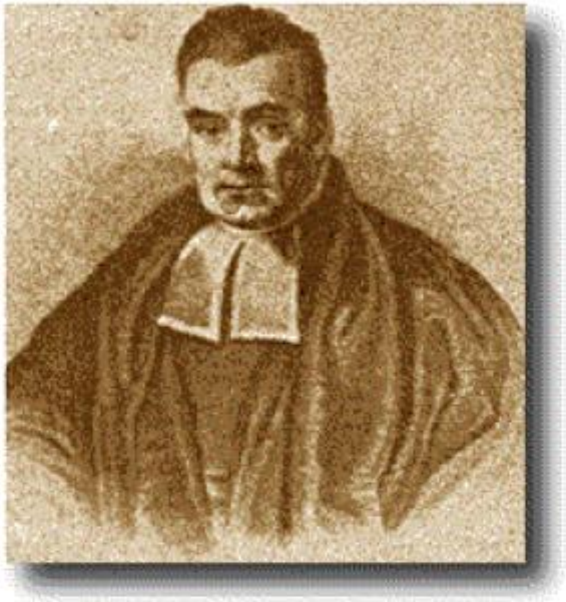
The Bayes Rule

- What we have just did leads to the following general expression:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)}$$

This is Bayes Rule

Bayes, Thomas (1763) An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions of the Royal Society of London*, **53:370-418**



The Bayes Decision Rule for Minimum Error

- The a posteriori probability of a sample

$$P(Y=i \mid X) = \frac{p(X \mid Y=i)P(Y=i)}{p(X)} = \frac{\pi_i p_i(X)}{\sum_i \pi_i p_i(X)} \equiv q_i(X)$$

- Bayes Test:

- Likelihood Ratio:

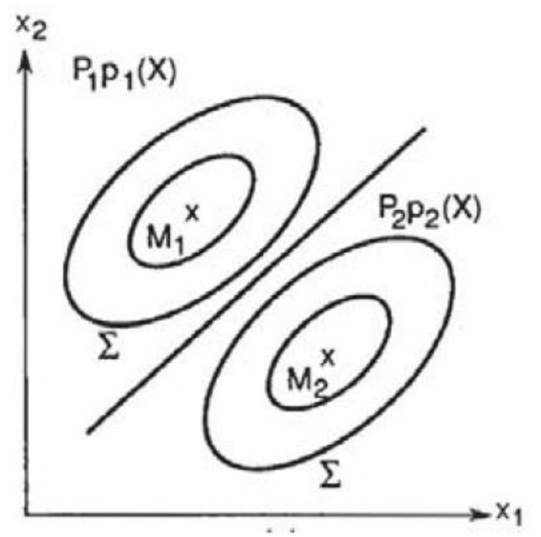
$$\ell(X) =$$

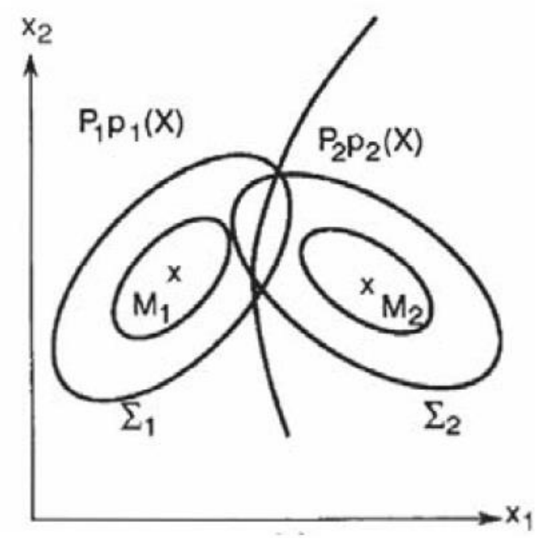
- Discriminant function:

$$h(X) =$$

Example of Decision Rules

- When each class is a normal …





- We can write the decision boundary analytically in some cases … homework!!

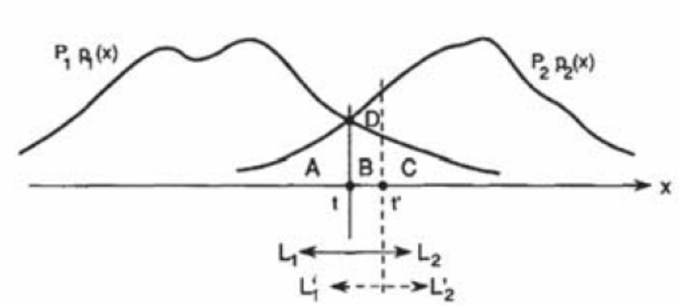
Bayes Error

- We must calculate the *probability of error*
  - the probability that a sample is assigned to the wrong class
- Given a datum X, what is the *risk*?

$$r(X) = \min[q_1(X), q_2(X)]$$

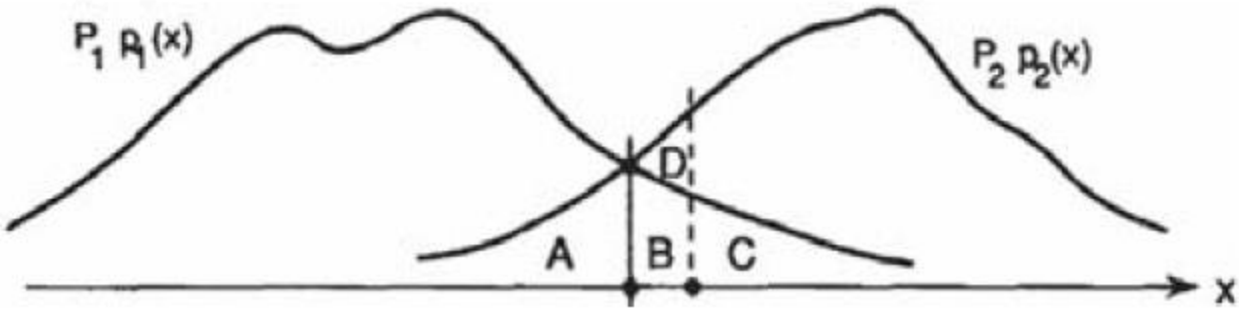
- The Bayes error (the expected risk):

$$\begin{aligned}
\epsilon &= E[r(X)] = \int r(x)p(x)dx \\
&= \int \min[\pi_i p_1(x), \pi_2 p_2(x)]dx \\
&= \pi_1 \int_{L_1} p_1(x)dx + \pi_2 \int_{L_2} p_2(x)dx \\
&= \pi_1 \epsilon_1 + \pi_2 \epsilon_2
\end{aligned}$$



More on Bayes Error

- Bayes error is the lower bound of probability of classification error



- Bayes classifier is the theoretically best classifier that minimizes probability of classification error
- Computing Bayes error is in general a very complex problem. Why?
  - Density estimation:
  - Integrating density function:

$$\epsilon_1 = \int_{\ln(\pi_1/\pi_2)}^{+\infty} p_1(x)dx \qquad \epsilon_2 = \int_{-\infty}^{\ln(\pi_1/\pi_2)} p_2(x)dx$$

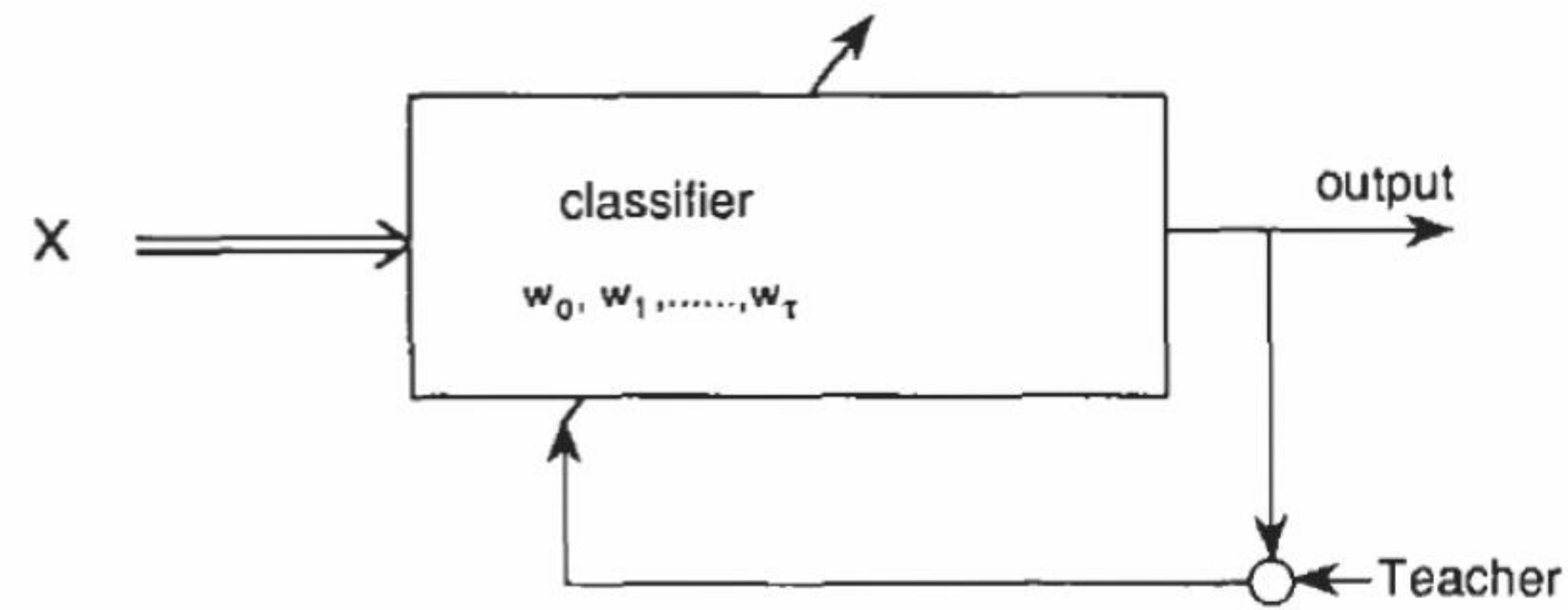
Learning Classifier

- The decision rule:

$$h(X) = -\ln p_1(X) + \ln p_2(X) \underset{<}{\overset{>}{}} \ln \frac{\pi_1}{\pi_2}$$

- Learning strategies
  - Generative Learning
    - Parametric
    - Nonparametric
  - Discriminative Learning
    - Parametric
    - Nonparametric
  - Instance-based Learning (Store all past experience in memory)
    - A special case of nonparametric classifier

Supervised Learning



- K-Nearest-Neighbor Classifier:

  where the h(X) is represented by all the data, and by an algorithm

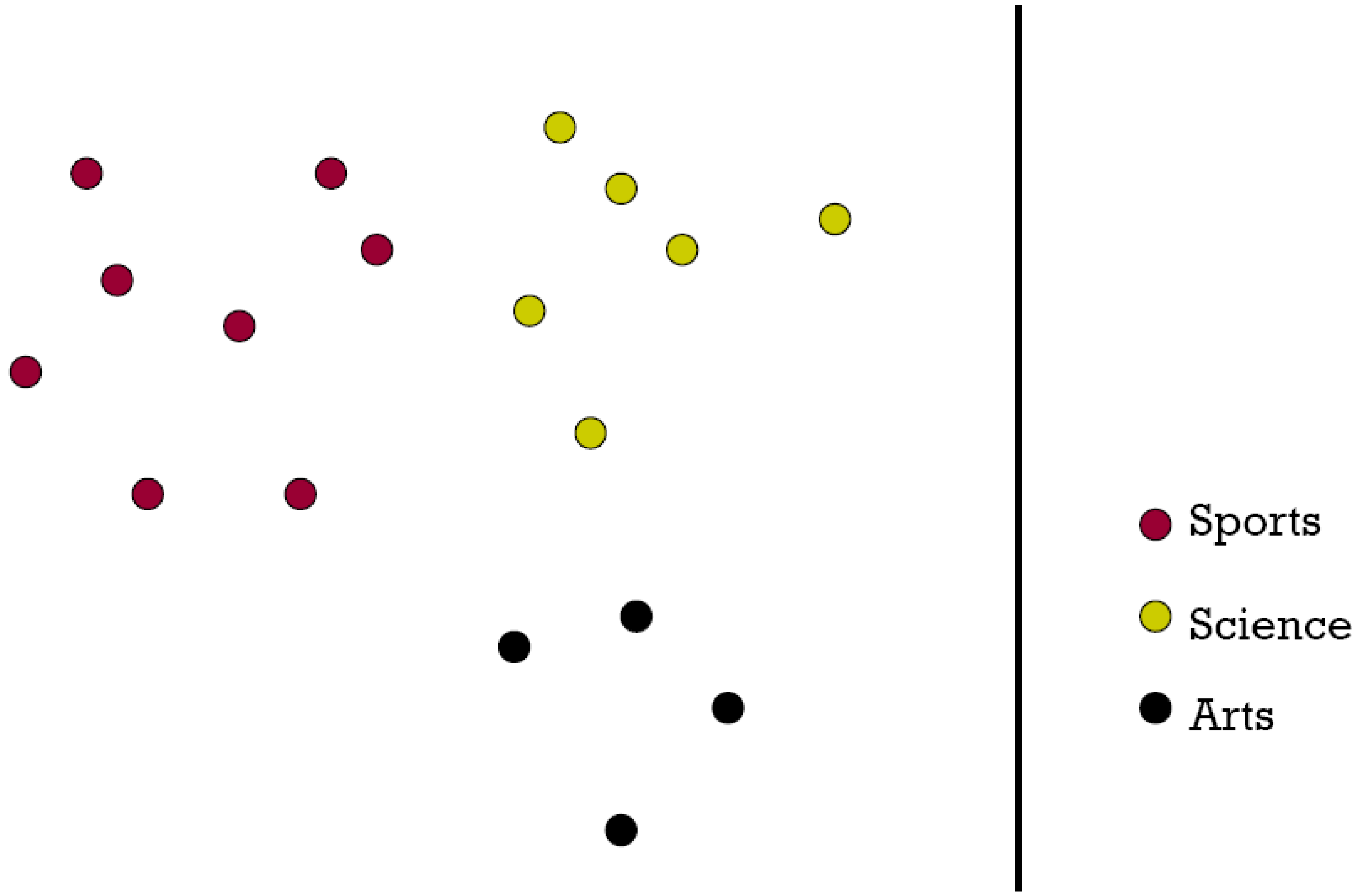
Recall: Vector Space Representation

- Each document is a vector, one component for each term (=word).

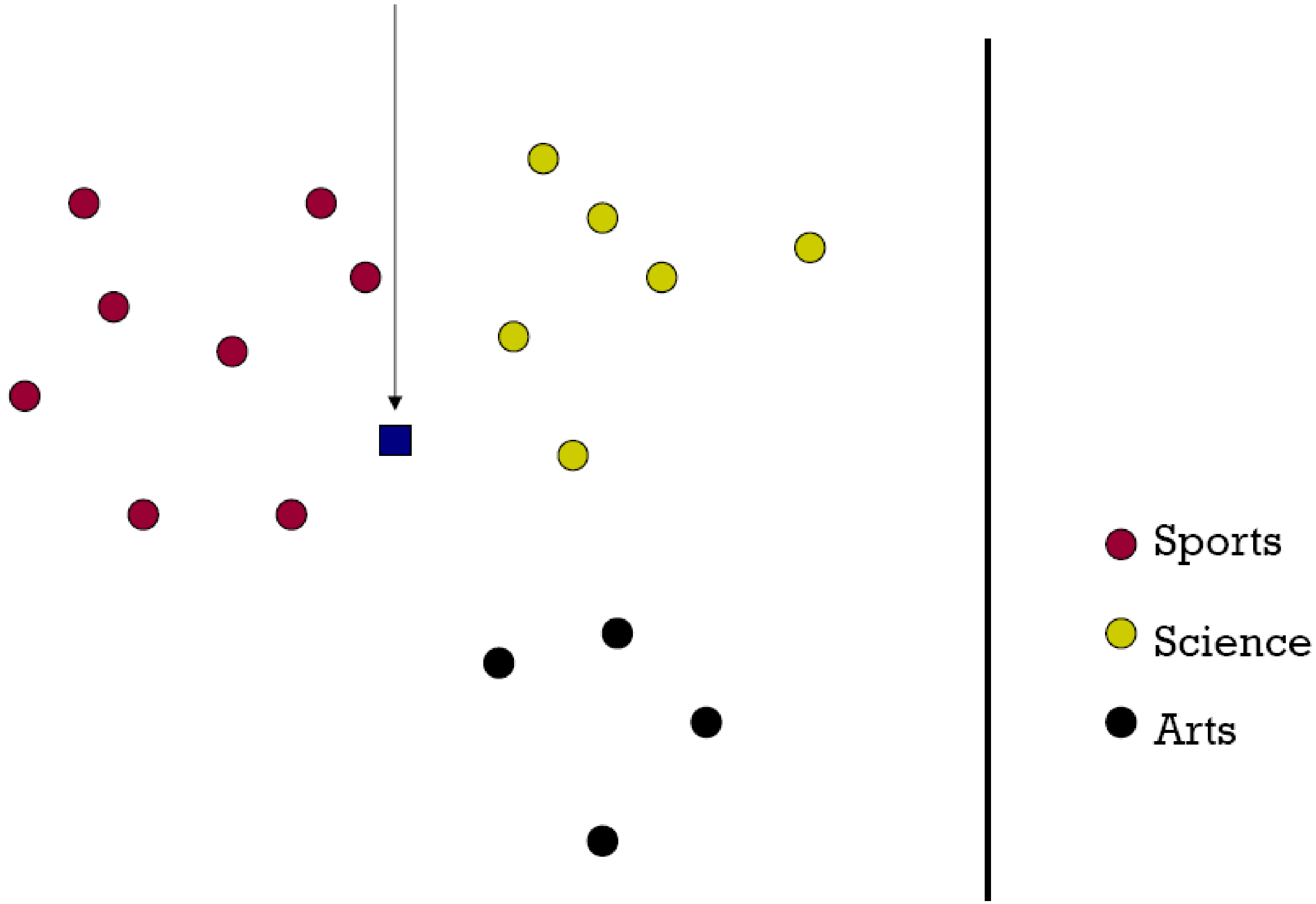
	Doc 1	Doc 2	Doc 3	...
Word 1	3	0	0	...
Word 2	0	8	1	...
Word 3	12	1	10	...
...	0	1	3	...
...	0	0	0	...

- Normalize to unit length.
- High-dimensional vector space:
  - Terms are axes, 10,000+ dimensions, or even 100,000+
  - Docs are vectors in this space

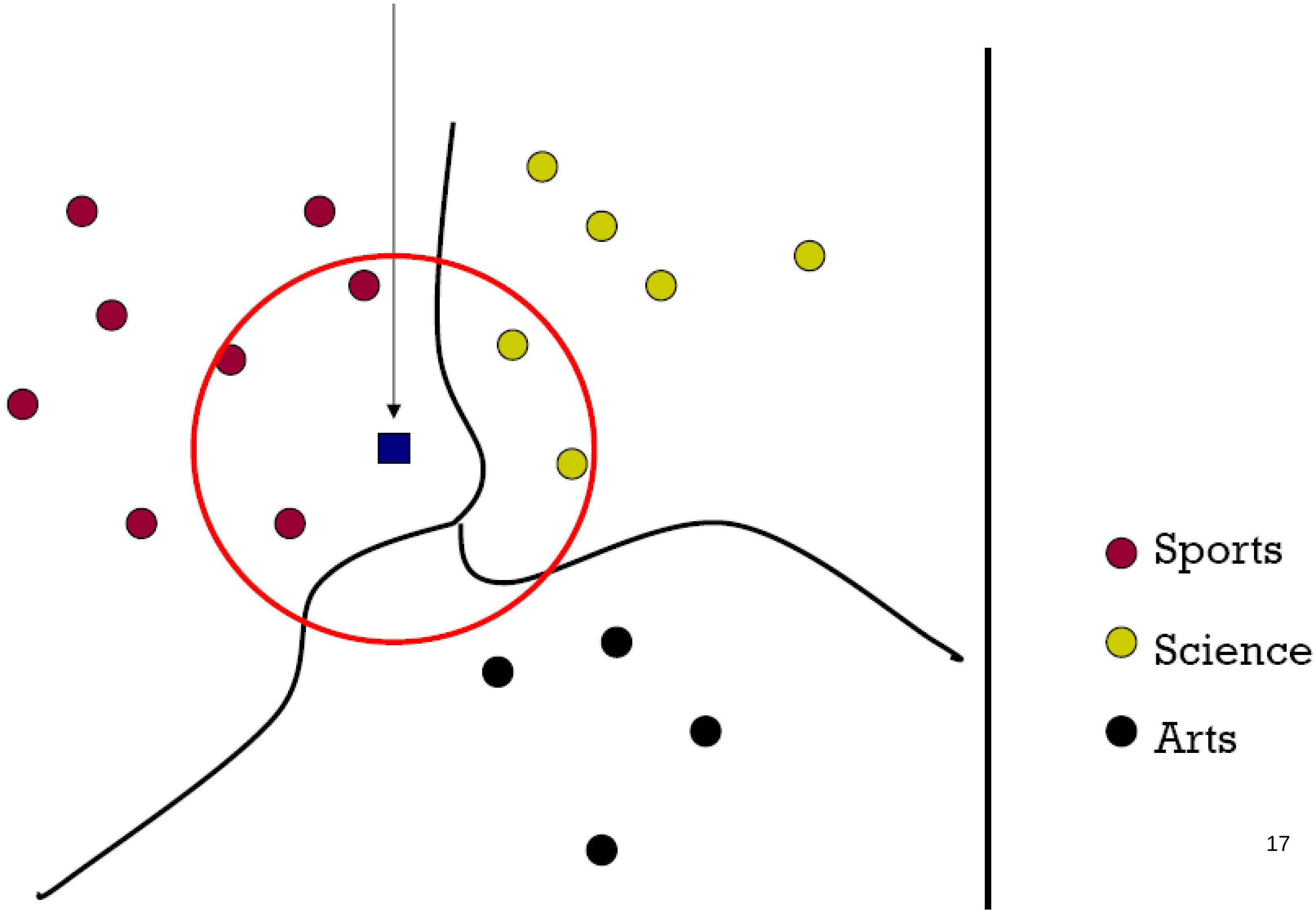
Classes in a Vector Space



Test Document = ?

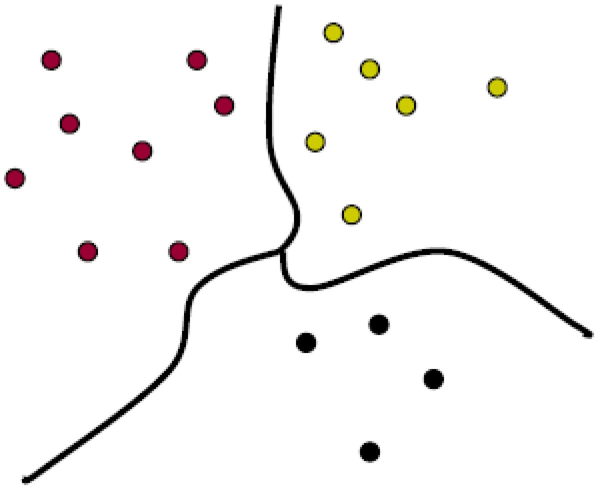


K-Nearest Neighbor (kNN) classifier



kNN Is Close to Optimal

- Cover and Hart 1967
- Asymptotically, the error rate of 1-nearest-neighbor classification is less than twice the Bayes rate [error rate of classifier knowing model that generated data]
- In particular, asymptotic error rate is 0 if Bayes rate is 0.
- Decision boundary:

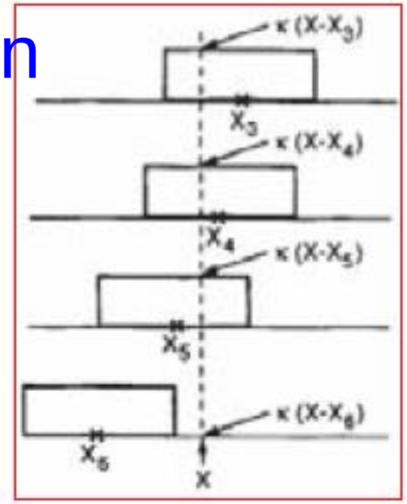


Where does kNN come from?

- How to estimation *p*(*X*) ?
- Nonparametric density estimation
  - Parzen density estimate

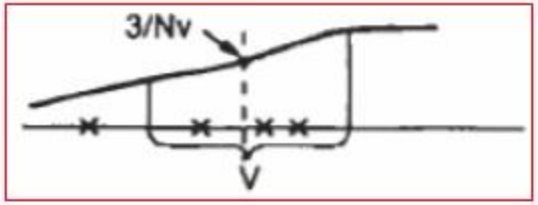
E.g. (Kernel density est.):

$$\hat{p}(X) = \frac{1}{N}\sum_{i=1}^{N}\kappa(X - x_i)$$



More generally:

$$\hat{p}(X) = \frac{1}{N}\frac{k(X)}{V}$$



Where does kNN come from?

- Nonparametric density estimation
  - Parzen density estimate $$\hat{p}(X) = \frac{1}{N}\frac{k(X)}{V}$$
  - kNN density estimate $$\hat{p}(X) = \frac{1}{N}\frac{(k-1)}{V(X)}$$
- Bayes classifier based on kNN density estimator:

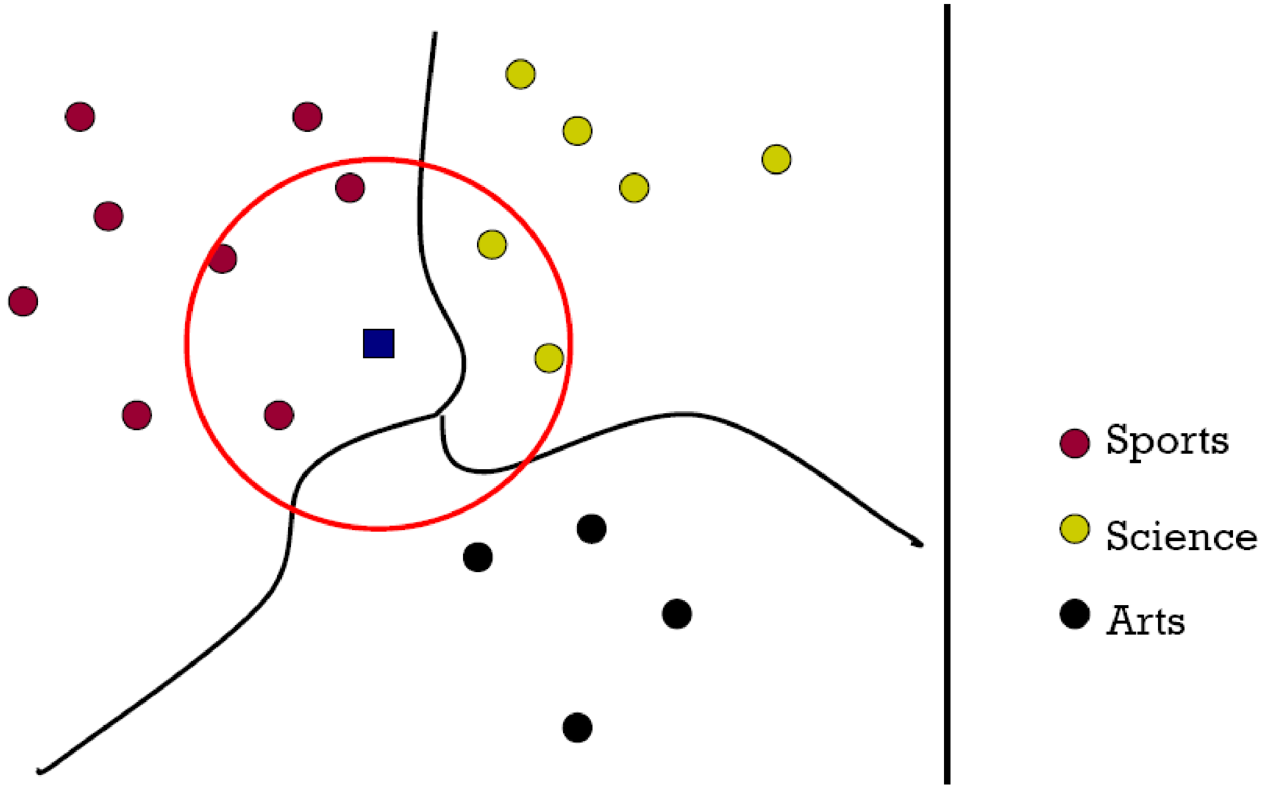
$$h(X) \quad = \quad -\ln\frac{p_1(X)}{p_2(X)} = -\ln\frac{(k_1-1)N_2V_2(X)}{(k_2-1)N_1V_1(X)} \gtrless \ln\frac{\pi_1}{\pi_2}$$

  - Voting kNN classifier

    Pick K_1 and K_2 implicitly by picking $K_1+K_2=K$, $V_1=V_2$, $N_1=N_2$

Voting kNN

- The procedure



kNN is an instance of Instance-Based Learning

- What makes an Instance-Based Learner?
  - A distance metric
  - How many nearby neighbors to look at?
  - A weighting function (optional)
  - How to relate to the local points?

Euclidean Distance Metric

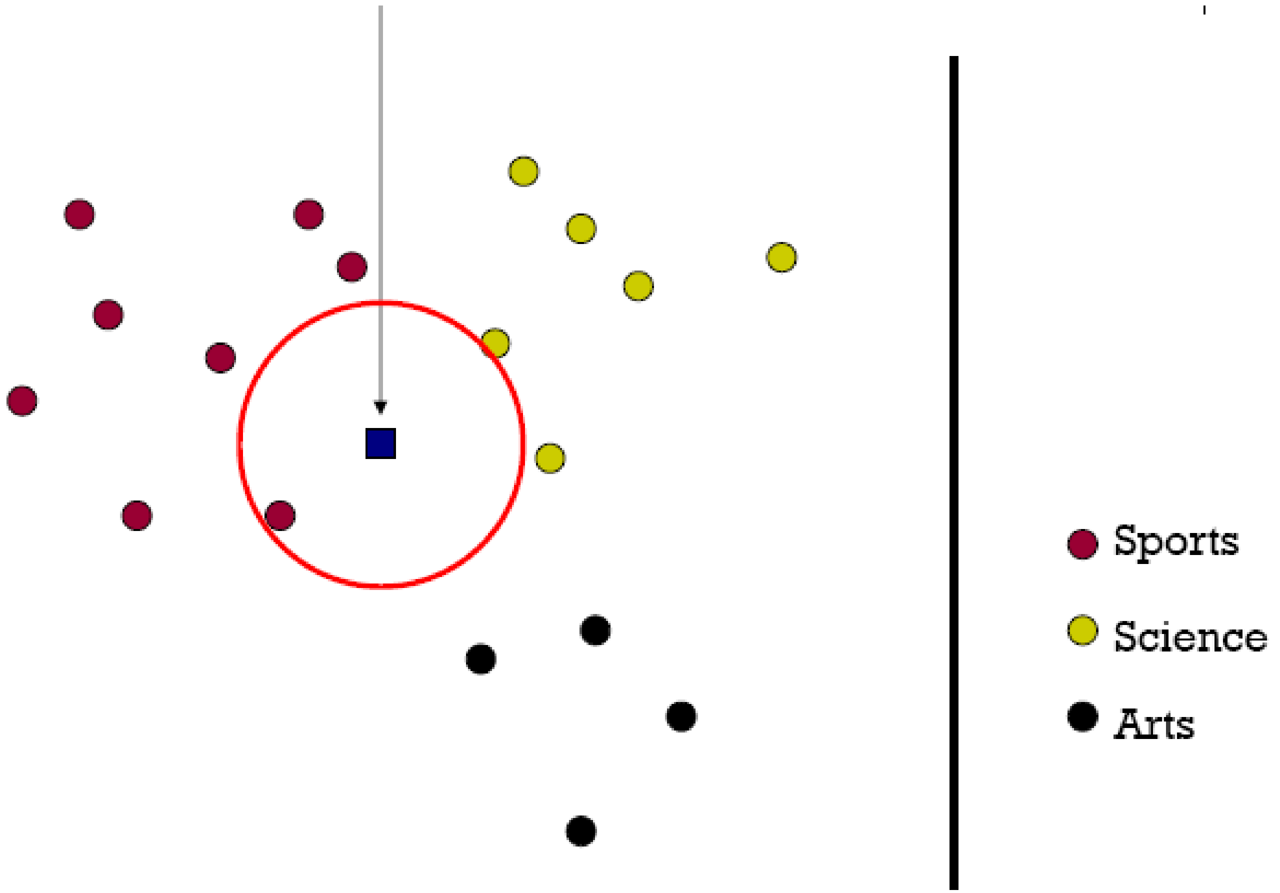
$$D(x,x') = \sqrt{\sum_i \sigma_i^2 (x_i - x_i')^2}$$

- Or equivalently,

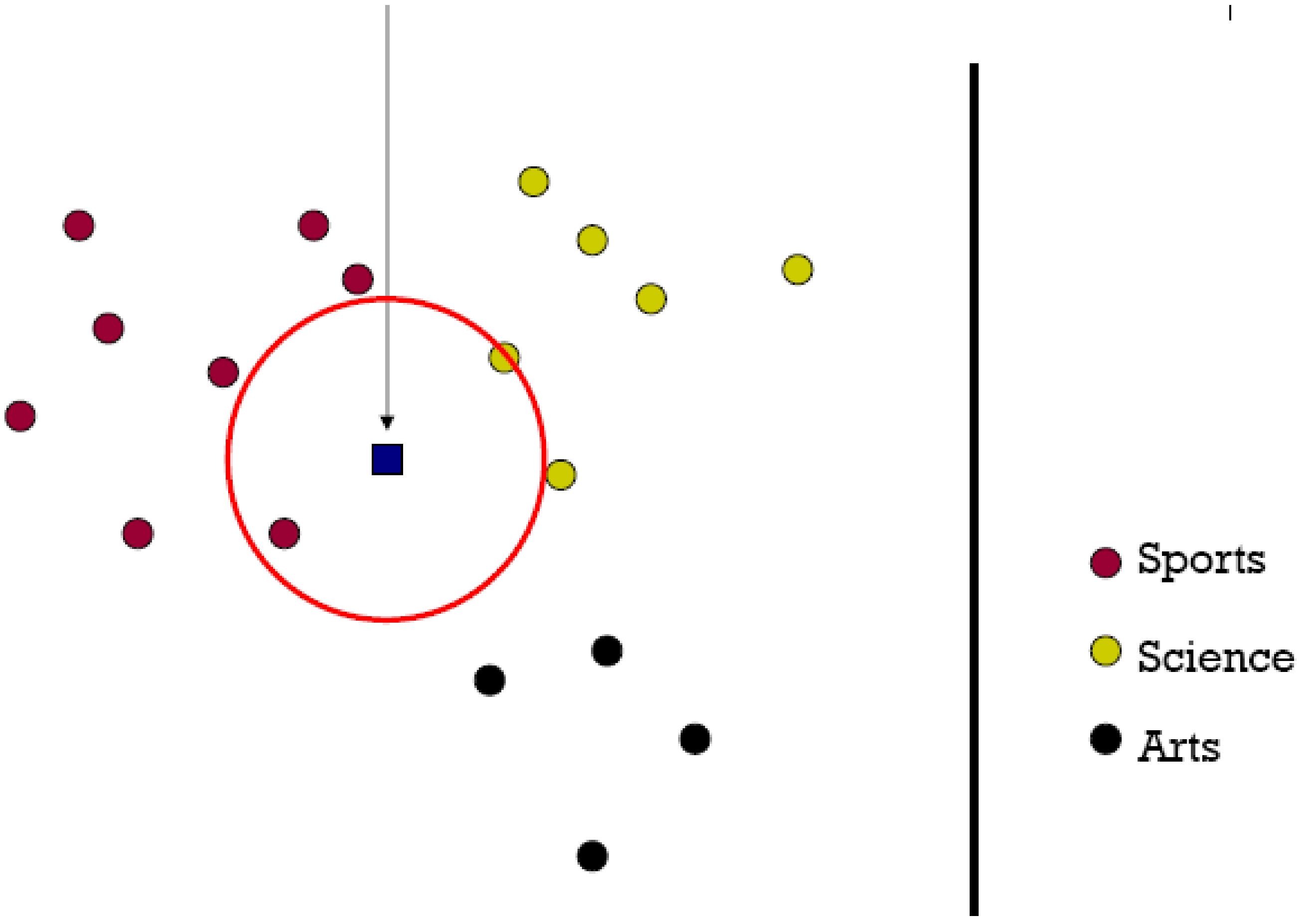
$$D(x,x') = \sqrt{(x-x')^T \Sigma (x-x')}$$

- Other metrics:
  - L1 norm: |x-x'|= $\sum_{i=1}^{n} |x_i - x'|$
  - L∞ norm: max |x-x'| (elementwise …)
  - Mahalanobis: where Σ is full, and symmetric
  - Correlation
  - Angle
  - Hamming distance, Manhattan distance

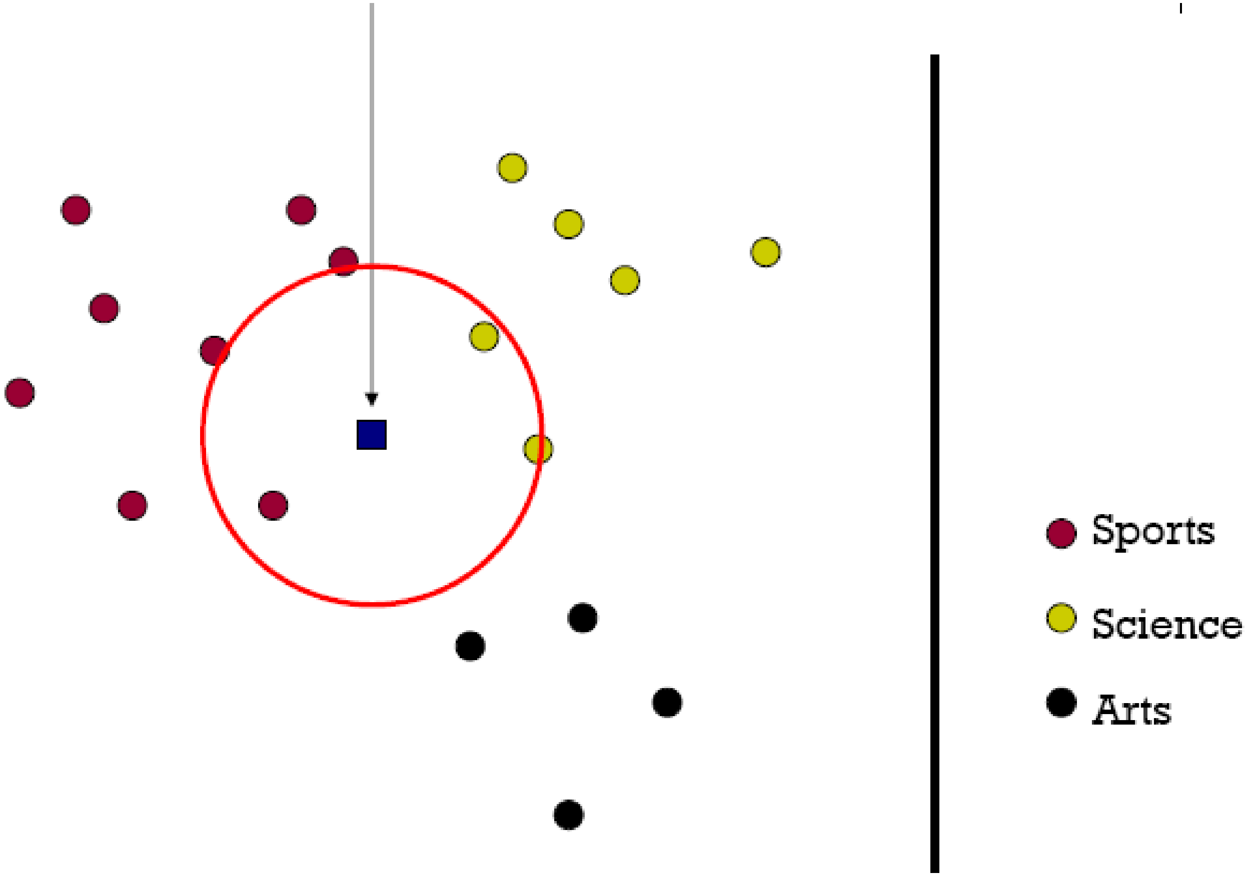
1-Nearest Neighbor (kNN) classifier



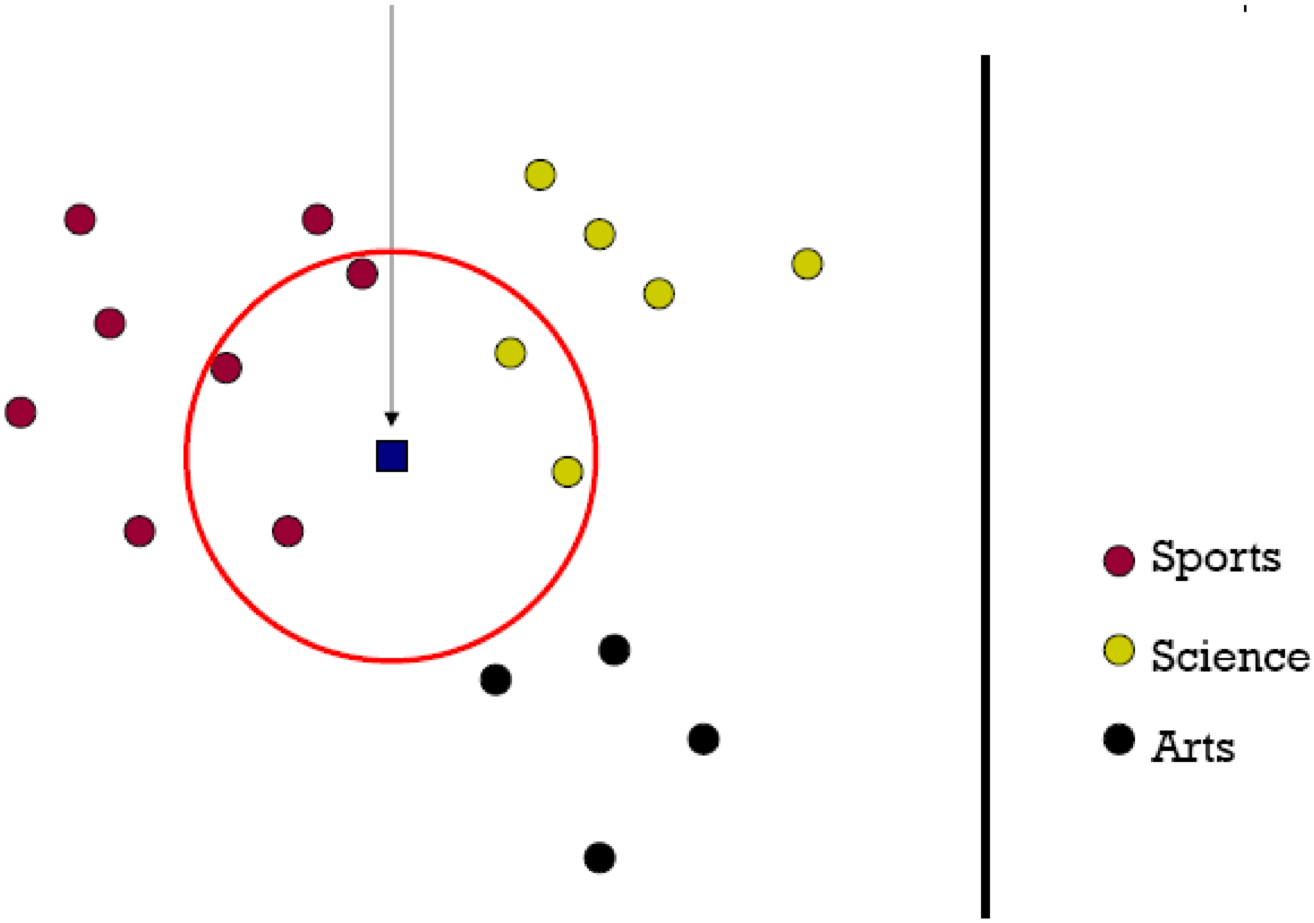
2-Nearest Neighbor (kNN) classifier



3-Nearest Neighbor (kNN) classifier



5-Nearest Neighbor (kNN) classifier



Nearest-Neighbor Learning Algorithm

- Learning is just storing the representations of the training examples in *D*.
- Testing instance *x*:
  - Compute similarity between *x* and all examples in *D*.
  - Assign *x* the category of the most similar example in *D*.
- Does not explicitly compute a generalization or category prototypes.
- Also called:
  - Case-based learning
  - Memory-based learning
  - Lazy learning

Case Study:
kNN for Web Classification

- Dataset
  - **20 News Groups (20 classes)**
  - **Download :(http://people.csail.mit.edu/jrennie/20Newsgroups/)**
  - **61,118 words, 18,774 documents**
  - **Class labels descriptions**

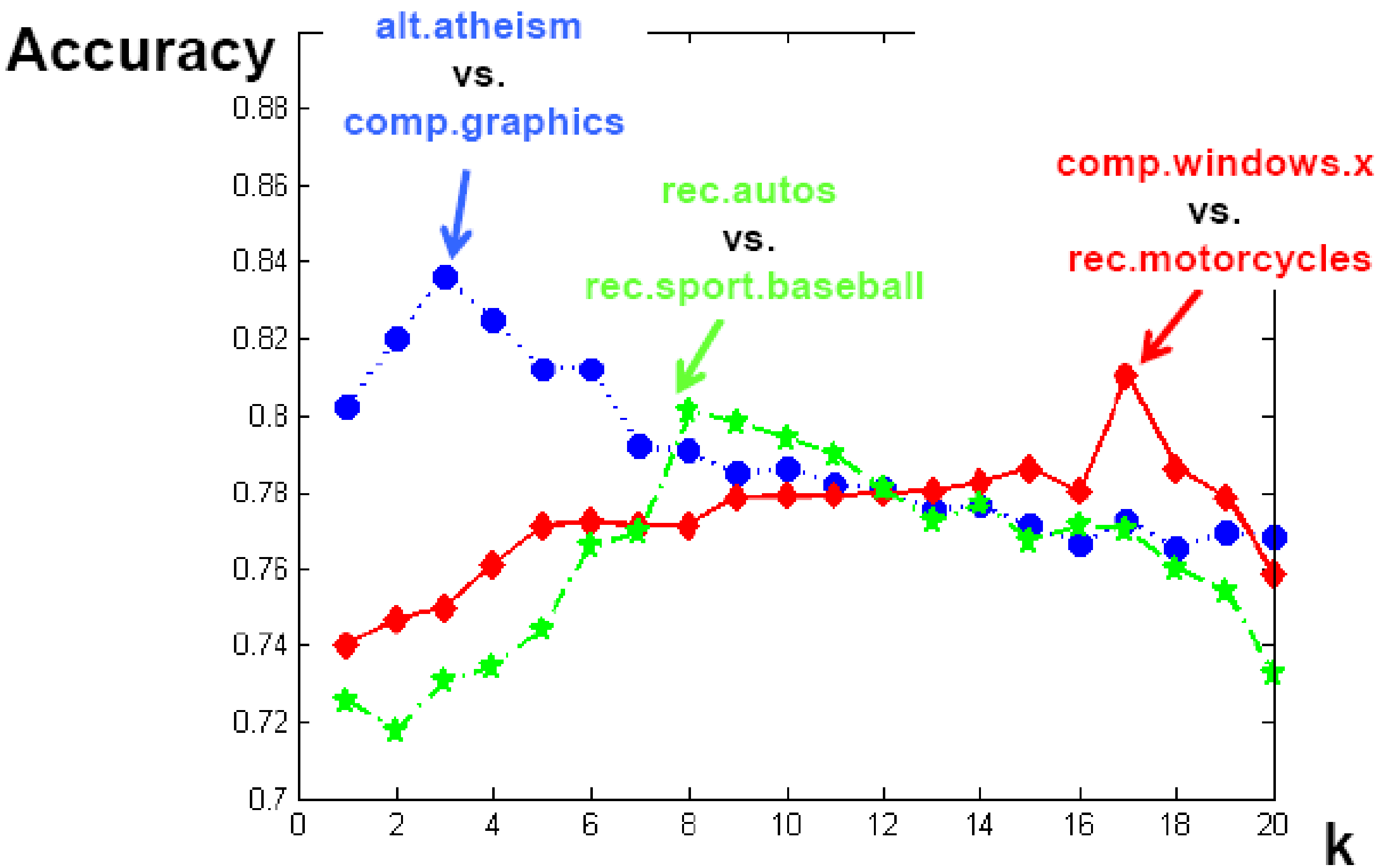
comp.graphics comp.os.ms-windows.misc comp.sys.ibm.pc.hardware comp.sys.mac.hardware comp.windows.x	rec.autos rec.motorcycles rec.sport.baseball rec.sport.hockey	sci.crypt sci.electronics sci.med sci.space
misc.forsale	talk.politics.misc talk.politics.guns talk.politics.mideast	talk.religion.misc alt.atheism soc.religion.christian

Experimental Setup

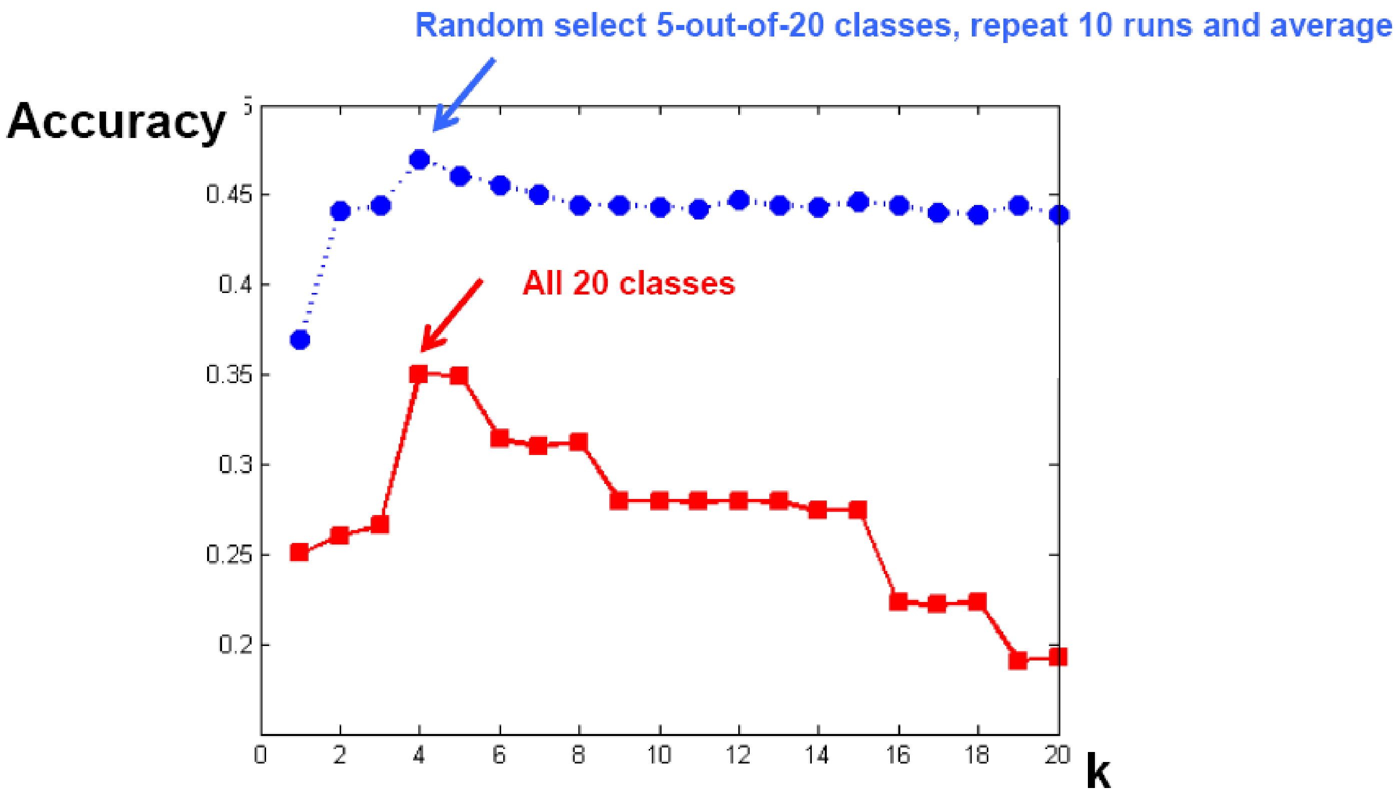
- Training/Test Sets:
  - 50%-50% randomly split.
  - 10 runs
  - report average results
- Evaluation Criteria:

$$Accuracy = \frac{\sum_{i \in test\ set} \mathrm{I}(predict_i = true\ label_i)}{\#\ of\ test\ samples}$$

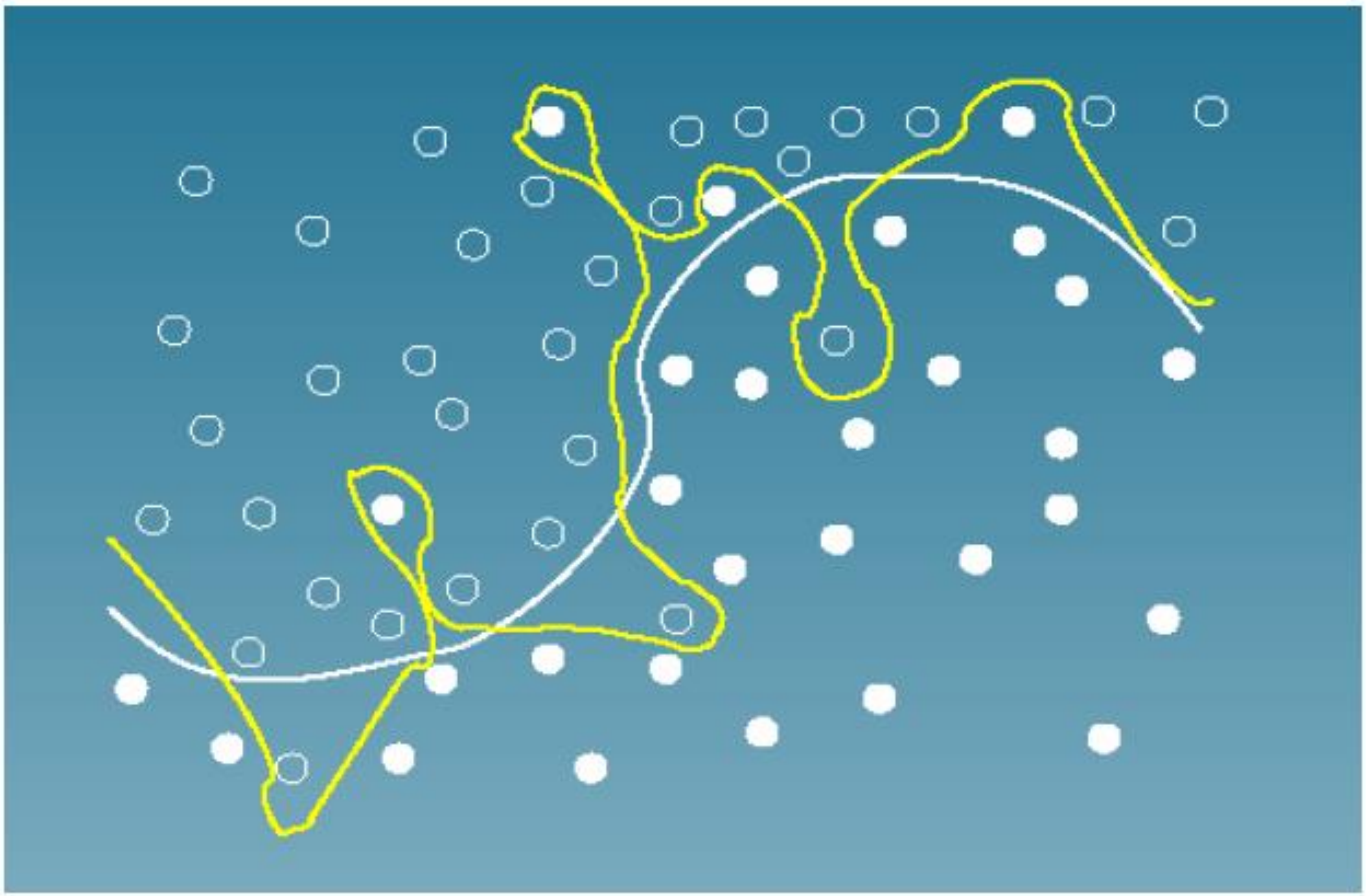
Results: Binary Classes



Results: Multiple Classes



Is kNN ideal? … more later



Effect of Parameters

- Sample size
  - The more the better
  - Need efficient search algorithm for NN
- Dimensionality
  - Curse of dimensionality
- Density
  - How smooth?
- Metric
  - The relative scalings in the distance metric affect region shapes.
- Weight
  - Spurious or less relevant points need to be downweighted
- K

Summary

- ***Bayes classifier*** is the best classifier which minimizes the probability of classification error.
- Nonparametric and parametric classifier
- A nonparametric classifier does not rely on any assumption concerning the structure of the underlying density function.
- A classifier becomes the ***Bayes classifier*** if the density estimates converge to the true densities
    - when an infinite number of samples are used
    - The resulting error is the ***Bayes error,*** the smallest achievable error given the underlying distributions.