Machine Learning

10-701/15-781, Spring 2008

Theory of Classification and Nonparametric Classifier

Eric Xing

Lecture 2, January 16, 2006

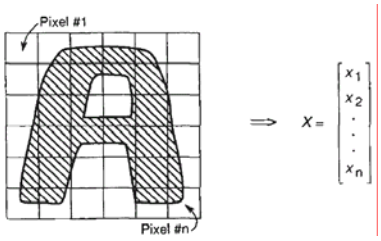
Reading: Chap. 2,5 CB and handouts

Outline

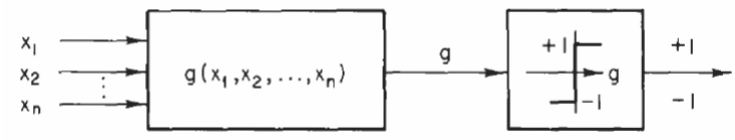
- What is theoretically the best classifier
- Bayesian decision rule for Minimum Error
- Nonparametric Classifier (Instance-based learning)
  - Nonparametric density estimation
  - K-nearest-neighbor classifier
  - Optimality of kNN

Classification

- Representing data:

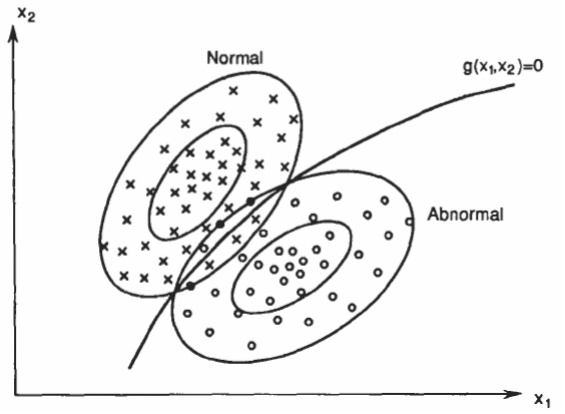


- Hypothesis (classifier)



Decision-making as dividing a high-dimensional space

- Distributions of samples from normal and abnormal machine



Basic Probability Concepts

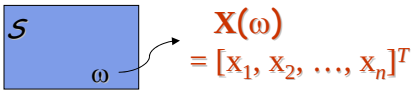
- A *sample space* $\mathcal{S}$ is the set of all possible outcomes of a conceptual or physical, repeatable experiment. ($\mathcal{S}$ can be finite or infinite.)
  - E.g., $\mathcal{S}$ may be the set of all possible outcomes of a dice roll: $\mathcal{S} \equiv \{1,2,3,4,5,6\}$

  - E.g., $\mathcal{S}$ may be the set of all possible nucleotides of a DNA site: $\mathcal{S} \equiv \{\mathrm{A},\mathrm{T},\mathrm{C},\mathrm{G}\}$

  - E.g., $\mathcal{S}$ may be the set of all possible positions time-space positions of a aircraft on a radar screen: $\mathcal{S} \equiv \{0, R_{\max}\} \times \{0, 360^{\circ}\} \times \{0, +\infty\}$

Random Variable

- A *random variable* is a function that associates a unique numerical value (a token) with every outcome of an experiment. (The value of the r.v. will vary from trial to trial as the experiment is repeated)



  - Discrete r.v.:
    - The outcome of a dice-roll
    - The outcome of reading a nt at site *i*: X_i
  - Binary event and indicator variable:
    - Seeing an "A" at a site $\Rightarrow X$=1, o/w X=0.
    - This describes the true or false outcome a random *event*.
    - Can we describe richer outcomes in the same way? (i.e., X=1, 2, 3, 4, for being A, C, G, T) --- think about what would happen if we take expectation of X.
  - Unit-Base Random vector

    X_i=[X_{iA}, X_{iT}, X_{iG}, X_{iC}]', X_i=[0,0,1,0]' $\Rightarrow$ seeing a "G" at site *i*
  - Continuous r.v.:
    - The outcome of **recording** the **true** location of an aircraft: X_{true}
    - The outcome of **observing** the **measured** location of an aircraft X_{obs}

Discrete Prob. Distribution

- (In the discrete case), a probability distribution P on $\mathcal{S}$ (and hence on the domain of X) is an assignment of a non-negative real number $P(s)$ to each $s \in \mathcal{S}$ (or each valid value of x) such that $\Sigma_{s \in \mathcal{S}} P(s)=1$. $(0 \le P(s) \le 1)$
  - intuitively, $P(s)$ corresponds to the *frequency* (or the likelihood) of getting s in the experiments, if repeated many times
  - call $\theta_s = P(s)$ the *parameters* in a discrete probability distribution
- A probability distribution on a sample space is sometimes called a *probability model*, in particular if several different distributions are under consideration
  - write models as M_1, M_2, probabilities as $P(X|M_1)$, $P(X|M_2)$
  - e.g., M_1 may be the appropriate prob. dist. if X is from "fair dice", M_2 is for the "loaded dice".
  - M is usually a two-tuple of {dist. family, dist. parameters}

Discrete Distributions

- Bernoulli distribution: Ber(p)

$$P(x) = \begin{cases} 1-p & \text{for } x = 0 \\ p & \text{for } x = 1 \end{cases} \quad \Rightarrow \quad P(x) = p^x (1-p)^{1-x}$$

- Multinomial distribution: Mult(1, θ)
  - Multinomial (indicator) variable:

$$X = \begin{bmatrix} X_1 \\ X_2 \\ X_3 \\ X_4 \\ X_5 \\ X_6 \end{bmatrix}, \quad \text{where} \quad X_j = [0,1], \text{ and } \sum_{j \in \{1,\ldots,6\}} X_j = 1$$

$$X_j = 1 \text{ w.p. } \theta_j, \quad \sum_{j \in \{1,\ldots,6\}} \theta_j = 1 \ .$$

$$p(x(j)) = P(\{X_j = 1, \text{where } j \text{ index the dice-face}\})$$
$$= \theta_j = \theta_A^{x_A} \times \theta_C^{x_C} \times \theta_G^{x_G} \times \theta_T^{x_T} = \prod_k \theta_k^{x_k} = \theta^x$$

Discrete Distributions

- Multinomial distribution: Mult(n, θ)
  - Count variable:

$$X = \begin{bmatrix} x_1 \\ \vdots \\ x_K \end{bmatrix}, \qquad \text{where } \sum_j x_j = n$$

$$p(x) = \frac{n!}{x_1! x_2! \cdots x_K!} \theta_1^{x_1} \theta_2^{x_2} \cdots \theta_K^{x_K} = \frac{n!}{x_1! x_2! \cdots x_K!} \theta^x$$

"Arts" "Budgets" "Children" "Education"

NEW MILLION CHILDREN SCHOOL
FILM TAX WOMEN STUDENTS
SHOW PROGRAM PEOPLE SCHOOLS
MUSIC BUDGET CHILD EDUCATION
MOVIE BILLION YEARS TEACHERS
PLAY FEDERAL FAMILIES HIGH
MUSICAL YEAR WORK PUBLIC
BEST SPENDING PARENTS TEACHER
ACTOR NEW SAYS BENNETT
FIRST STATE FAMILY MANIGAT
YORK PLAN WELFARE NAMPHY
OPERA MONEY MEN STATE
THEATER PROGRAMS PERCENT PRESIDENT
ACTRESS GOVERNMENT CARE ELEMENTARY
LOVE CONGRESS LIFE HAITI

The William Randolph Hearst Foundation will give $1.25 million to Lincoln Center, Metropolitan Opera Co., New York Philharmonic and Juilliard School. "Our board felt that we had a real opportunity to make a mark on the future of the performing arts with these grants an act every bit as important as our traditional areas of support in health, medical research, education and the social services," Hearst Foundation President Randolph A. Hearst said Monday in announcing the grants. Lincoln Center's share will be $200,000 for its new building, which will house young artists and provide new public facilities. The Metropolitan Opera Co. and New York Philharmonic will receive $400,000 each. The Juilliard School, where music and the performing arts are taught, will get $250,000. The Hearst Foundation, a leading supporter of the Lincoln Center Consolidated Corporate Fund, will make its usual annual $100,000 donation, too.

Continuous Prob. Distribution

- A continuous random variable $\mathbf{X}$ can assume any value in an interval on the real line or in a region in a high dimensional space
  - **A random vector** $\mathbf{X}=[x_1, x_2, \ldots, x_n]^T$ usually corresponds to a real-valued measurements of some property, e.g., length, position, …
  - It is not possible to talk about the probability of the random variable assuming a particular value --- $P(\mathbf{x}) = 0$
  - Instead, we talk about the probability of the random variable assuming a value within a given interval, or half interval
    - $P(X \in [x_1, x_2])$,

      $P(X < x) = P(X \in [-\infty, x])$
    - Arbitrary Boolean combination of basic propositions

Continuous Prob. Distribution

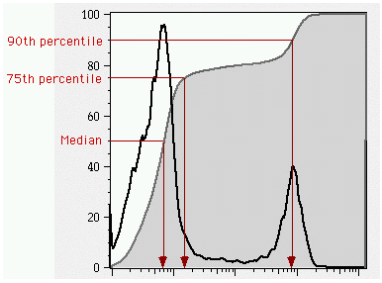
- The probability of the random variable assuming a value within some given interval from x_1 to x_2 is defined to be the area under the graph of the probability density function between x_1 and x_2.
  - Probability mass: $P(X \in [x_1, x_2]) = \int_{x_1}^{x_2} p(x)dx$,

    note that $\int_{-\infty}^{+\infty} p(x)dx = 1.$
  - Cumulative distribution function (CDF):

    $$P(x) = P(X < x) = \int_{-\infty}^{x} p(x')dx'$$
  - Probability density function (PDF):

    $$p(x) = \frac{d}{dx}P(x)$$

    $$\int_{-\infty}^{+\infty} p(x)dx = 1; \quad p(x) > 0, \forall x$$

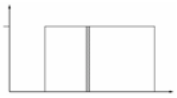


Car flow on Liberty Bridge (cooked up!)

Continuous Distributions

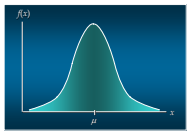
- Uniform Probability Density Function

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



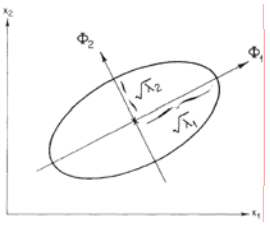
- Normal (Gaussian) Probability Density Function

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



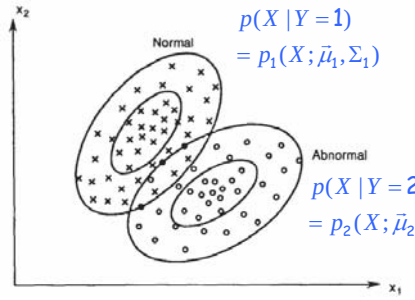
  - The distribution is symmetric, and is often illustrated as a bell-shaped curve.
  - Two parameters, μ (mean) and σ (standard deviation), determine the location and shape of the distribution.
  - The highest point on the normal curve is at the mean, which is also the median and mode.
  - The mean can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

$$P(Y \mid X) = \frac{P(X \mid Y)\, p(Y)}{P(X)}$$

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:

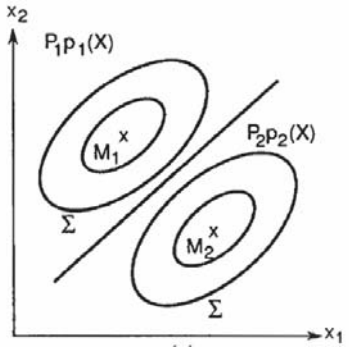
$$\ell(X) =$$

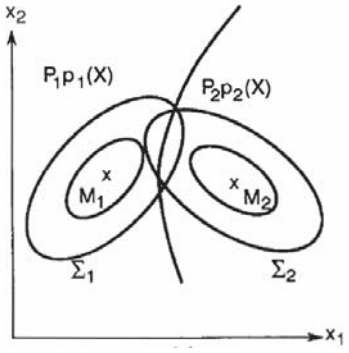
- Discriminant function:

$$h(X) =$$

Example of Decision Rules

- When each class is a normal …





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

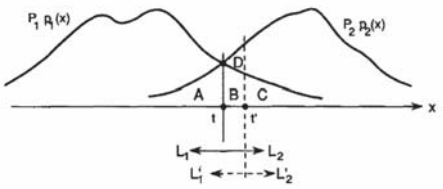
Bayes Error

- 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*?

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

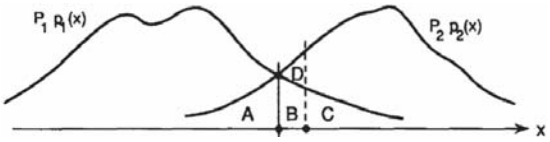
- The Bayes error (the expected risk):

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



More on Bayes Error

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



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

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

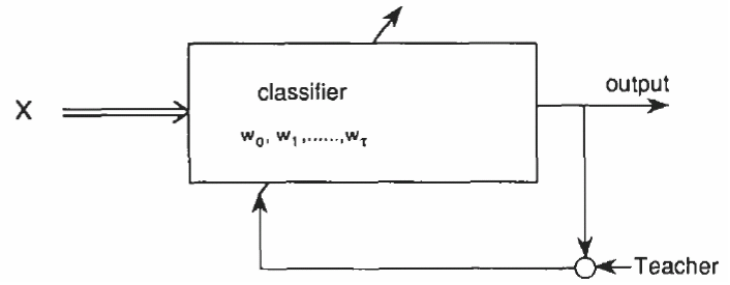
Learning Classifier

- 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
  - Instance-base (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

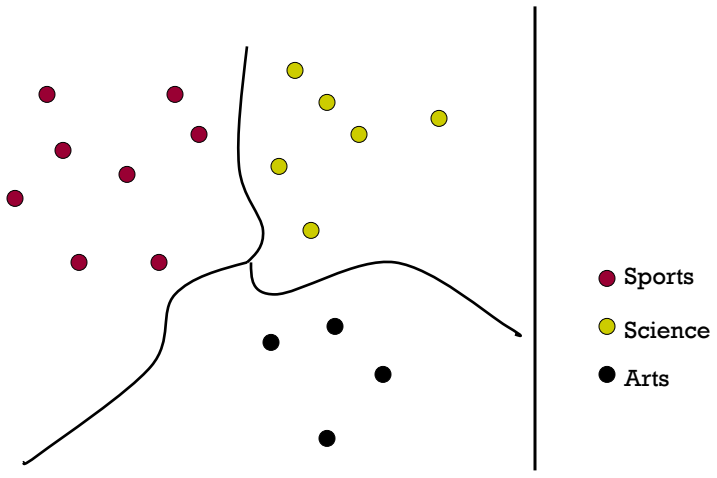
Recall: Vector Space Representation

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

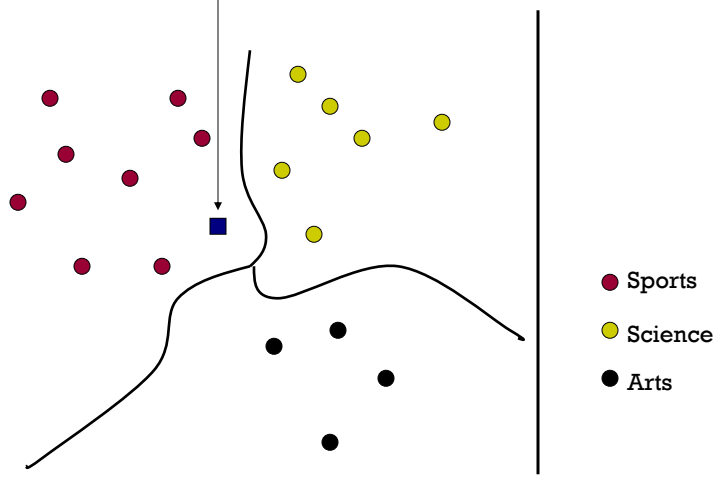
	Doc 1	Doc 2	Doc 3	...
Word 1	3	0	0	...
Word 2	0	8	1	...
Word 3	12	1	10	...
...	0	1	3	...
...	0	0	0	...

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

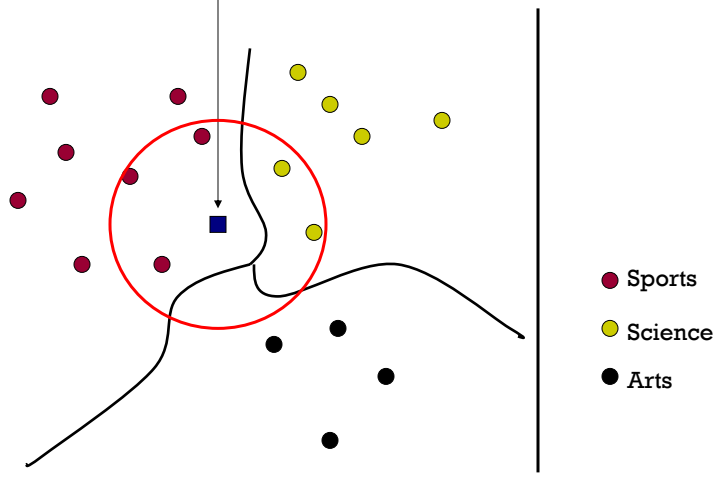
Classes in a Vector Space



Test Document = ?

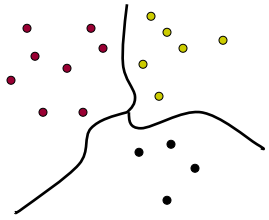


K-Nearest Neighbor (kNN) classifier



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

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

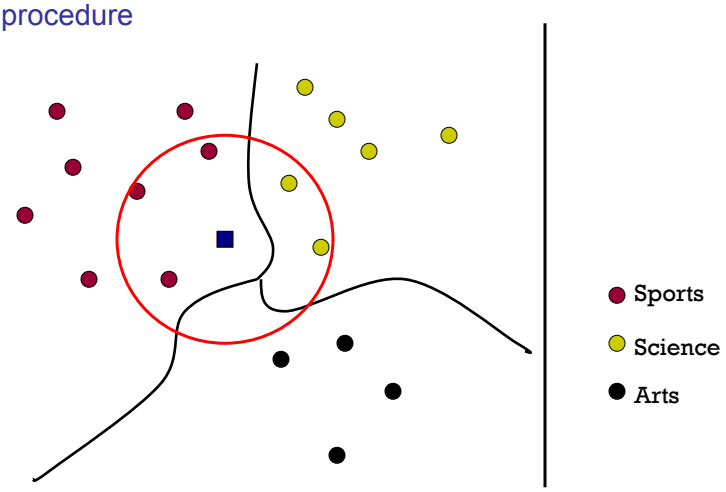
  - kNN density estimate

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

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

Voting kNN

- The procedure



Asymptotic Analysis

- Condition risk: $r_k(X,X_{NN})$
  - Test sample X
  - NN sample X_{NN}
  - Denote the event X is class I as $X \leftrightarrow I$
  - Assuming k=1

$$\begin{aligned} r_1(X, X_{NN}) &= Pr\Big\{\{X \leftrightarrow 1 \ \& \ X_{NN} \leftrightarrow 2\} \text{ or } \{X \leftrightarrow 2 \ \& \ X_{NN} \leftrightarrow 1\} | X, X_{NN}\Big\} \\ &= Pr\Big\{\{X \leftrightarrow 1 \ \& \ X_{NN} \leftrightarrow 2\}\Big\} + Pr\Big\{\{X \leftrightarrow 2 \ \& \ X_{NN} \leftrightarrow 1\} | X, X_{NN}\Big\} \\ &= q_1(X)q_2(X_{NN}) + q_2(X)q_1(X_{NN}) \end{aligned}$$

  - When an infinite number of samples is available, X_{NN} will be so close to X

$$r_1^*(X) = 2q_1(X)q_2(X) = 2\xi(X)$$

Asymptotic Analysis, cont.

- Recall conditional Bayes risk:

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

$$= \frac{1}{2} - \frac{1}{2}\sqrt{1 - 4\xi(X)}$$

$$= \sum_{i=1}^{\infty} \frac{1}{i}\binom{2i-2}{i-1}\xi^i(X)$$

This is called the MacLaurin series expansion

- Thus the asymptotic condition risk

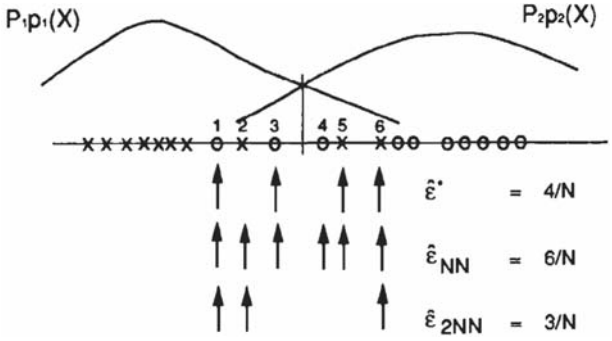
$$r_1^*(X) = 2\xi(X) \leq 2r^*(X)$$

- It can be shown that $\epsilon_1^* \leq 2\epsilon^*$
  - This is remarkable, considering that the procedure does not use any information about the underlying distributions and only the class of the single nearest neighbor determines the outcome of the decision.

In fact

$$\frac{1}{2}\epsilon^* \leq \epsilon_{2NN}^* \leq \epsilon_{4NN}^* \leq \ldots \leq \epsilon^* \leq \ldots \leq \epsilon_{3NN}^* \leq \epsilon_{NN}^* \leq 2\epsilon^*$$

- Example:



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

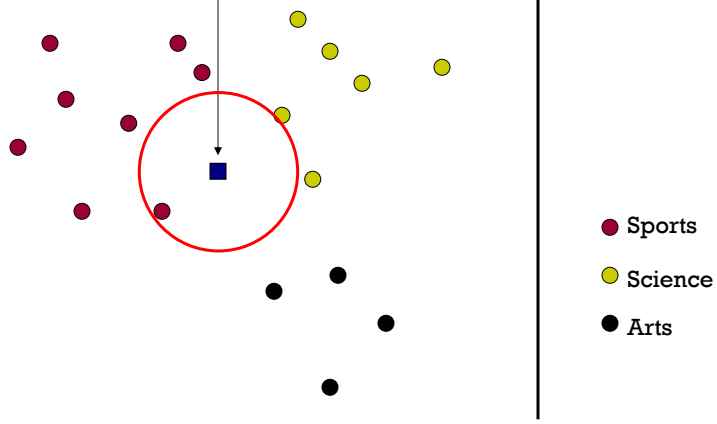
$$D(x, x') = \sqrt{\sum_i \sigma_i^2 (x_i - x_i')^2}$$

- Or equivalently,

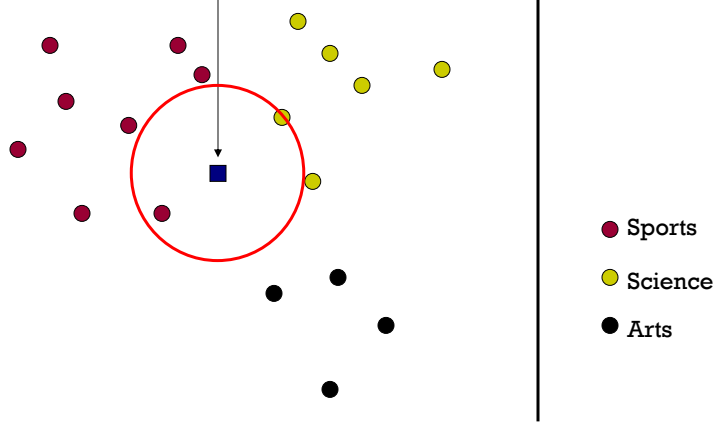
$$D(x, x') = \sqrt{(x - x')^T \Sigma (x - x')}$$

- Other metrics:
  - L_1 norm: |x-x'|
  - L_∞ norm: max |x-x'| (elementwise …)
  - Mahalanobis: where Σ is full, and symmetric
  - Correlation
  - Angle
  - Hamming distance, Manhattan distance
  - …

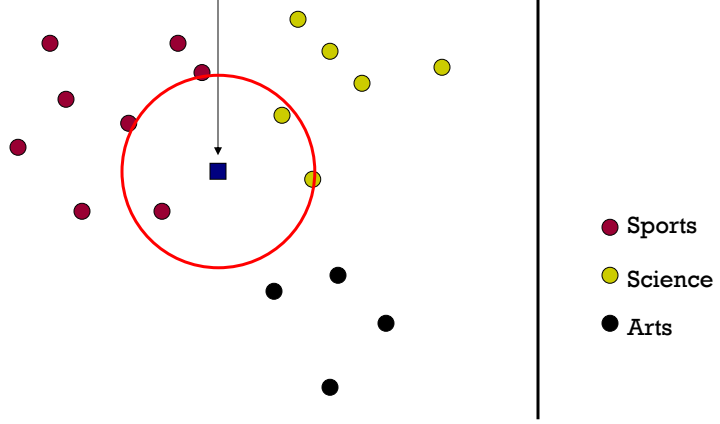
1-Nearest Neighbor (kNN) classifier



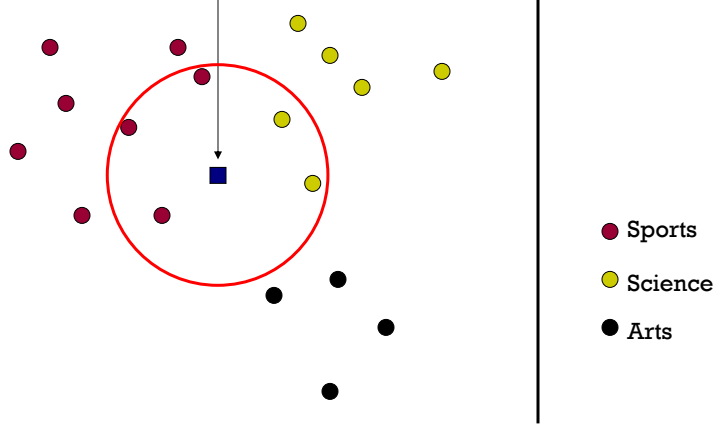
2-Nearest Neighbor (kNN) classifier



3-Nearest Neighbor (kNN) classifier



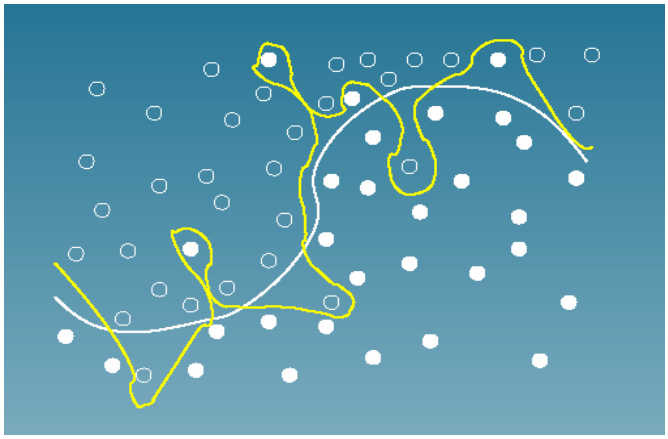
5-Nearest Neighbor (kNN) classifier



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

Is kNN ideal? … more later



Summary

- ***Bayes classifier*** is the best classifierwhich 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.