Nearest Neighbor Methods Understanding

By Shekh Firoz Alam

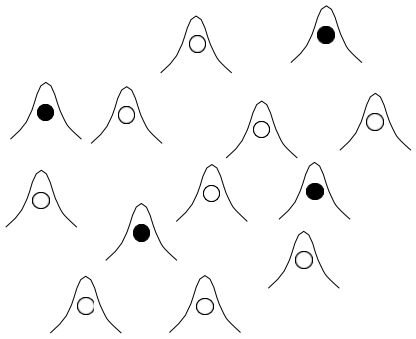
Generalization bounds using VC dimension

$$\text{error}_{true}(h) \leq \text{error}_{train}(h) + \sqrt{\frac{VC(H)\left(\ln \frac{2m}{VC(H)} + 1\right) + \ln \frac{4}{\delta}}{m}}$$

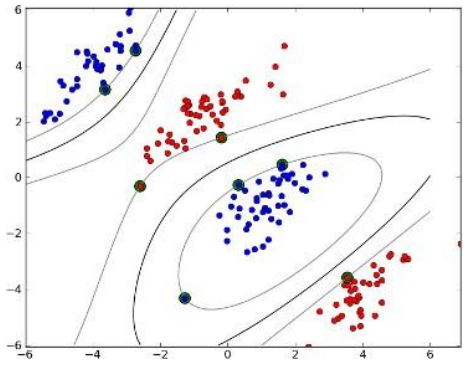
- Linear classifiers:
  - VC(H) = d+1, for *d* features plus constant term *b*
- SVM with Gaussian Kernel
  - VC(H) = ∞

$$K(\vec{u}, \vec{v}) = \exp\left(-\frac{||\vec{u} - \vec{v}||_2^2}{2\sigma^2}\right)$$

Euclidean distance, squared



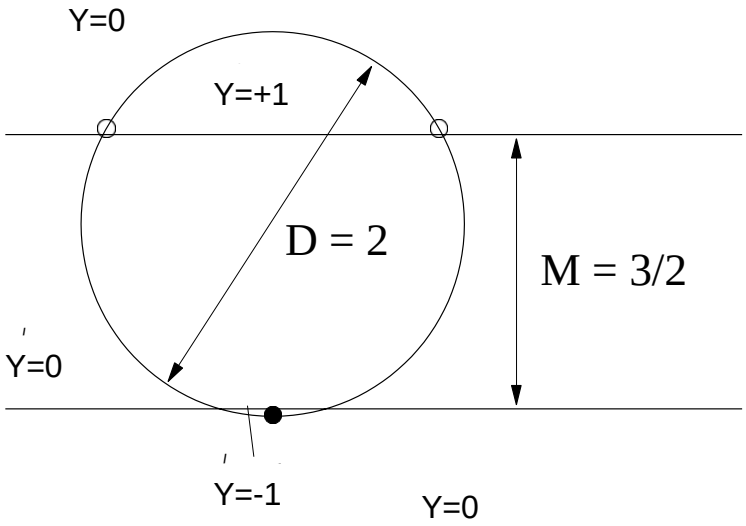
[Figure from Chris Burges]



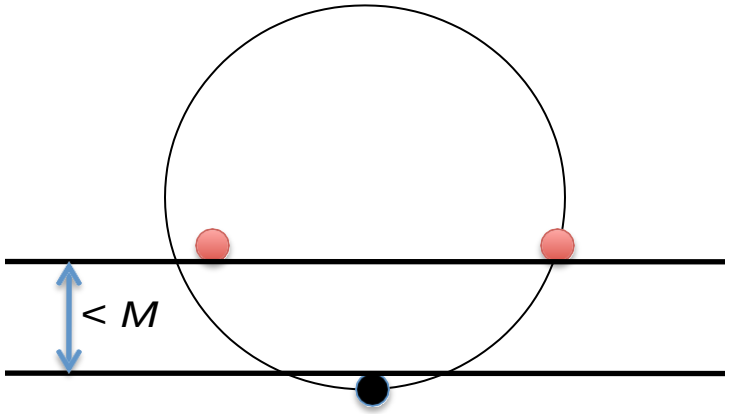
[Figure from mblondel.org]

Gap tolerant classifiers

- Suppose data lies in R^d in a ball of diameter **D**
- Consider a hypothesis class H of linear classifiers that can only classify point sets with margin at least **M**
- What is the largest set of points that H can shatter?



Cannot shatter these points:



$$\text{VC dimension} = \min\left(d, \frac{D^2}{M^2}\right)$$

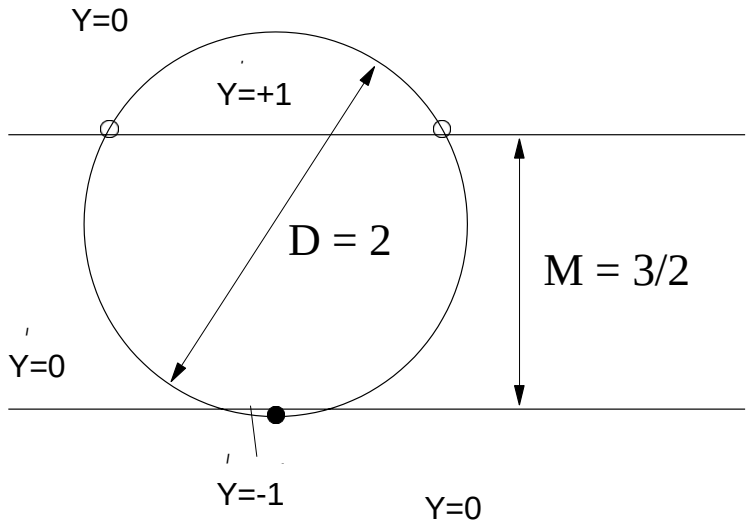
$$M = 2\gamma = 2\frac{1}{||w||}$$

SVM attempts to *minimize* $||w||^2$, which *minimizes* VC---dimension!!!

[Figure from Chris Burges]

Gap tolerant classifiers

- Suppose data lies in R^d in a ball of diameter **D**
- Consider a hypothesis class H of linear classifiers that can only classify point sets with margin at least **M**
- What is the largest set of points that H can shatter?



VC dimension = $\min\left(d, \frac{D^2}{M^2}\right)$

$$K(\vec{u},\vec{v}) = \exp\left(-\frac{||\vec{u}-\vec{v}||_2^2}{2\sigma^2}\right)$$

What is R=D/2 for the Gaussian kernel?

$$R = \max_x ||\varphi(x)|| = \max_x \sqrt{\varphi(x)\cdot\varphi(x)} = \max_x \sqrt{K(x,x)} = 1 \quad !!!$$

What is $||w||^2$?

$$||w||^2 = \frac{2}{M}^2$$

$$||w||^2 = ||\sum_i \alpha_i y_i \varphi(x_i)||^2 = \sum_i \sum_j \alpha_i \alpha_j y_i y_j K(x_i, x_j)$$

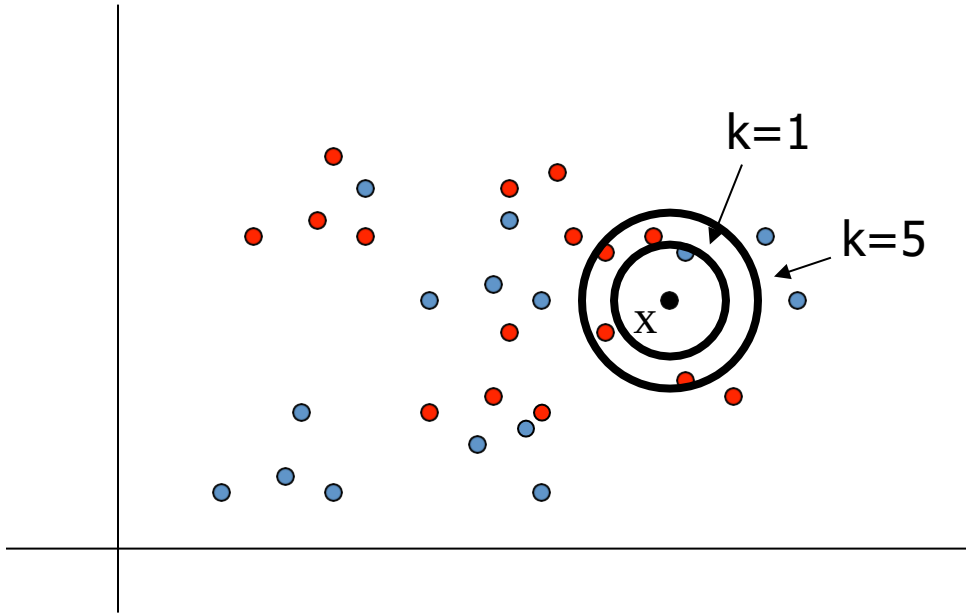
[Figure from Chris Burges]

Nearest Neighbor Algorithm

- Learning Algorithm:
  - Store training examples
- Prediction Algorithm:
  - To classify a new example **x** by finding the training example $(\mathbf{x}^i, y^i)$ that is *nearest* to **x**
  - Guess the class $y = y^i$

K-Nearest Neighbor Methods

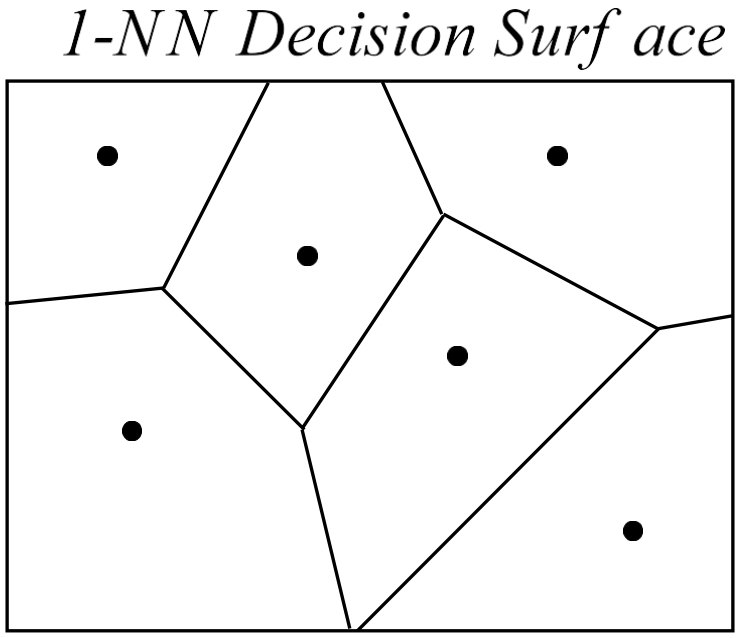
- To classify a new input vector x, examine the k-closest training data points to x and assign the object to the most frequently occurring class



common values for k: 3, 5

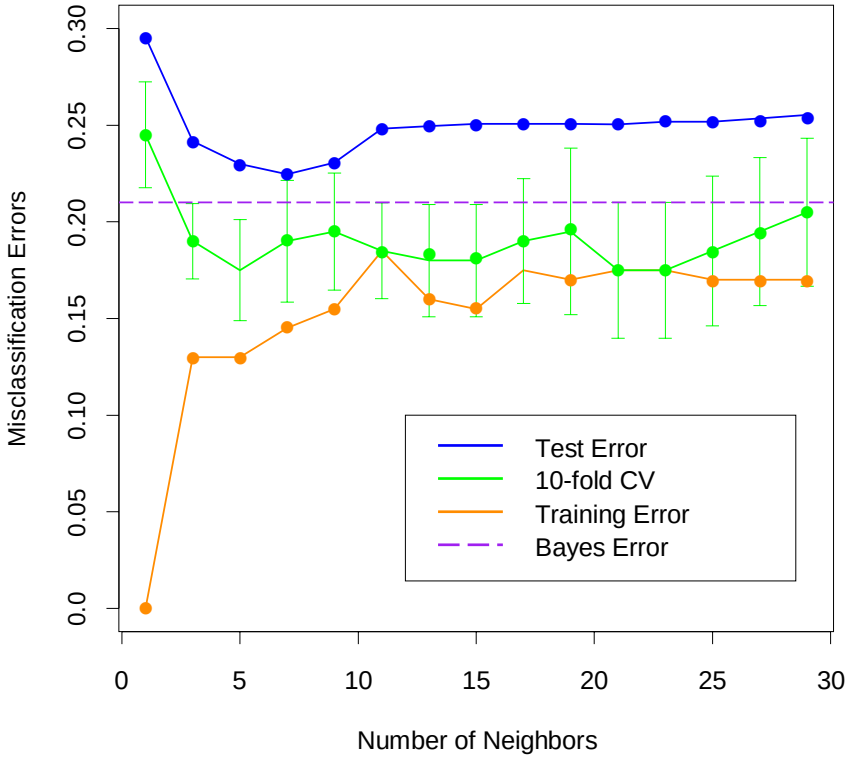
Decision Boundaries

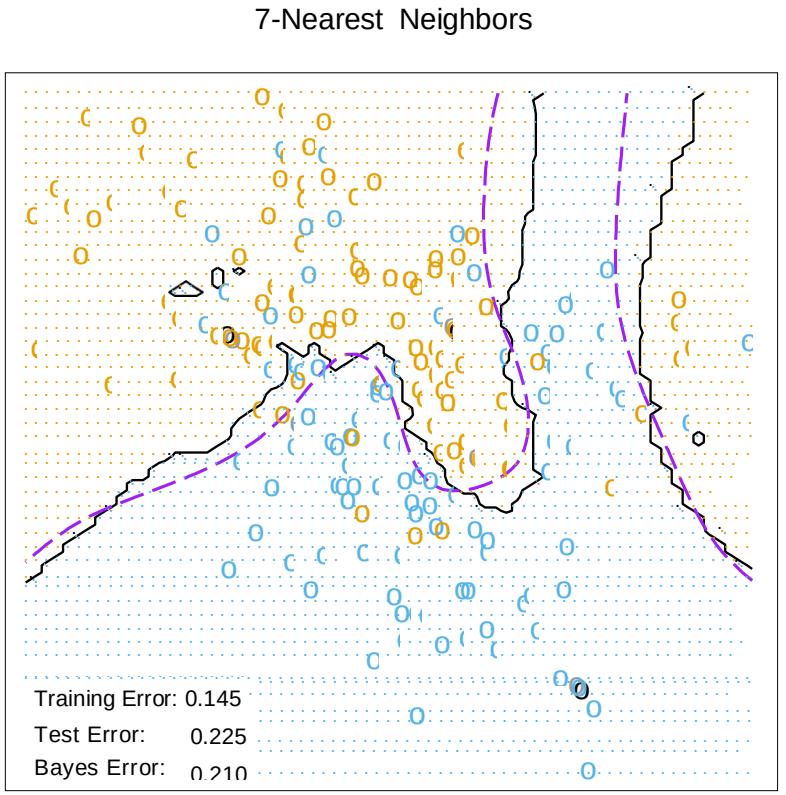
- The nearest neighbor algorithm does not explicitly compute decision boundaries. However, the decision boundaries form a subset of the Voronoi diagram for the training data.



- The more examples that are stored, the more complex the decision boundaries can become

Example results for k---NN





[Figures from Hastie and Tibshirani, Chapter 13]

Nearest Neighbor

When to Consider

- Instance map to points in R^n
- Less than 20 attributes per instance
- Lots of training data

Advantages

- Training is very fast
- Learn complex target functions
- Do not lose information

Disadvantages

- Slow at query time
- Easily fooled by irrelevant attributes

Issues

- Distance measure
  - Most common: Euclidean
- Choosing k
  - Increasing k reduces variance, increases bias
- For high-dimensional space, problem that the nearest neighbor may not be very close at all!
- Memory-based technique. Must make a pass through the data for each classification. This can be prohibitive for large data sets.

Distance

- Notation: object with p measurements

$$x^i = (x^i_1, x^i_2, \ldots, x^i_p)$$

- Most common distance metric is *Euclidean* distance:

$$d_E(x^i, x^j) = \left(\sum_{k=1}^{p} (x^i_k - x^j_k)^2 \right)^{\frac{1}{2}}$$

- ED makes sense when different measurements are commensurate; each is variable measured in the same units.

- If the measurements are different, say length and weight, it is not clear.

Standardization

When variables are not commensurate, we can standardize them by dividing by the sample standard deviation. This makes them all equally important.

The estimate for the standard deviation of x_k :

$$\hat{\sigma}_k = \left(\frac{1}{n} \sum_{i=1}^{n} \left(x_k^i - \bar{x}_k \right)^2 \right)^{\frac{1}{2}}$$

where x_k is the sample mean:

$$\bar{x}_k = \frac{1}{n} \sum_{i=1}^{n} x_k^i$$

Weighted Euclidean distance

Finally, if we have some idea of the relative importance of each variable, we can weight them:

$$d_{WE}(i,j) = \left(\sum_{k=1}^{p} w_k (x_k^i - x_k^j)^2 \right)^{\frac{1}{2}}$$

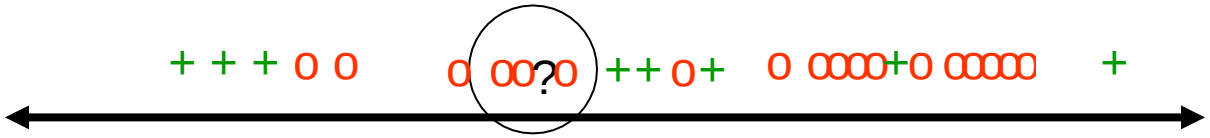
The Curse of Dimensionality

- Nearest neighbor breaks down in high-dimensional spaces because the "neighborhood" becomes very large.

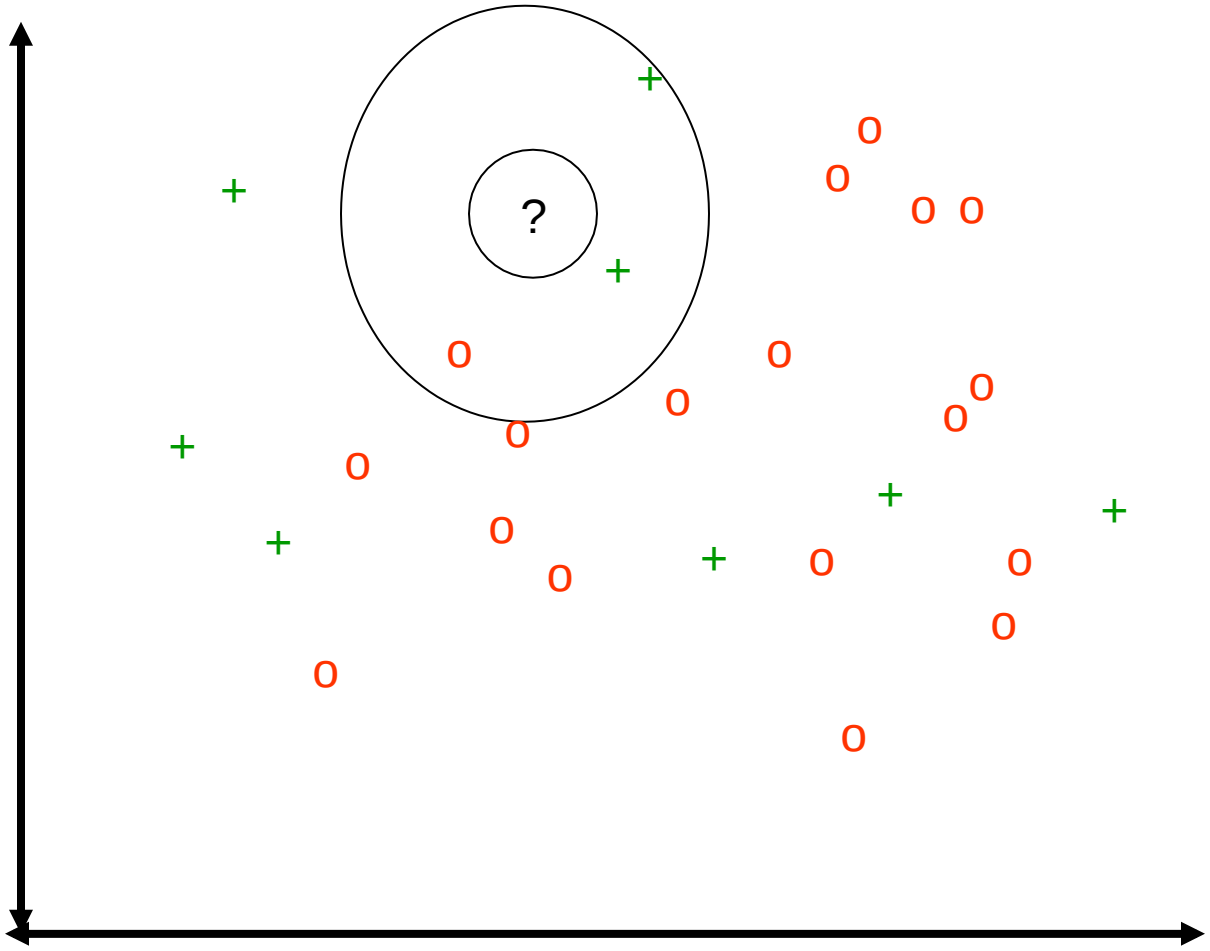
- Suppose we have 5000 points uniformly distributed in the unit hypercube and we want to apply the 5-nearest neighbor algorithm.

- Suppose our query point is at the origin.
  - 1D –
    - On a one dimensional line, we must go a distance of 5/5000 = 0.001 on average to capture the 5 nearest neighbors
  - 2D –
    - In two dimensions, we must go sqrt(0.001) to get a square that contains 0.001 of the volume
  - D –
    - In D dimensions, we must go $(0.001)^{1/D}$

K-NN and irrelevant features



K-NN and irrelevant features



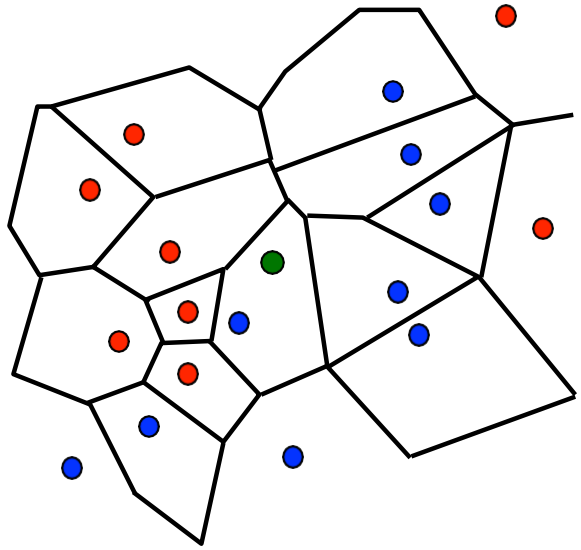
Nearest neighbor problem

- Problem: given sample $S = ((x_1, y_1), \ldots, (x_m, y_m))$, find the nearest neighbor of test point .
  - general problem extensively studied in computer science.
  - exact vs. approximate algorithms.
  - dimensionality N crucial.
  - better algorithms for small intrinsic dimension (e.g., limited doubling dimension).

[Slides from Mehyrar Mohri]

Efficient Indexing: N=2

- Algorithm:
  - compute Voronoi diagram in $O(m \log m)$.
  - point location data structure to determine NN.
  - complexity: $O(m)$ space, $O(\log m)$ time.



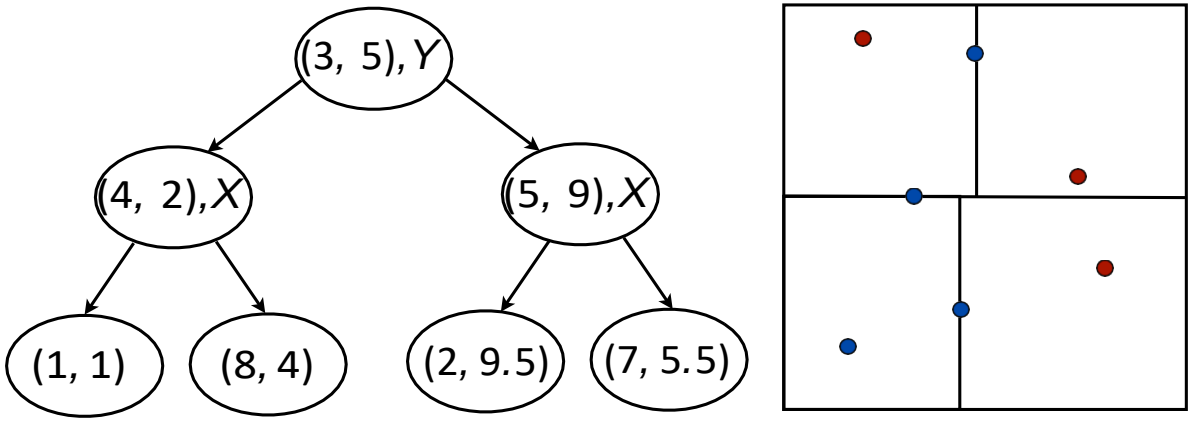
[Slides from Mehyrar Mohri]

Efficient Indexing: N>2

- Voronoi diagram: size in $O\ m^{N/2}$.
- Linear algorithm (no pre-processing):
  - compute distance $x - x_i$ for all $i \in [1, m]$.
  - complexity of distance computation: $\Omega(Nm)$.
  - no additional space needed.
- Tree-based data structures: pre-processing.
  - often used in applications:*k*-d trees (*k*-dimensional trees).

[Slides from Mehyrar Mohri]

Efficient Indexing for N>2: KD trees



Construction algorithm

- Algorithm: for each non-leaf node,
  - choose dimension (e.g., longest of hyperrectangle).
  - choose pivot (median).
  - split node according to (pivot, dimension).

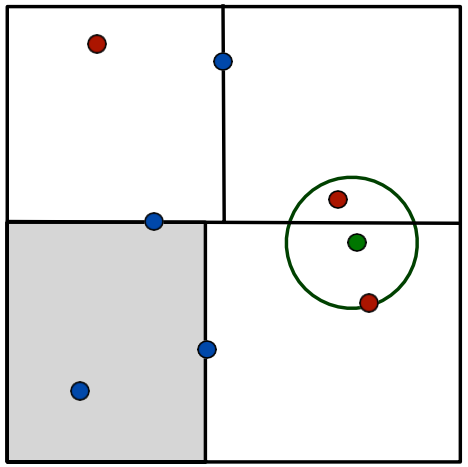
→ balanced tree, binary space partitioning.

[Slides from Mehyrar Mohri]

Efficient Indexing for N>2: KD trees

- Algorithm:
  - find region containing (starting from root node, move to child node based on node test).
  - save region point x_0 as current best.
  - move up tree and recursively search regions intersecting hypersphere $S(x, x - x_0)$:
    - update current best if current point is closer.
    - restart search with each intersecting sub-tree.
    - move up tree when no more intersecting sub-tree.

Search algorithm



[Slides from Mehyrar Mohri]

k-NN is similar to SVM with Gaussian kernel!

- Consider the following generalization of the *k*-NN algorithm (specialized to binary classification):

$$\hat{y}(x) \leftarrow \text{sign} \sum_{i=1}^{N} y_i d(x_i, x) \quad \text{with } d(x_i, x) = \frac{1}{||x_i - x||_2^2} \text{ or... } d(x_i, x) = \exp\left(-\frac{||x_i - x||_2^2}{2\sigma^2}\right)$$

- Looks at *all* training points (i.e., *k*=N), but weights the *i*'th training point's label by how far $\mathbf{x}_i$ is from $\mathbf{x}$

- Now compare this to classification with SVM and a Gaussian kernel:

$$\hat{y}(x) \leftarrow \text{sign} \sum_{i=1}^{N} \alpha_i y_i K(x_i, x) \qquad K(\vec{u}, \vec{v}) = \exp\left(-\frac{||\vec{u} - \vec{v}||_2^2}{2\sigma^2}\right) \qquad 0 \le \alpha_i \le C$$

- The discriminant functions are nearly identical! The SVM has parameters α_i that can be learned to maximize predictive accuracy

KNN Advantages

- Easy to program
- No optimization or training required
- Classification accuracy can be very good; can outperform more complex models