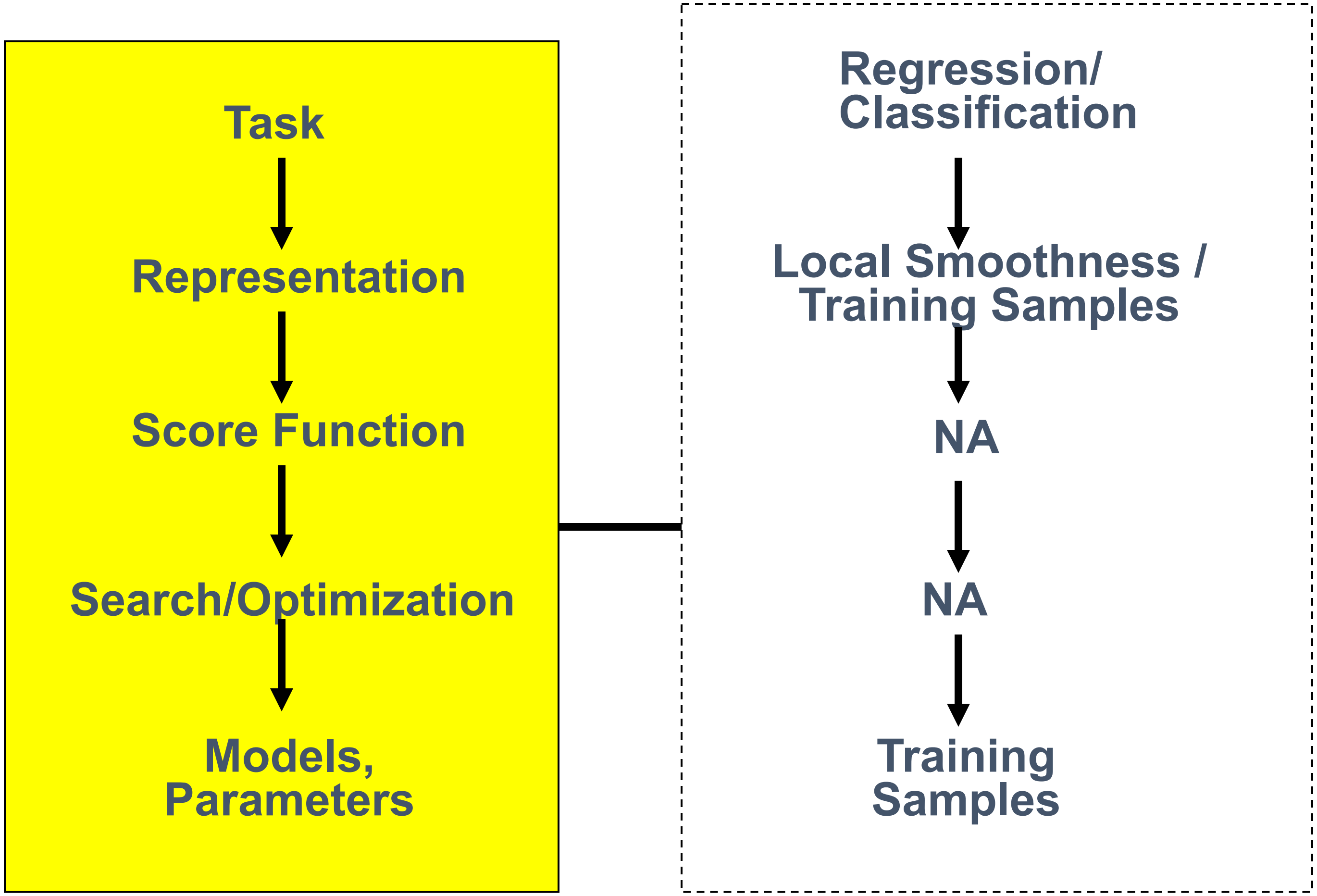
UVA CS 6316: Machine Learning

Lecture 9E: More about K-nearest-neighbor

Dr. Yanjun Qi

University of Virginia
Department of Computer Science

K-Nearest Neighbor



Task → Representation → Score Function → Search/Optimization → Models, Parameters

Regression/ Classification → Local Smoothness / Training Samples → NA → NA → Training Samples

Computational Scalable?

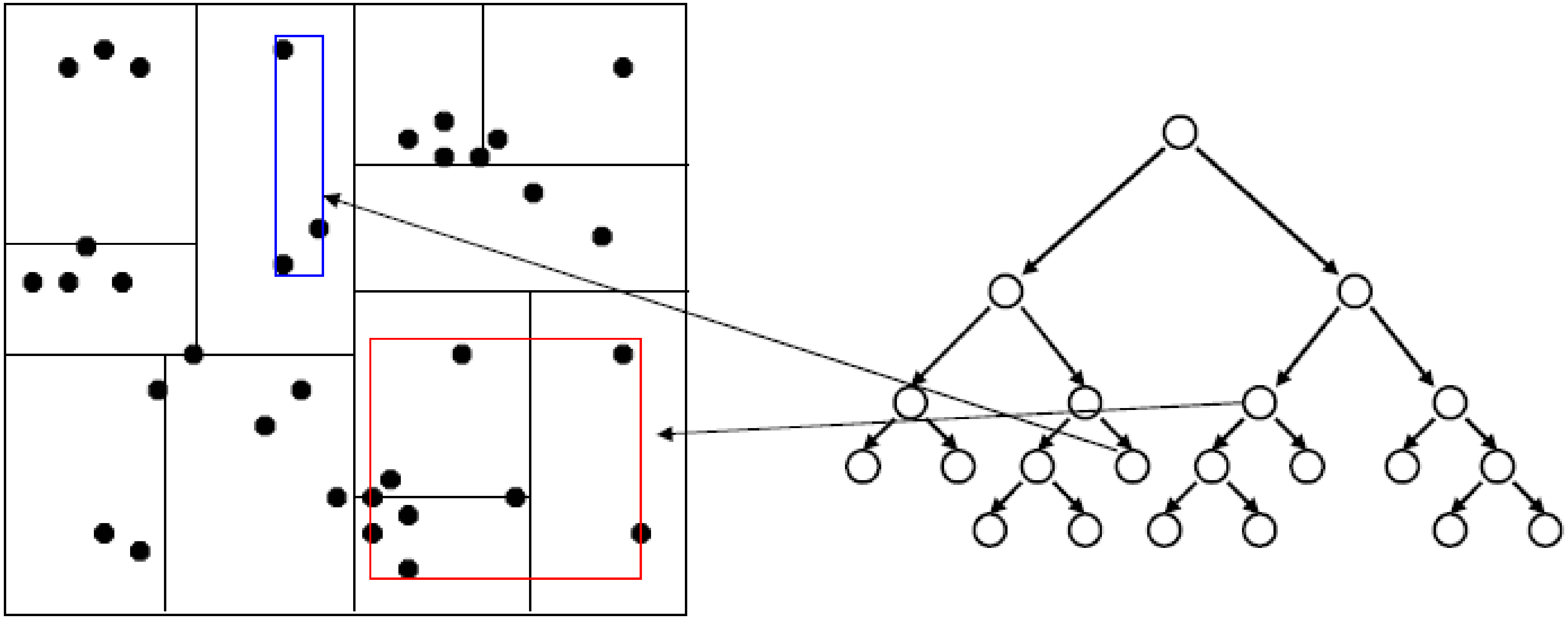
Computational Time Cost

	Train (n)	Test (m=1)
Linear Regression	$O(np^2+p^3)$	$O(p)$
KNN	$O(1)$	$O(np)+$ $O(\text{sort } n\text{-}k)$???

$p = 30{,}000$

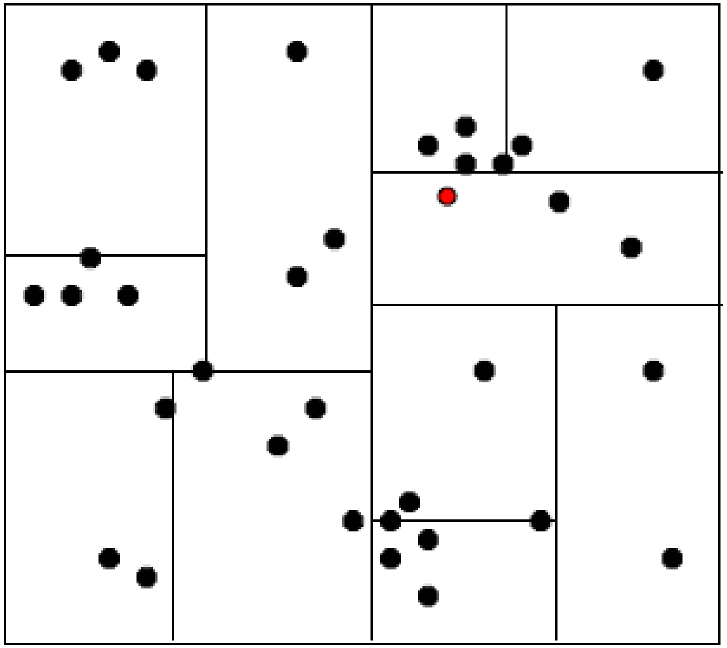
$n = 20{,}000$

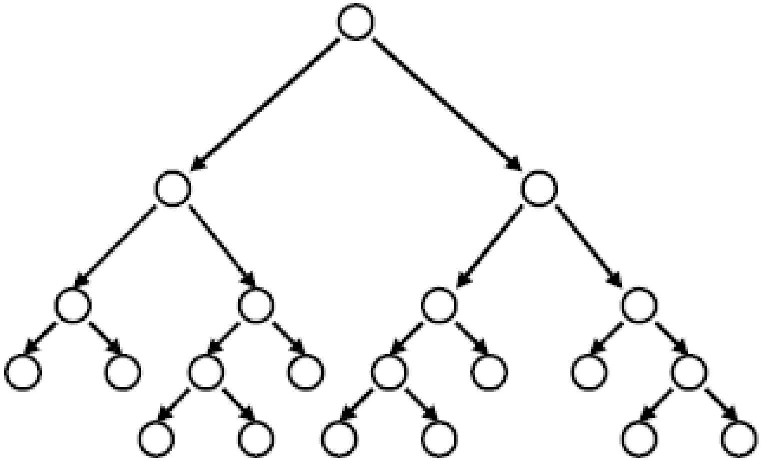
KD Tree for NN Search



- Each node contains
  - Children information
  - The tightest box that bounds all the data points within the node.

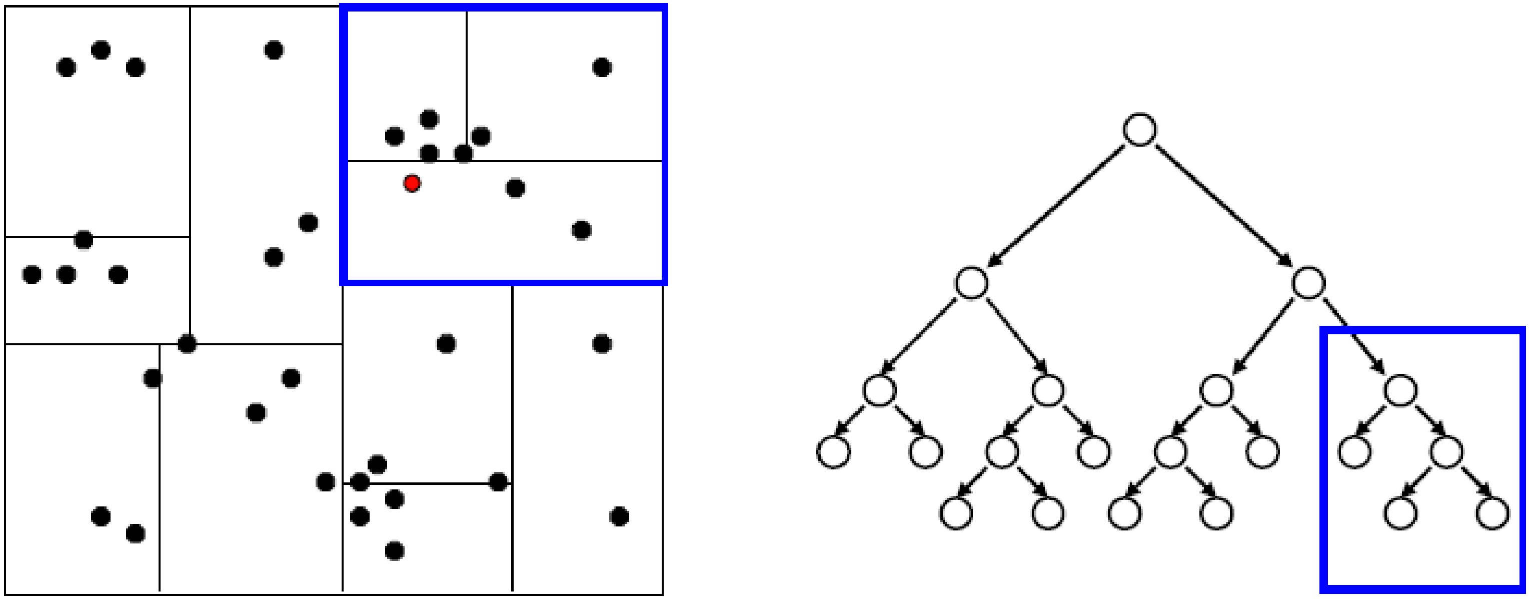
NN Search by KD Tree





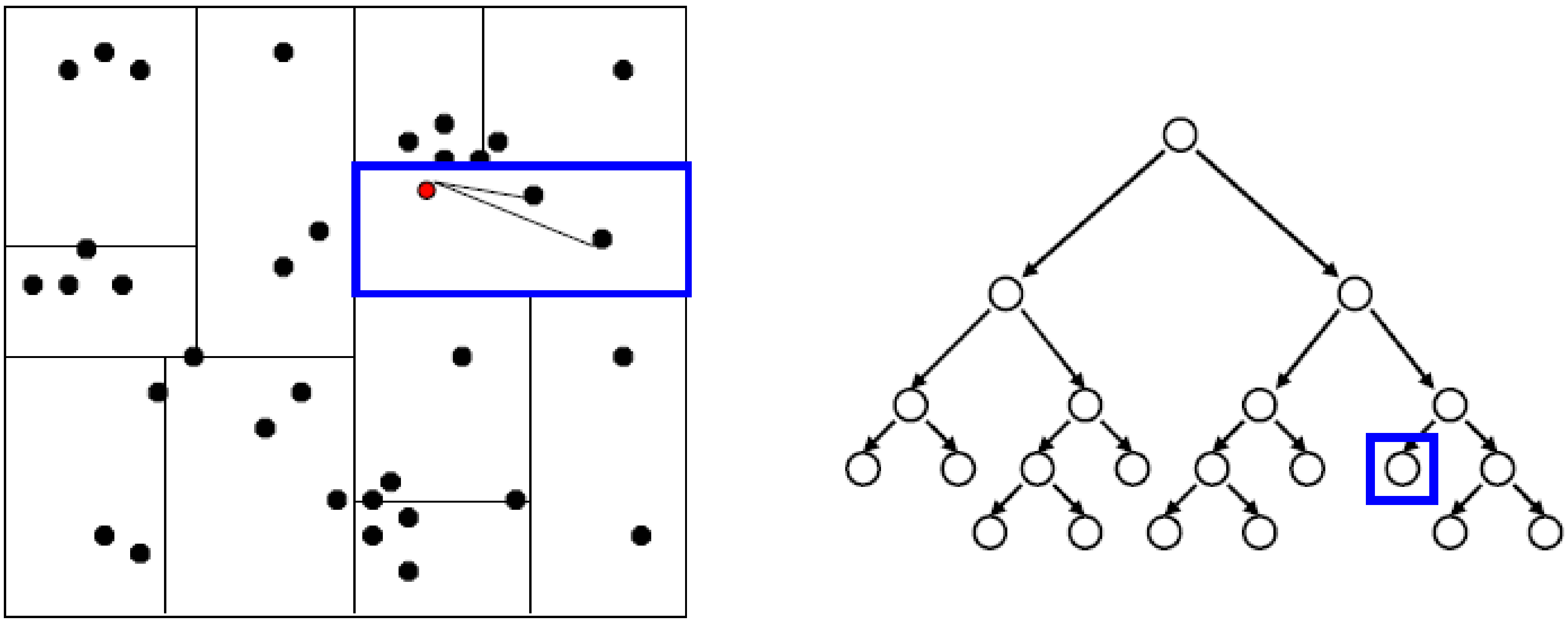
We traverse the tree looking for the nearest neighbor of the query point.

NN Search by KD Tree



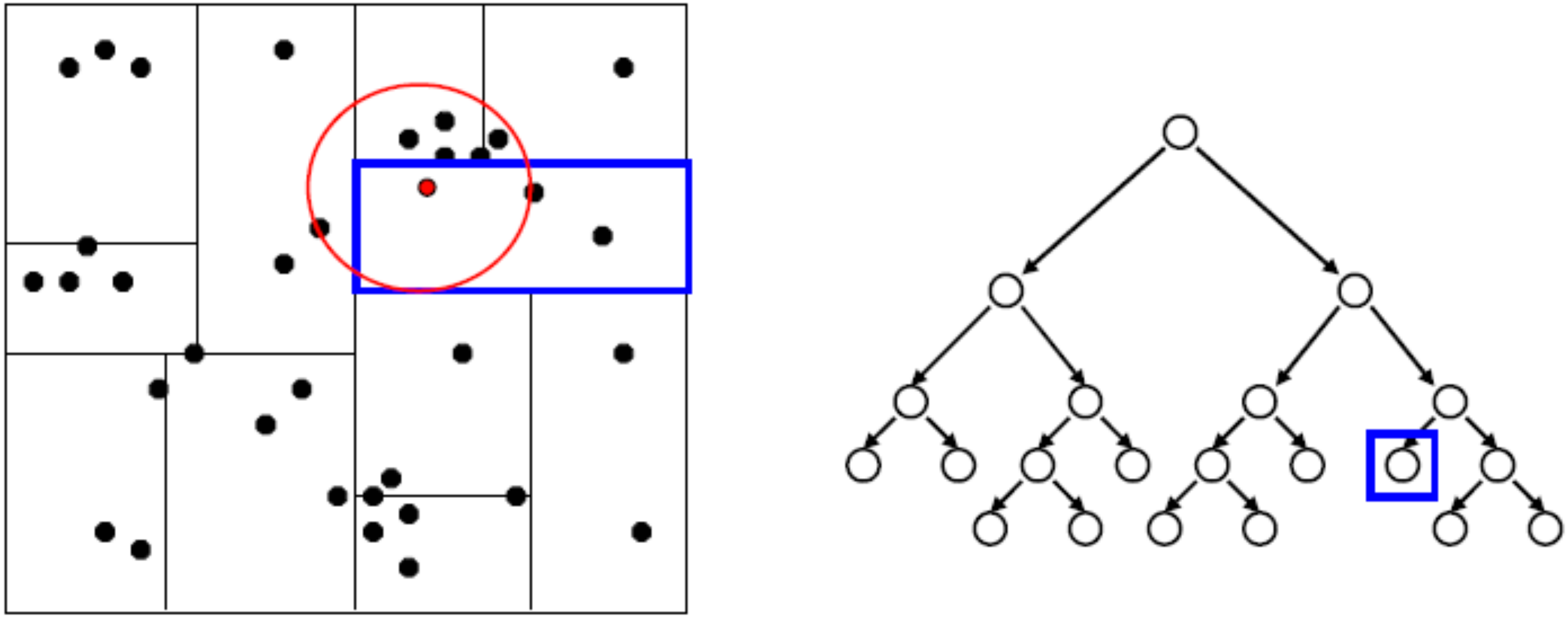
Examine nearby points first: Explore the branch of the tree that is closest to the query point first.

NN Search by KD Tree



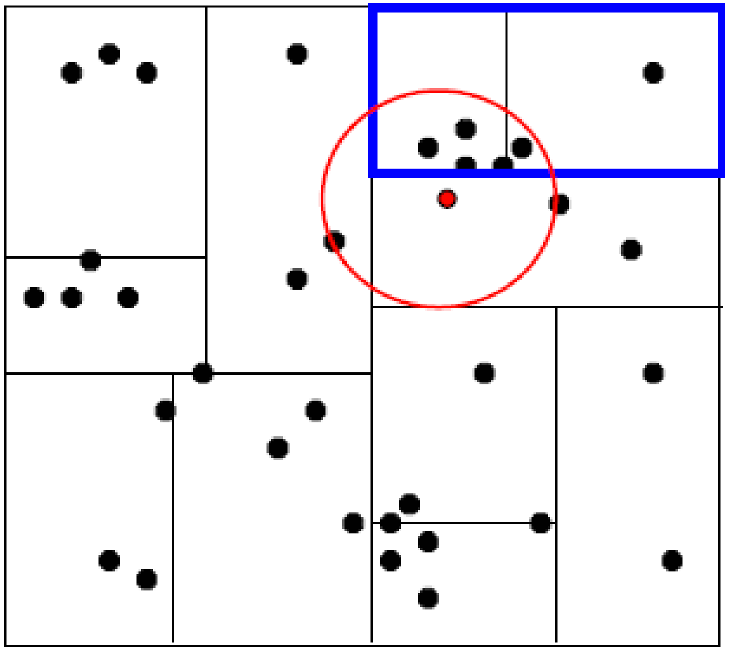
When we reach a leaf node: compute the distance to each point in the node.

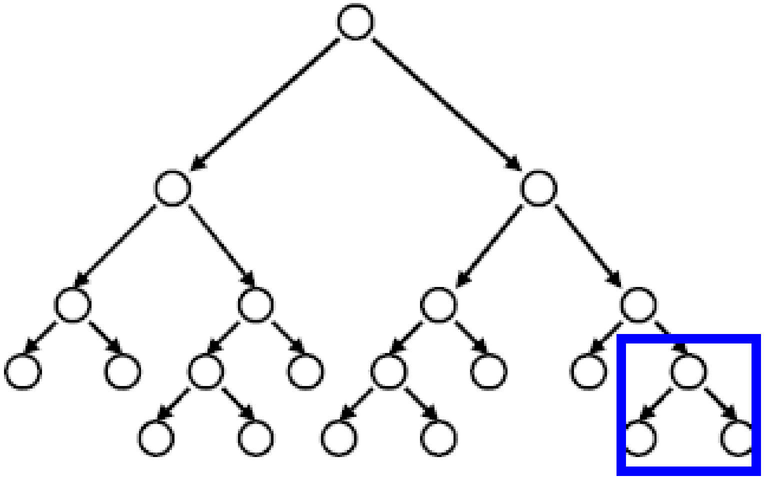
NN Search by KD Tree



When we reach a leaf node: compute the distance to each point in the node.

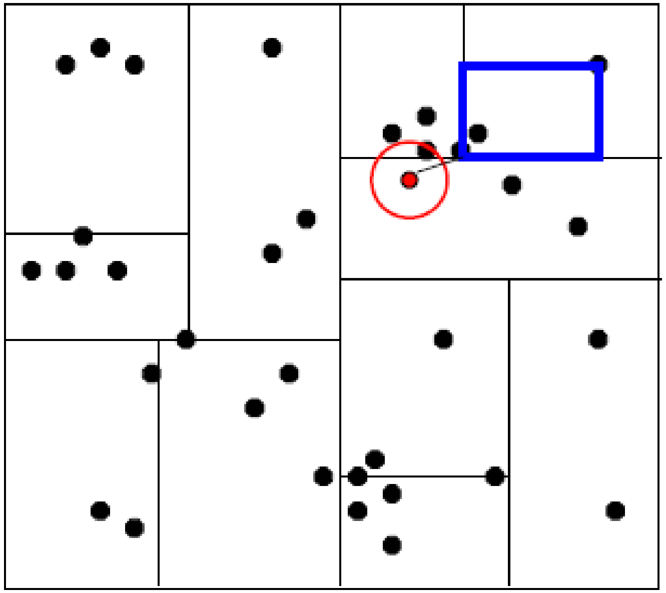
NN Search by KD Tree

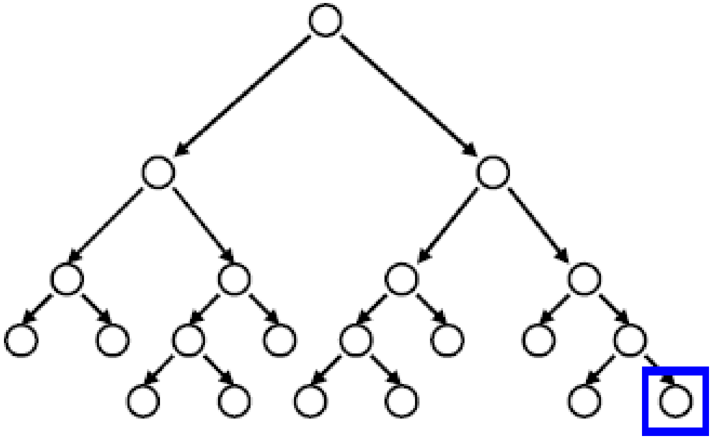




Then we can backtrack and try the other branch at each node visited.

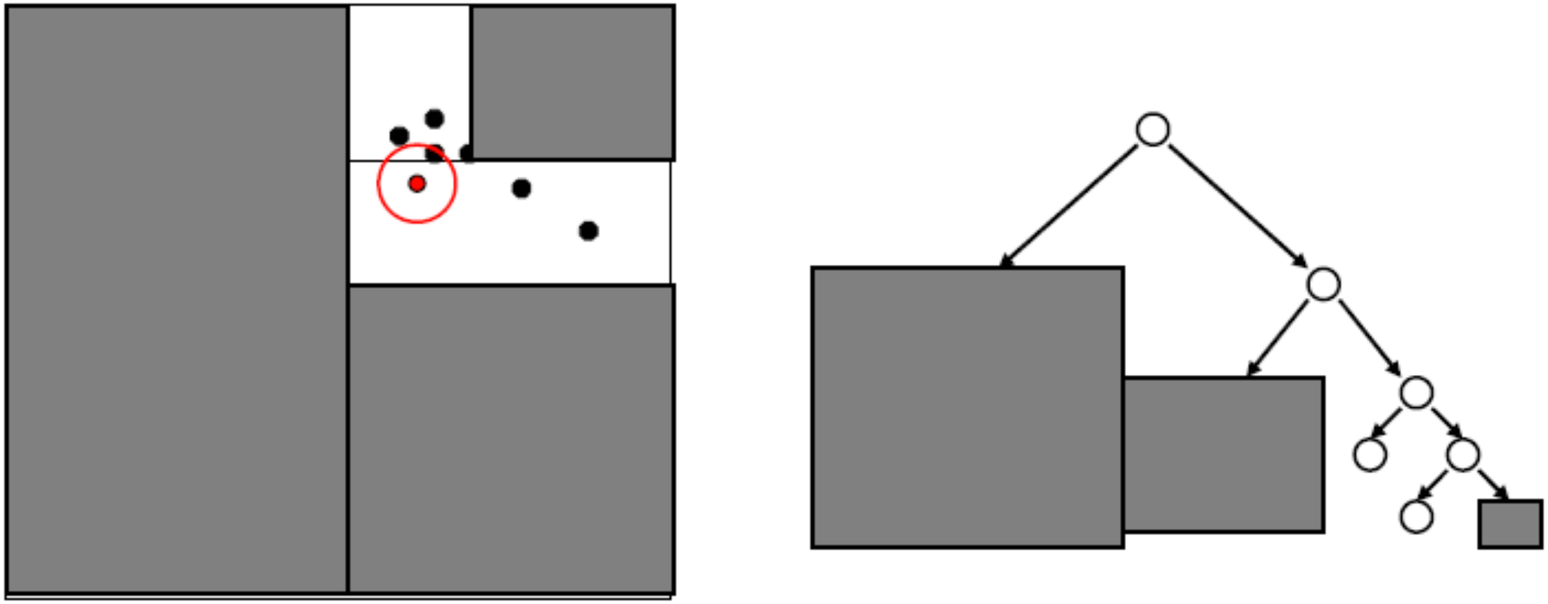
NN Search by KD Tree





Using the distance bounds and the bounds of the data below each node, we can prune parts of the tree that could NOT include the nearest neighbor.

NN Search by KD Tree



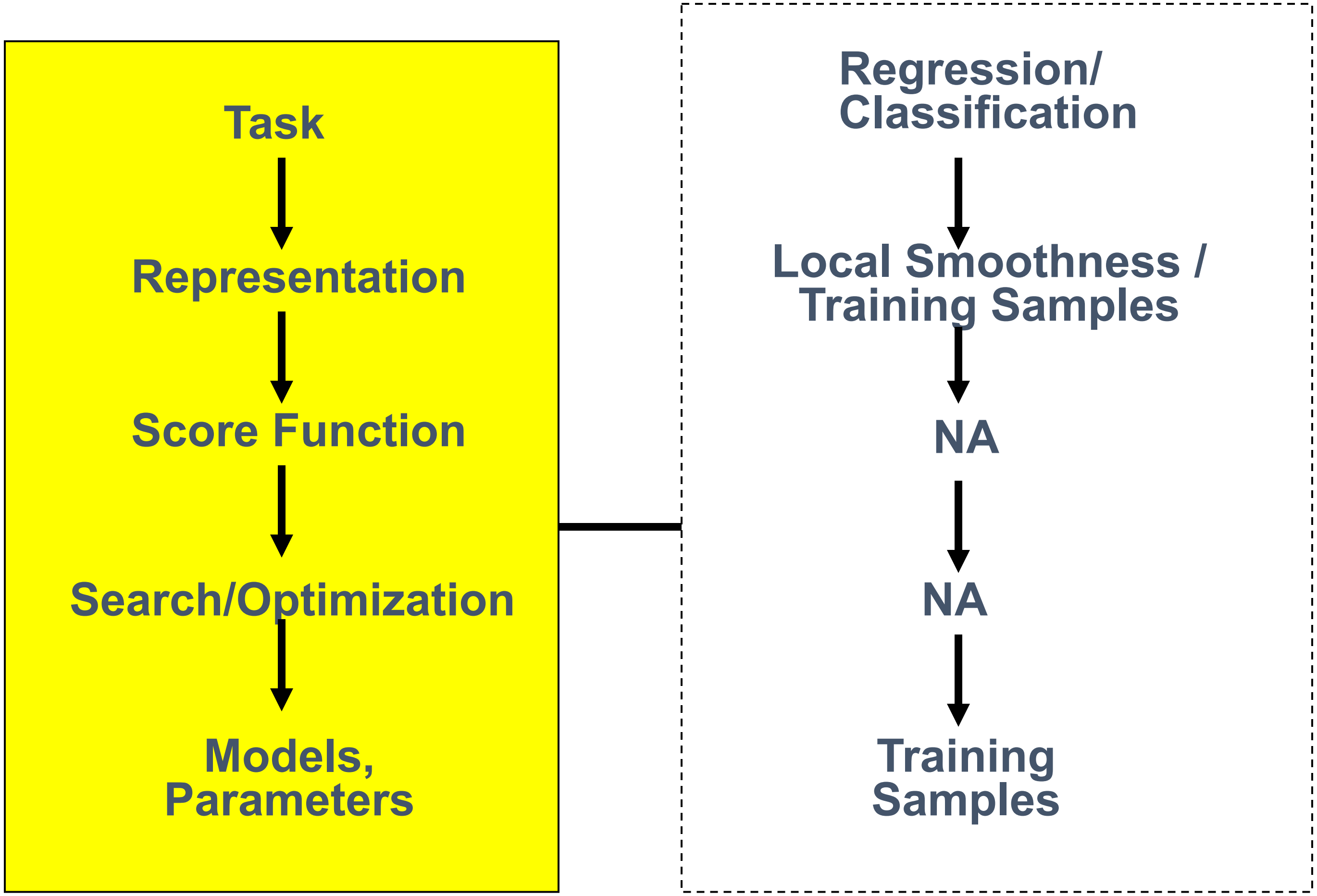
Using the distance bounds and the bounds of the data below each node, we can prune parts of the tree that could NOT include the nearest neighbor.

Curse of Dimensionality

- Imagine instances described by 20 attributes, but only 2 are relevant to target function
- Curse of dimensionality: nearest neighbor is easily mislead when high dimensional X
- Consider N data points uniformly distributed in a p-dimensional unit ball centered at origin. Consider the nn estimate at the original. The mean distance from the origin to the closest data point is:

$$d(p, N) = \left(1 - 2^{-1/N}\right)^{1/p} \approx 1 - \frac{\log N}{p}$$

K-Nearest Neighbor



Task

Representation

Score Function

Search/Optimization

Models, Parameters

Regression/ Classification

Local Smoothness / Training Samples

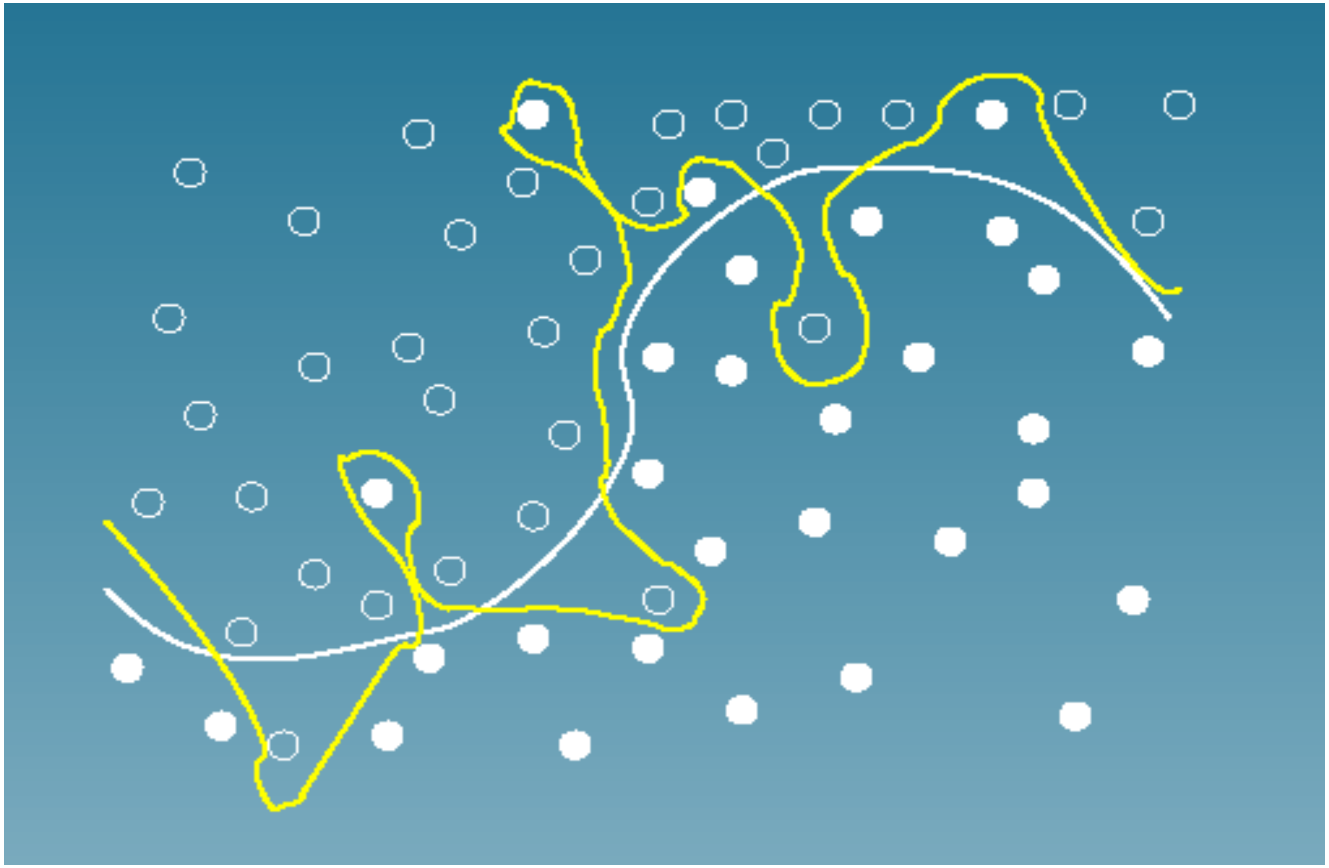
NA

NA

Training Samples

Asymptotically Sound?

Is kNN ideal? ... See Extra



© Eric Xing @ CMU, 2006-2008

Bayes Classifier

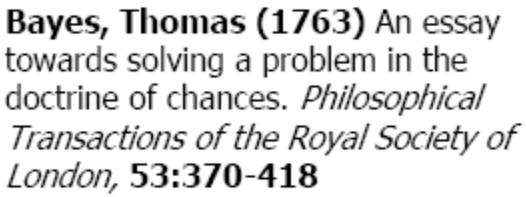
- ***Bayes classifier*** is the best classifier which minimizes the probability of classification error.
- A classifier becomes the ***Bayes classifier*** if the density estimates converge to the true densities
  - when an infinite number of samples are used
  - The resulting error is the ***Bayes error,*** the smallest achievable error given the underlying distributions.

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

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

- Bayes Test:

- Likelihood Ratio:

$$\ell(X)=$$

- Discriminant function:

$$h(X)=$$

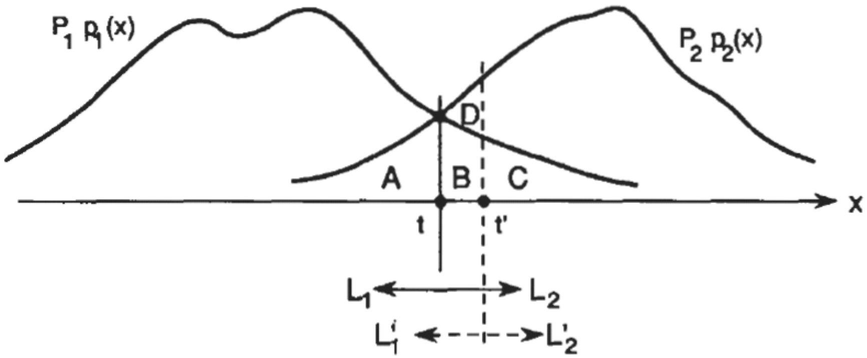
Bayes Error

- 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*?

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

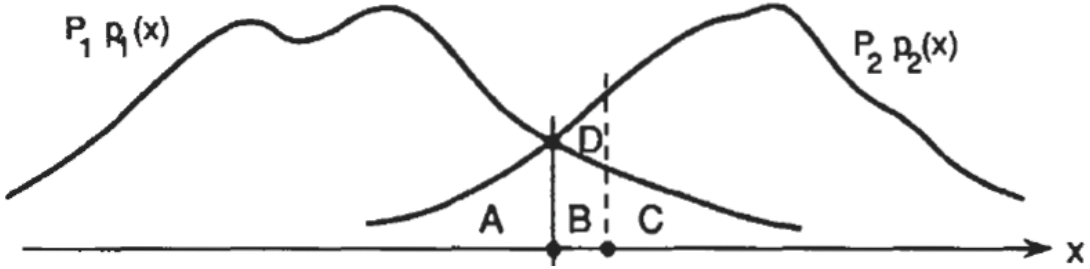
- The Bayes error (the expected risk):

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



More on Bayes Error

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

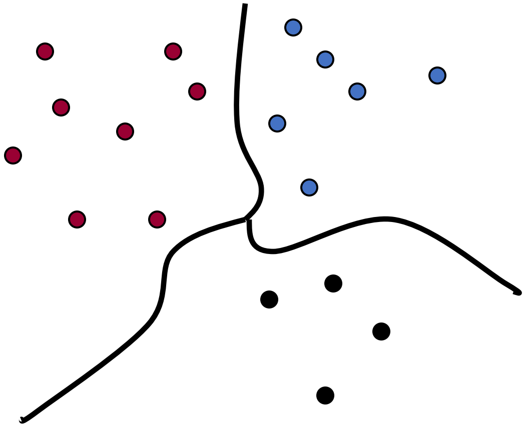


- Bayes classifier is the theoretically best classifier that minimizes probability of classification error
- Computing Bayes error is in general a very complex problem. Why?
  - Density estimation:
  - Integrating density function:

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

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:



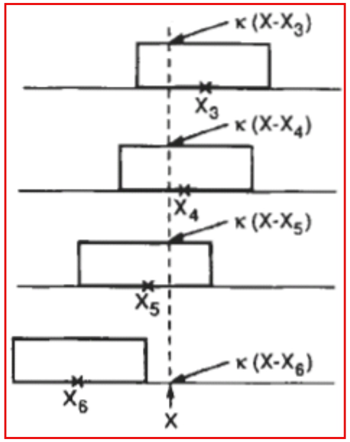
Where does kNN come from?

- How to estimation $p(X)$?

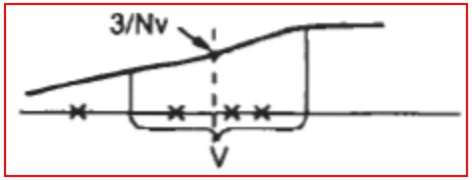
- Nonparametric density estimation
  - Parzen density estimate

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:

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

Voting kNN classifier

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

Asymptotic Analysis

- Condition risk: $\mathrm{r}_{\mathrm{k}}(X,X_{NN})$
  - Test sample X
  - NN sample X_{NN}
  - Denote the event X is class I as $X \leftrightarrow I$

  - Assuming k=1

$$r_1(X, X_{NN}) = Pr\Big\{\{X \leftrightarrow 1 \ \& \ X_{NN} \leftrightarrow 2\} \text{ or } \{X \leftrightarrow 2 \ \& \ X_{NN} \leftrightarrow 1\} | X, X_{NN}\Big\}$$

$$= Pr\Big\{\{X \leftrightarrow 1 \ \& \ X_{NN} \leftrightarrow 2\}\Big\} + Pr\{\{X \leftrightarrow 2 \ \& \ X_{NN} \leftrightarrow 1\} | X, X_{NN}\Big\}$$

$$= q_1(X)q_2(X_{NN}) + q_2(X)q_1(X_{NN})$$

  - 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

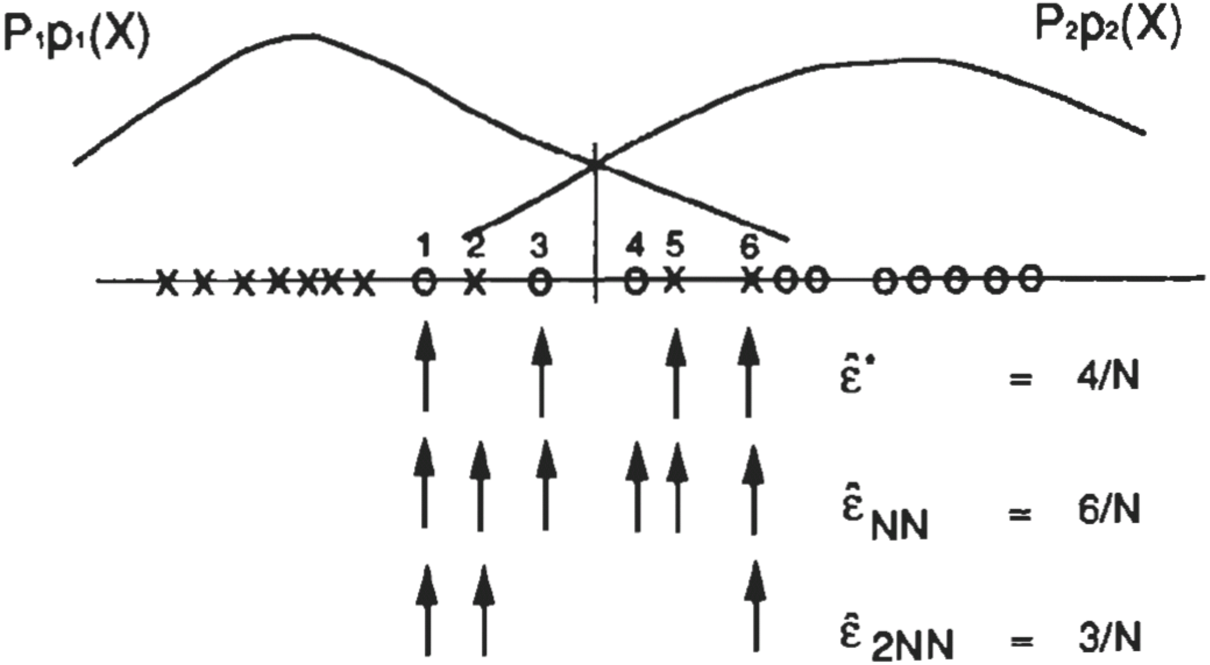
$$r_1^*(X) = 2\xi(X) \leq 2r^*(X)$$

- 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:


$P_1p_1(X)$
$P_2p_2(X)$
1
2
3
4
5
6
$\hat{\varepsilon}^*$ = 4/N
$\hat{\varepsilon}_{NN}$ = 6/N
$\hat{\varepsilon}_{2NN}$ = 3/N

References

- ❑ Prof. Tan, Steinbach, Kumar's "Introduction to Data Mining" slide
- ❑ Prof. Andrew Moore's slides
- ❑ Prof. Jin Rong's slides about kNN
- ❑ Prof. Eric Xing's slides
- ❑ Hastie, Trevor, et al. *The elements of statistical learning*. Vol. 2. No. 1. New York: Springer, 2009.