الجامعة السورية الخاصة
SYRIAN PRIVATE UNIVERSITY

Week 11

كلية الهندسة المعلوماتية

مقرر تعلم الآلة

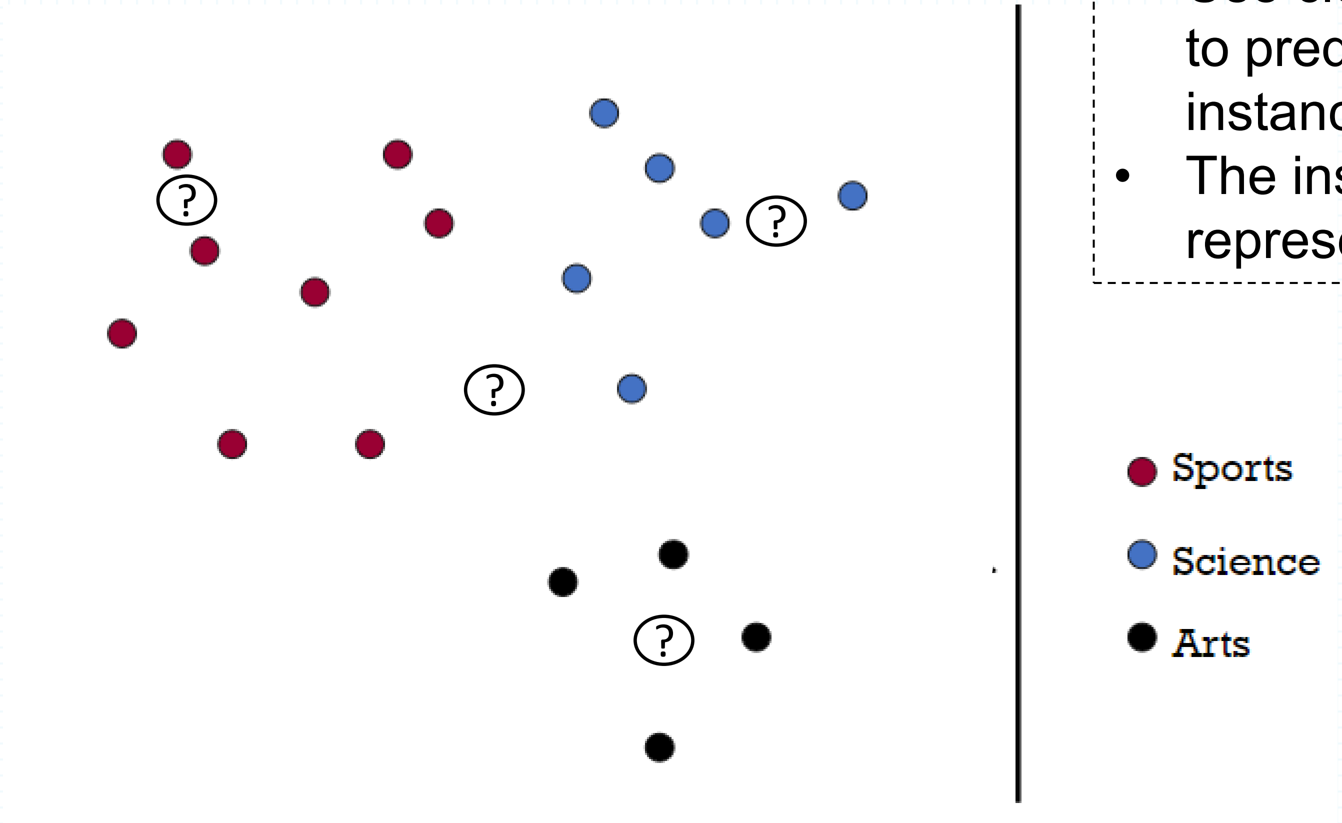
K Nearest Neighbor (KNN)

د. رياض سنبل

Basic Idea

- Each instance is represented as a vector of features.

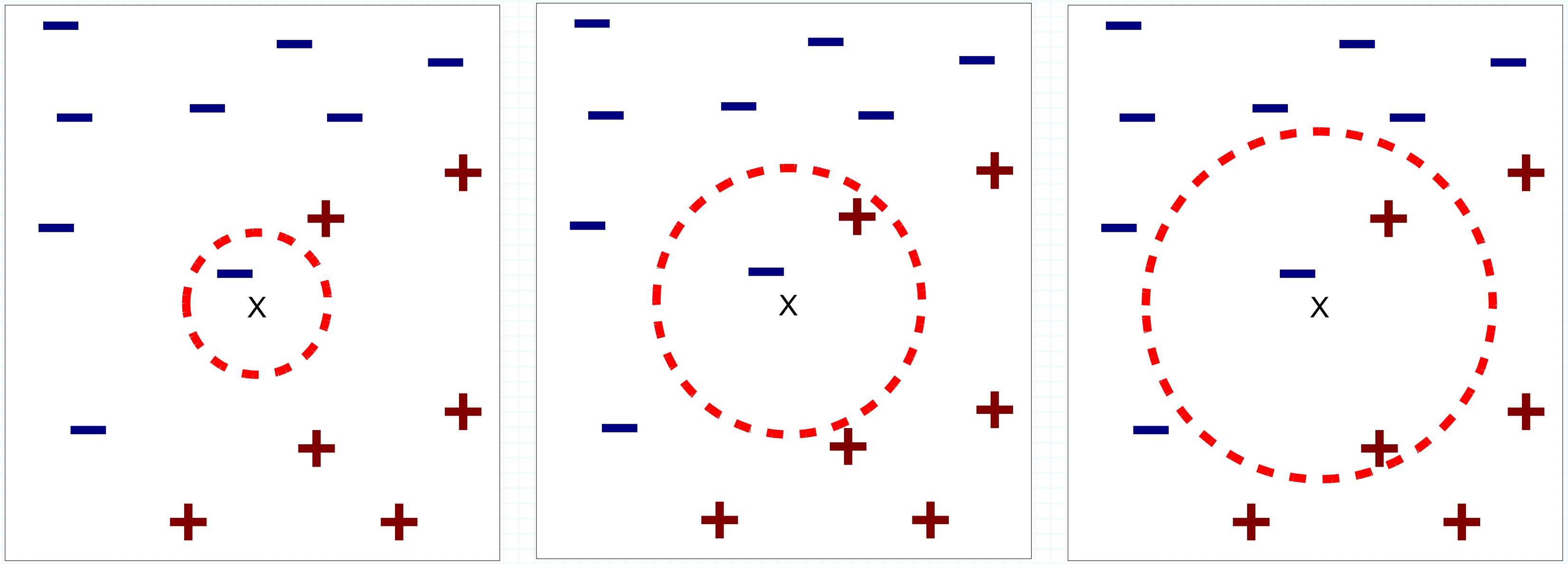
- Use closet training instances to predict the class of a new instance
- The instances themselves represent the knowledge!



Basic Idea

- *k*-NN classification rule is to assign to a test sample the majority category label of its *k* nearest training samples
- In practice, *k* is usually chosen to be odd, so as to avoid ties
- The *k* = 1 rule is generally called the nearest-neighbor classification rule

Definition of Nearest Neighbor

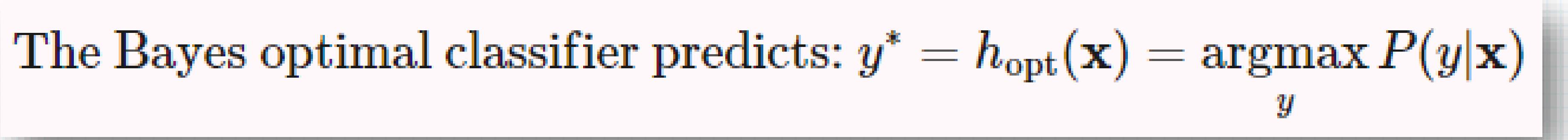


(a) 1-nearest neighbor (b) 2-nearest neighbor (c) 3-nearest neighbor

K-nearest neighbors of a record x are data points that have the k smallest distance to x

Bayes optimal classifier and NN

- Assume (and this is almost never the case) you knew P(y|x), then you would simply predict the most likely label.

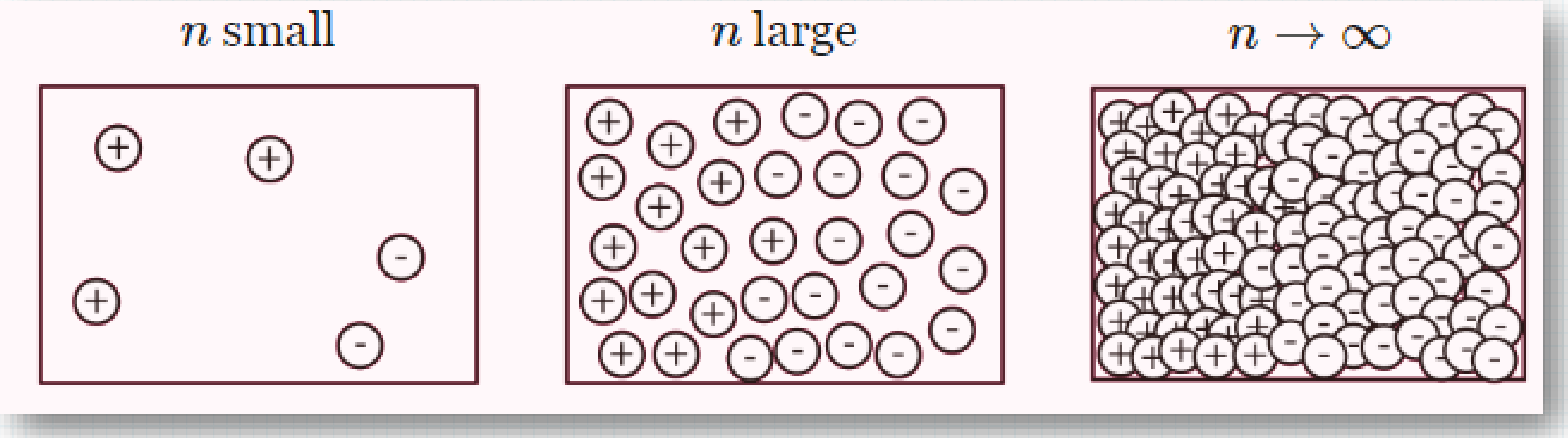


The Bayes optimal classifier predicts: $y^* = h_{\text{opt}}(\mathbf{x}) = \underset{y}{\operatorname{argmax}}\, P(y|\mathbf{x})$

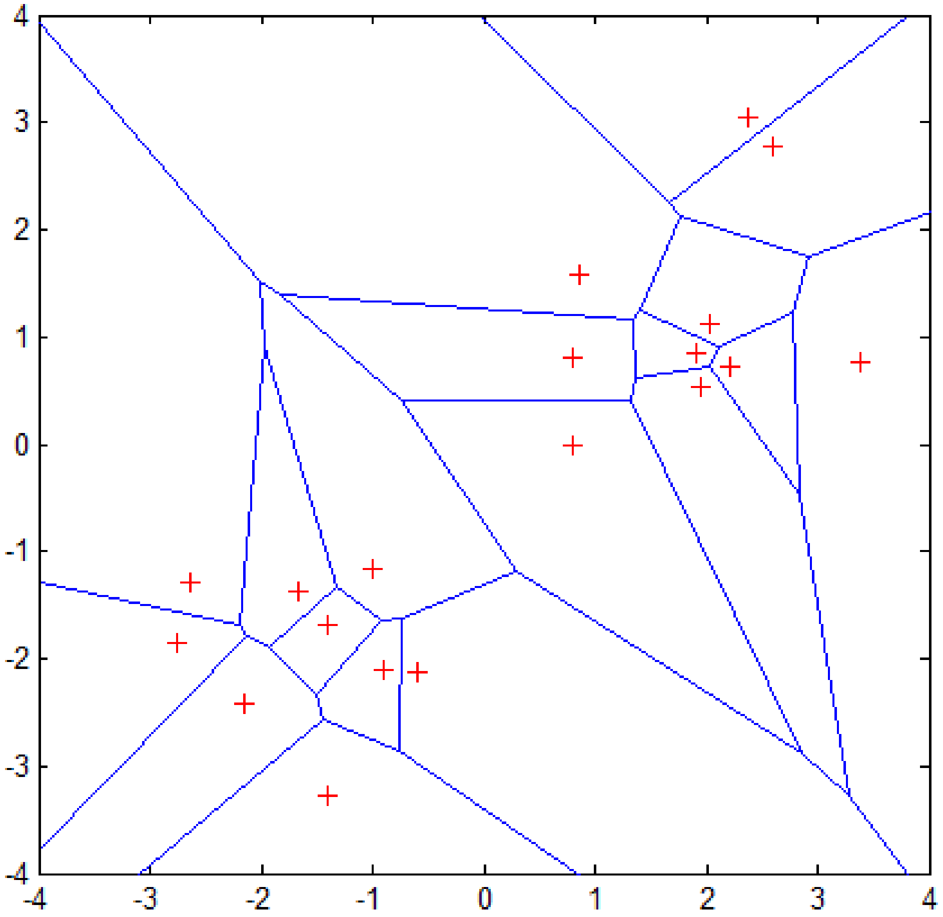
- Although the Bayes optimal classifier is as good as it gets, it still can make mistakes.
- **Why is the Bayes optimal classifier interesting, if it cannot be used in practice?** The reason is that it provides a highly informative lower bound of the error rate. With the same feature representation **no classifier can obtain a lower error.**

Bayes optimal classifier and NN

- As n→∞, the 1-NN classifier is only a factor 2 worse than the best possible classifier.



Voronoi Diagram

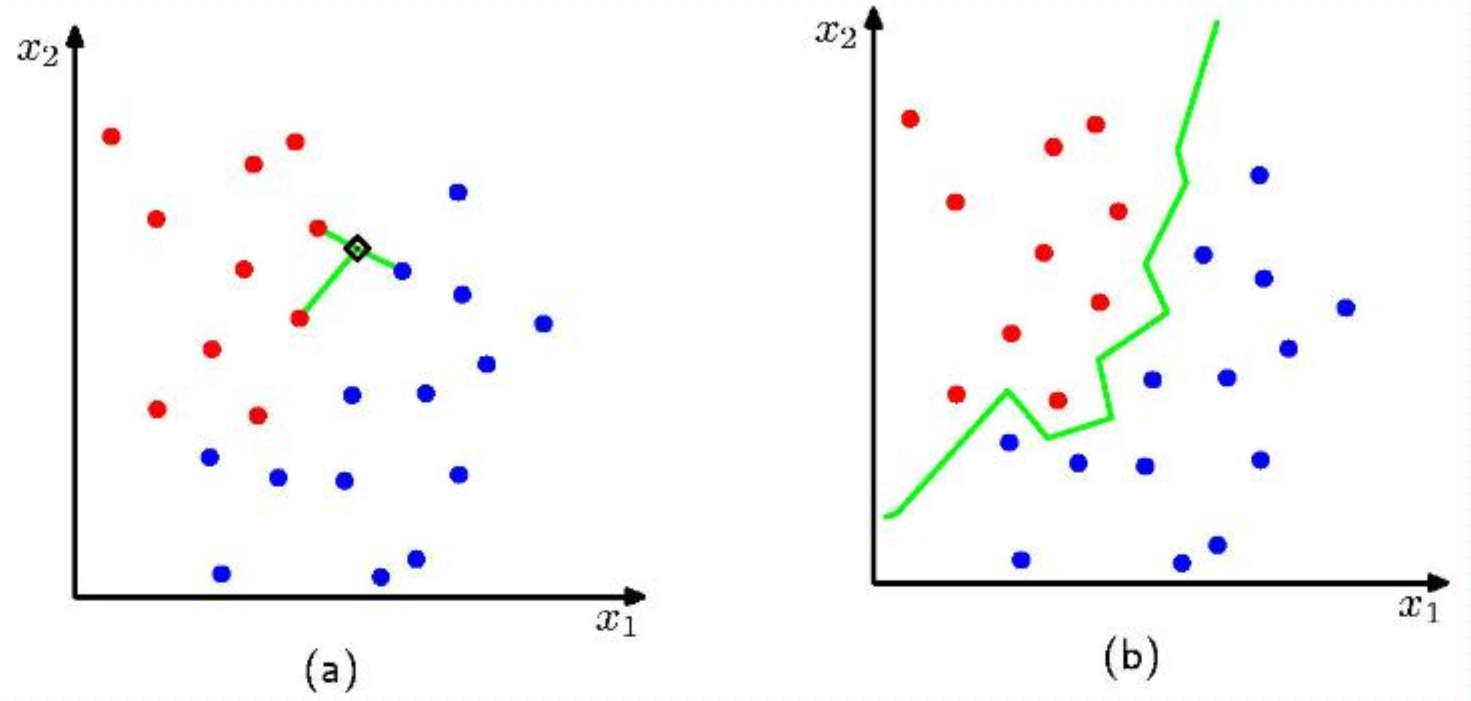


Properties:

1) All possible points within a sample's Voronoi cell are the nearest neighboring points for that sample
2) For any sample, the nearest sample is determined by the closest Voronoi cell edge

Decision boundary implemented by 3NN

- The boundary is always the perpendicular bisector of the line between two points (Vornoi tesselation)
  - k-nearest neighbors of a sample x are data points that have the k smallest distances to x



Nearest-Neighbor Classifiers: Issues

(1) The value of *k*, the number of nearest neighbors to retrieve

(2) Choice of Distance Metric to compute distance between records

(3) The weight of each neighbor.

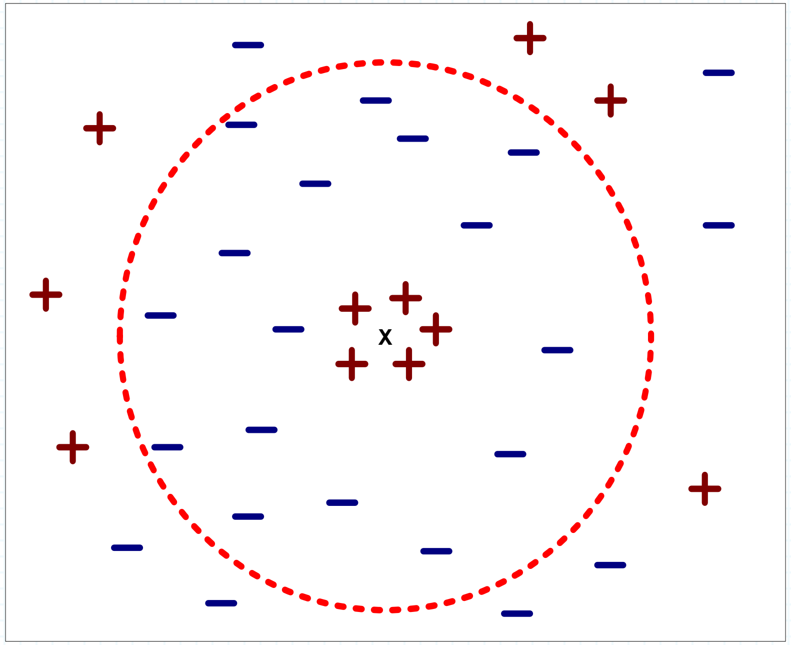
(4) Computational complexity

- Size of training set
- Dimension of data

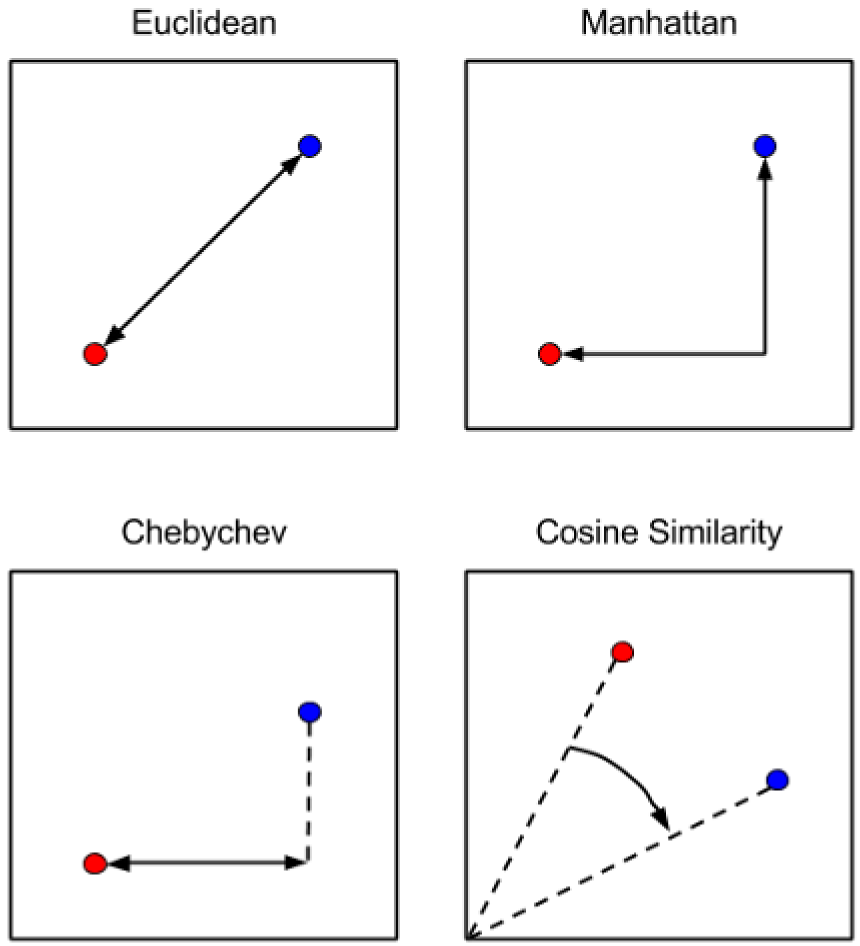
(1) Value of K

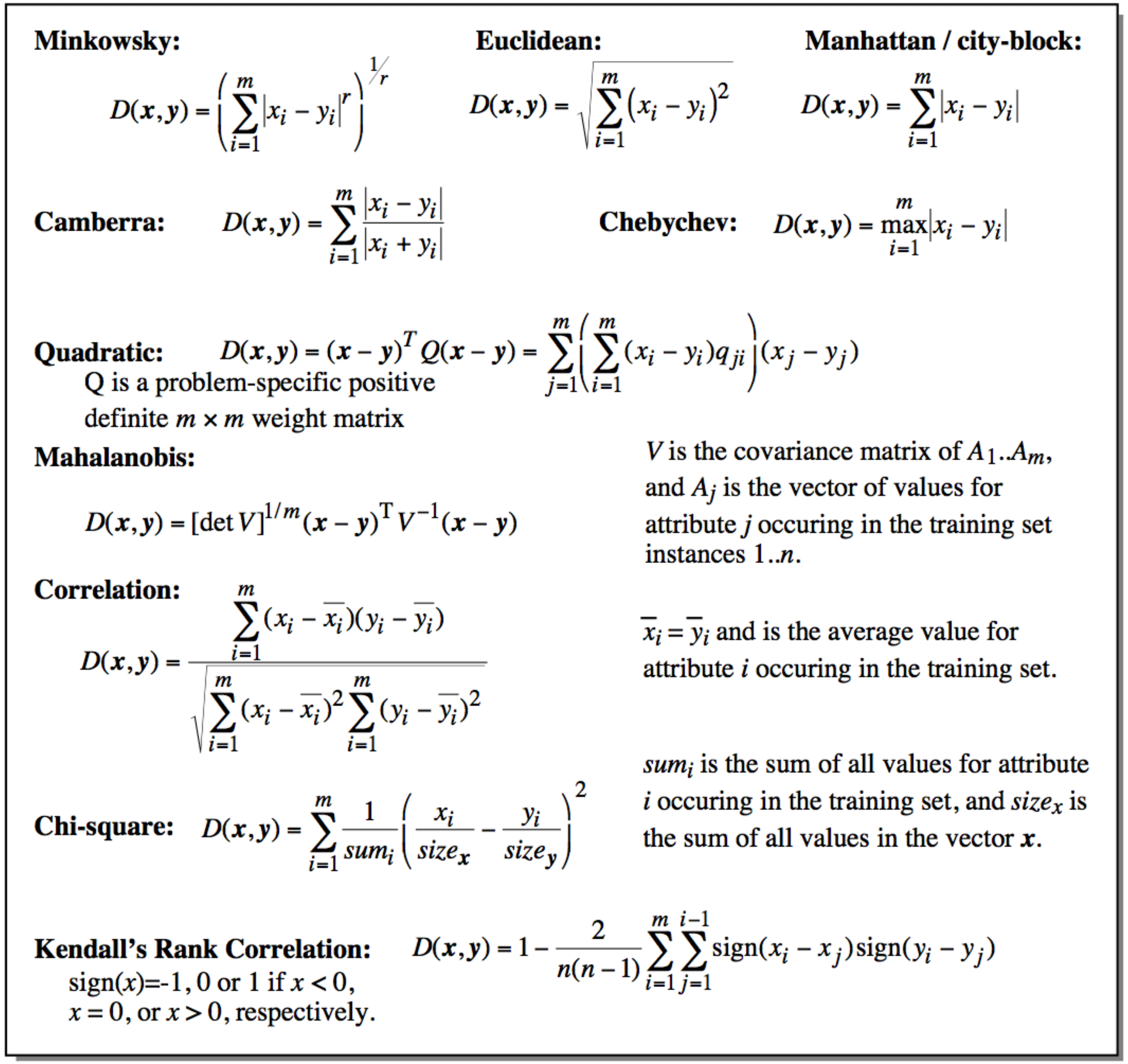
- Choosing the value of k:
  - If k is too small, sensitive to noise points
  - If k is too large, neighborhood may include points from other classes

Rule of thumb:
K = sqrt(N)
N: number of training points



(2) Distance Metrics





Minkowsky:

$$D(\mathbf{x},\mathbf{y}) = \left(\sum_{i=1}^{m} |x_i - y_i|^r\right)^{1/r}$$

Euclidean:

$$D(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{m} (x_i - y_i)^2}$$

Manhattan / city-block:

$$D(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} |x_i - y_i|$$

Camberra:

$$D(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} \frac{|x_i - y_i|}{|x_i + y_i|}$$

Chebychev:

$$D(\mathbf{x},\mathbf{y}) = \max_{i=1}^{m} |x_i - y_i|$$

Quadratic:

$$D(\mathbf{x},\mathbf{y}) = (\mathbf{x}-\mathbf{y})^T Q(\mathbf{x}-\mathbf{y}) = \sum_{j=1}^{m}\left(\sum_{i=1}^{m} (x_i - y_i) q_{ji}\right)(x_j - y_j)$$

Q is a problem-specific positive definite $m \times m$ weight matrix

Mahalanobis:

$$D(\mathbf{x},\mathbf{y}) = [\det V]^{1/m} (\mathbf{x}-\mathbf{y})^{\mathrm{T}} V^{-1} (\mathbf{x}-\mathbf{y})$$

V is the covariance matrix of $A_1..A_m$, and A_j is the vector of values for attribute j occuring in the training set instances $1..n$.

Correlation:

$$D(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{m} (x_i - \overline{x_i})(y_i - \overline{y_i})}{\sqrt{\sum_{i=1}^{m} (x_i - \overline{x_i})^2 \sum_{i=1}^{m} (y_i - \overline{y_i})^2}}$$

$\overline{x_i} = \overline{y_i}$ and is the average value for attribute i occuring in the training set.

Chi-square:

$$D(\mathbf{x},\mathbf{y}) = \sum_{i=1}^{m} \frac{1}{sum_i}\left(\frac{x_i}{size_{\mathbf{x}}} - \frac{y_i}{size_{\mathbf{y}}}\right)^2$$

sum_i is the sum of all values for attribute i occuring in the training set, and $size_x$ is the sum of all values in the vector $\mathbf{x}$.

Kendall's Rank Correlation:

$$D(\mathbf{x},\mathbf{y}) = 1 - \frac{2}{n(n-1)} \sum_{i=1}^{m}\sum_{j=1}^{i-1} \mathrm{sign}(x_i - x_j)\,\mathrm{sign}(y_i - y_j)$$

sign(x)=-1, 0 or 1 if $x < 0$, $x = 0$, or $x > 0$, respectively.

Figure 1. Equations of selected distance functions.
($\mathbf{x}$ and $\mathbf{y}$ are vectors of m attribute values).

(2) Distance Measure: Scale Effects

- Different features may have different measurement scales
  - E.g., patient weight in kg (range [50,200]) vs. blood protein values in ng/dL (range [-3,3])
- Consequences
  - Patient weight will have a much greater influence on the distance between samples
  - May bias the performance of the classifier

(2) Distance Measure: Standardization

- Transform raw feature values into z-scores

$$z_{ij} = \frac{x_{ij} - \mu_j}{\sigma_j}$$

  - x_{ij} is the value for the *i*th sample and *j*th feature
  - μ_j is the average of all x_{ij} for feature *j*
  - σ_j is the standard deviation of all x_{ij} over all input samples

- Range and scale of z-scores should be similar (providing distributions of raw feature values are alike)