METU Informatics Institute

Min720

Pattern Classification with Bio-Medical Applications

Part 6: Nearest and k-nearest Neighbor Classification

Nearest Neighbor (NN) Rule &
k-Nearest Neighbor (k-NN) Rule

Non-parametric classification rules:

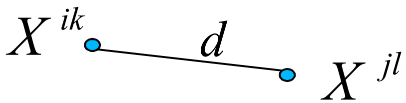
- Linear and generalized discriminant functions
- Nearest Neighbor & k-NN rules

NN Rule

1-NN: A direct classification using learning samples

Assume we have M learning samples from all categories

$$X^{11}, X^{12}, \ldots\ldots, \quad X^{jk}, \ldots\ldots\ldots \quad \ldots.$$



Xjk: k'th sample from i'th category

Assume a distance measure between samples

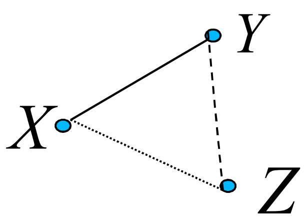
$$d(X^{ik}, X^{jl})$$

A general distance metric should obey the following rules:

$$d(X^{ij}, X^{ij}) = 0$$

$$d(X^{ij}, X^{jl}) = d(X^{jl}, X^{ij})$$

$$d(X,Y) \leq d(X,Z) + d(Z,Y)$$



Most standard: Euclidian Distance

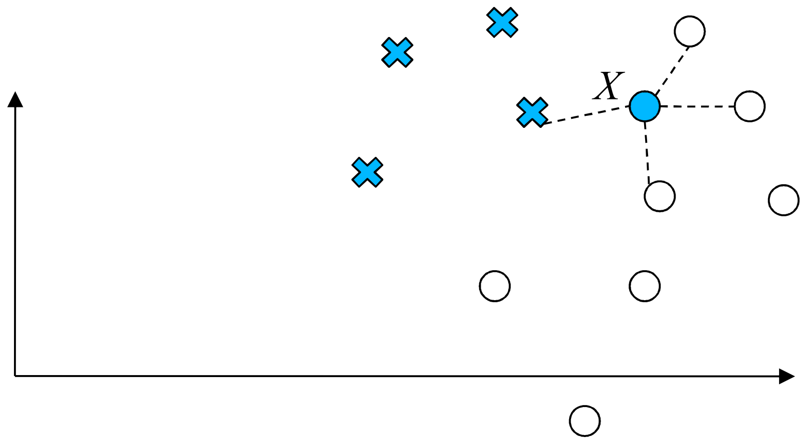
$$d(X,Y) = \left\| X - Y \right\| = \left[\sum_{i=1}^{n} (x_i - y_i)^2 \right]^{1/2} = \left[(X - Y)^T (X - Y) \right]^{1/2}$$

1-NN Rule: Given an unknown sample X

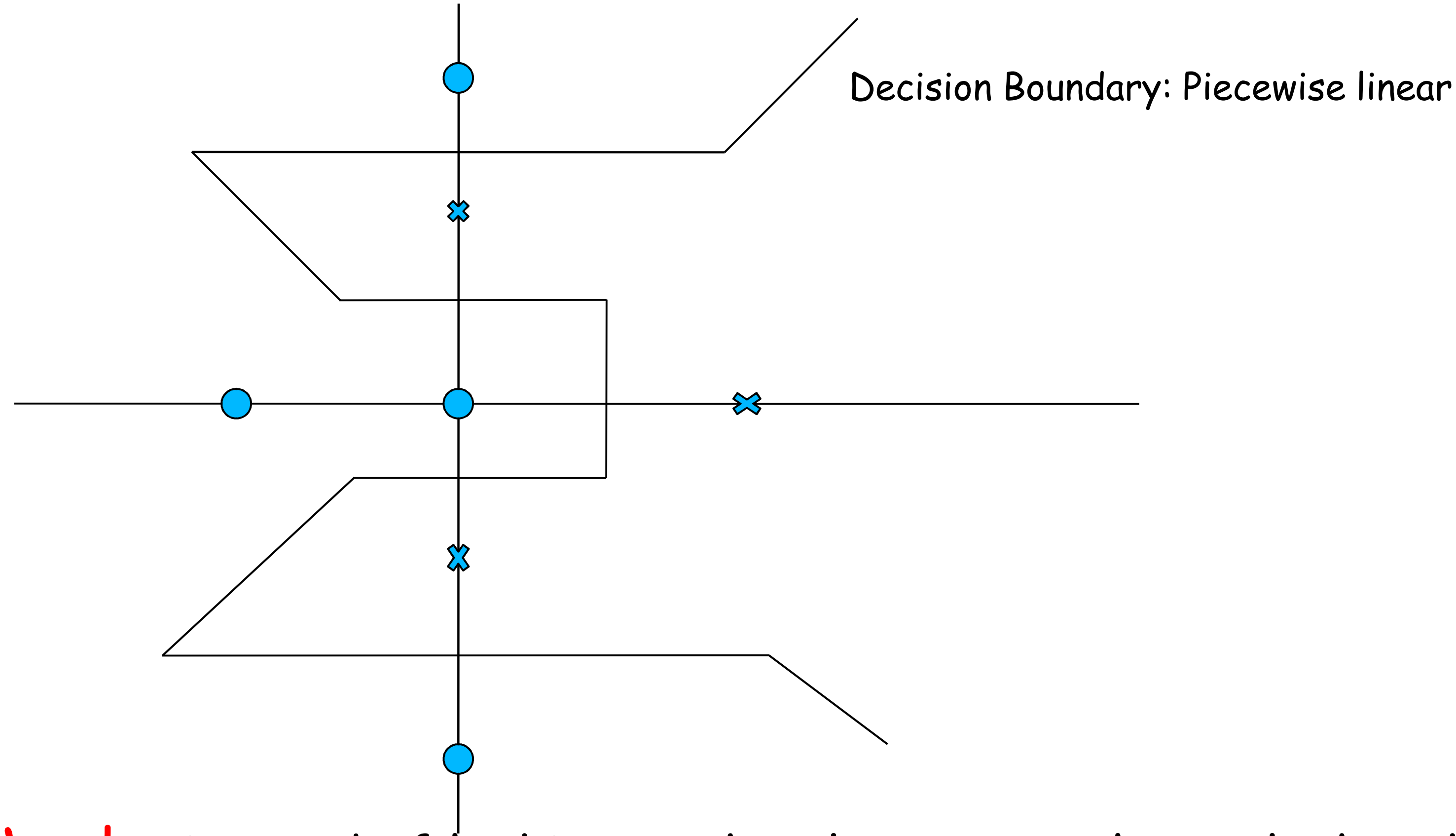
α_i if $d(X, X^{ik}) < d(X, X^{jl})$

For $jl \neq ik$

That is, assign X to category ω_i if the closest neighbor of X is from category i.



Example: Find the decision boundary for the problem below.



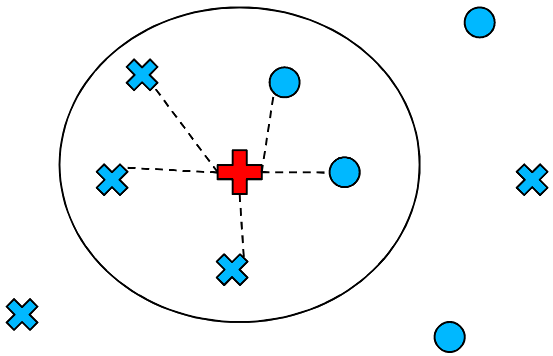
k-NN rule: instead of looking at the closest sample, we look at k nearest neighbors to X and we take a vote. The largest vote wins. k is usually taken as an odd number so that no ties occur.

• k-NN rule is shown to approach optimal classification when k becomes very large but $\frac{k}{M} \to 0$

• k-I NN (NN with a <u>reject option)</u>

Decide if majority is higher than a given threshold I. Otherwise reject.

k=5



If we threshold I=4 then,

For above example  is rejected to be classified.

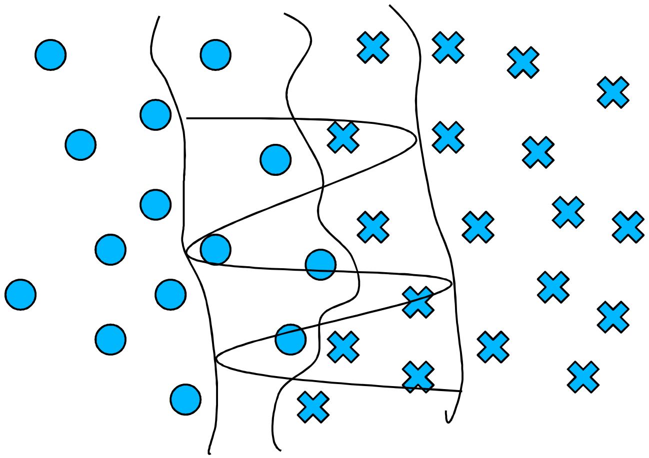
• Analysis of NN rule is possible when $M \to \infty$ and it was shown that it is no worse than twice of the minimum-error classification (in error rate).

EDITING AND CONDENSING

• NN rule becomes very attractive because of its simplicity and yet good performance.

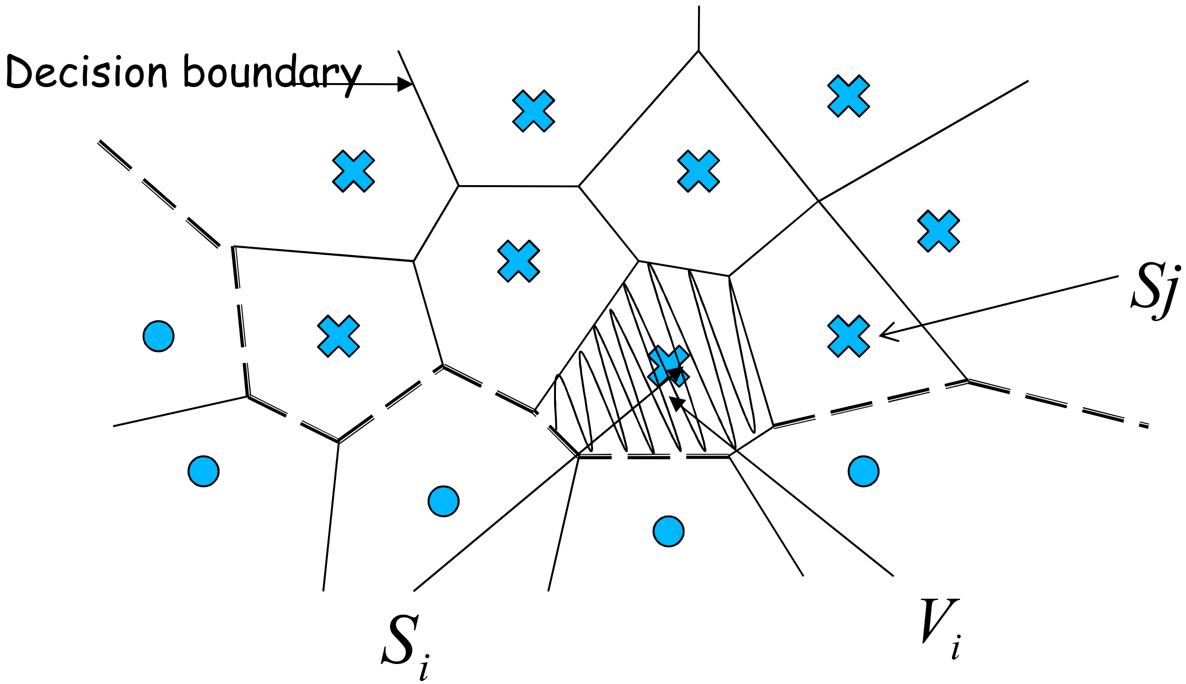
• So, it becomes important to reduce the computational costs involved.

• Do an intelligent elimination of the samples.



•Remove samples that do not contribute to the decision boundary.

Voronoi Diagrams



- V_i is a polygon such that any point that falls in V_i is closer to S_i than any other sample Sj .
-

• So the editing rule consists of throwing away all samples that do not have a Voronoi polygon that has a common boundary belonging to a sample from other category.

NN Editing Algorithm

- Consider the Voronoi diagrams for all samples
- Find Voronoi neighbors of sample X'
- If any neighbor is from other category, keep X'. Else remove from the set.
- Construct the Voronoi diagram with the remaining samples and use it for classification.

Advantage of NNR:

- No learning – no estimation
- so easy to implement

Disadvantage:

- Classification is more expensive. So people found ways to reduce the cost of NNR.

Analysis of NN and k-NN rules:

Possible when

- When $n \to \infty$, X(unknown sample) and X'(nearest neighbor) will get very close. Then,
- $P(\omega_i | X) \to P(\omega_i | X')$ that is, we are selecting the category ω_i with probability $P(\omega_i | X)$ (a-posteriori probability).

Error Bounds and Relation with Bayes Rule:

- Assume
  E^* - Error bound for Bayes Rule(a number between 0 and 1)
  E_1 - Error bound for 1-NN
  E_k - Error bound for k-NN

- It can be shown that

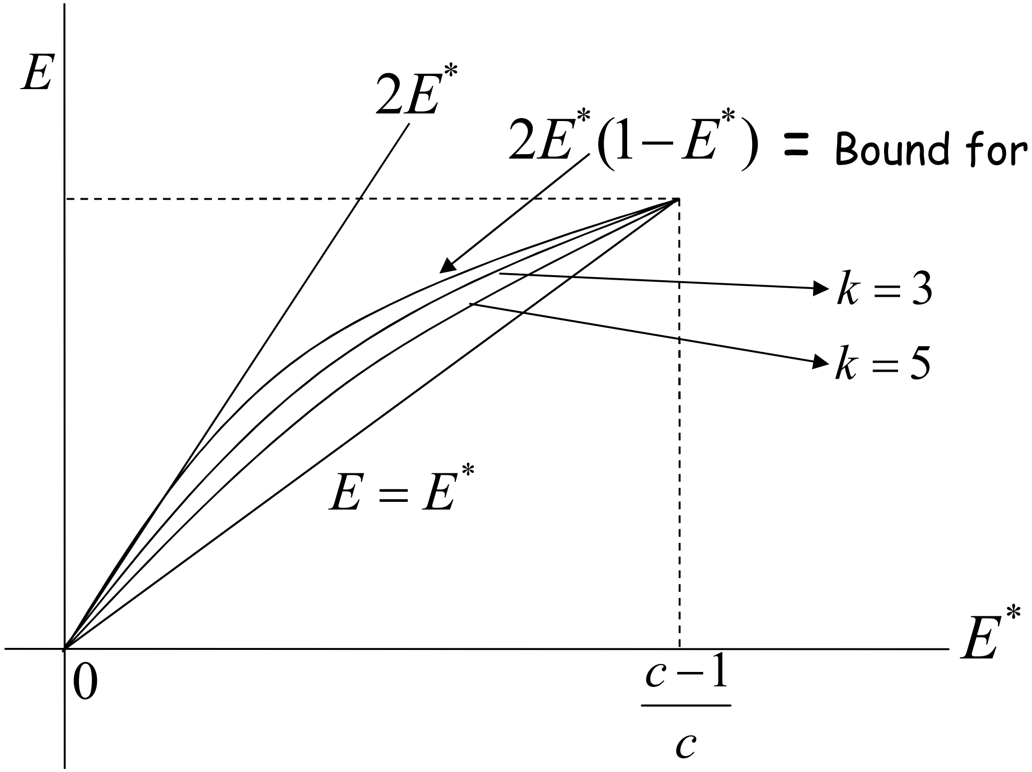
$$E^* \leq E_1 \leq 2E^*(1 - E^*) \leq 2E^*$$

for 2 categories and

$$E^* \leq E_1 \leq 2E^*\left(1 - \frac{c}{2(c-1)}E^*\right) \leq 2E^*$$

for c categories

- Always better than twice the Bayes error rate!



•Highest error occurs when $P_1(x) = P_2(x) = \ldots\ldots.$
(all densities are the same) then,

$$P(\omega_i) = \frac{1}{c} \qquad P(error) = 1 - \frac{1}{c} = \frac{c-1}{c}$$

E

Distance Measures(Metrices)

Non-negativity $D(x, y) \geq 0$

Reflexivity $D(x, y) = 0$ when only x=y

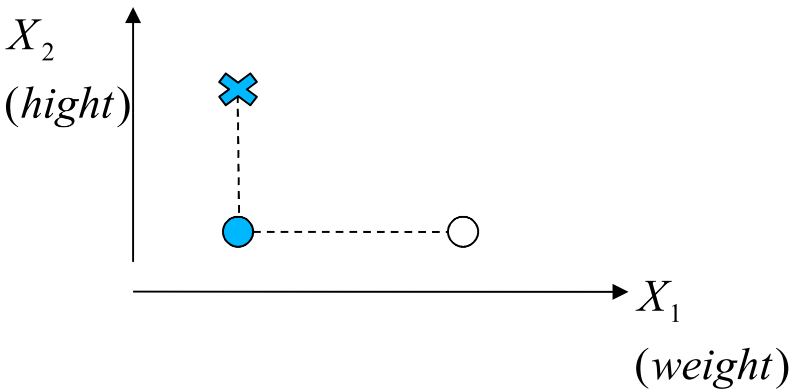
Symmetry $D(x, y) = D(y, x)$

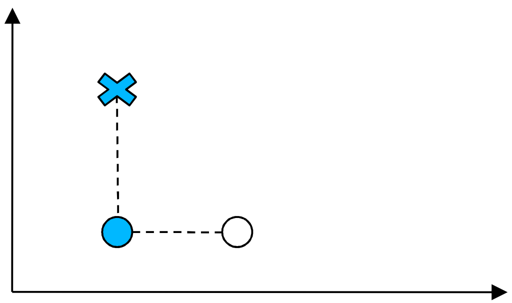
Triangle inequality $D(x, z) + D(z, y) \geq D(x, y)$

Euclidian distance satisfies these, but not always a meaningful measure.

Consider 2 features with different units scaling problem.

Solution: Scaling: Normalize the data by re-scaling (Reduce the range to 0-1





Scale of X_1 changed to half.

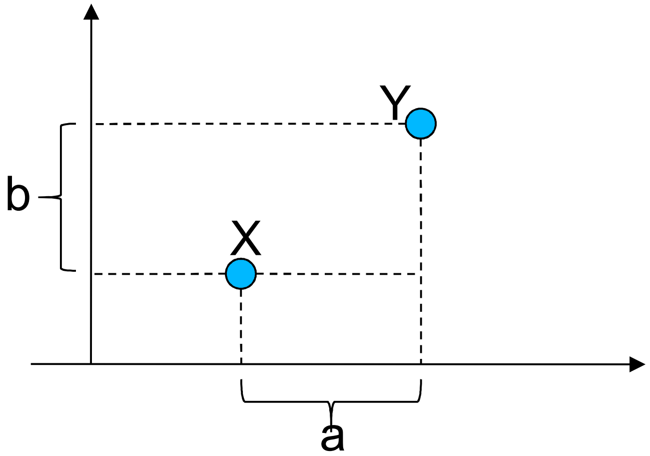
Minkowski Metric

A general definition for distance measures

$$L_k(x,y) = \left(\sum_{i=1}^{d} |x_i - y_i|^k \right)^{1/k}$$

L_1 norm – City block (Manhattan) distance: useful in digital problems

L_2 norm – Euclidian



$$L_1 = a + b$$

So use different k's depending of your problem.

Computational Complexity

Consider n samples in d dimensions
in crude 1-nn rule, we measure the distance from X to all samples. $O(dn^2)$ for classification. (for Bayes, $O(d^2)$) (n=number of samples:d=dim.))

To reduce the costs, several approaches
-Partial Distance: Compare the partially calculated distance to already found closest sample.

-Search Tree approach
-Editing(condensing) already discussed.