# NEAREST NEIGHBORS METHODS

Week09

# $k$-NN

# Review: Types of Classifiers

- A classifier is a function that assigns to a sample, $\mathbf{x}$ a class label $\hat{y}$

$$\hat{y} = f(\mathbf{x})$$

- A probabilistic classifier obtains conditional distributions $\Pr(Y|\mathbf{x})$, meaning that for a given $\mathbf{x} \in \boldsymbol{X}$, they assign probabilities to all $y \in Y$
  - Hard classification

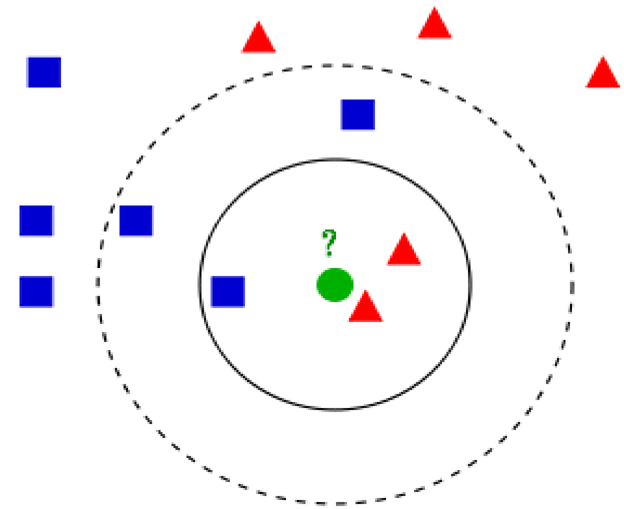
$$\hat{y} = \arg\max_{y} \Pr(Y = y|\,\mathbf{x})$$

**Any other classifiers not belonging to a probabilistic approach?**

# $k$-Nearest Neighbors($k$NN)

- Nonparametric method used for classification and regression

- For classification
  - Output class of data sample is determined by output class of its $k$-nearest neighbors
  - Majority vote
    - assign the output class to the most common class among $k$-nearest neighbors




- For regression
  - Output value of data sample is determined by output value of its $k$-nearest neighbors of the data sample
  - Output value is the average value of $k$-nearest neighbors
    - There are several different ways to calculate average

# ※ What is Nonparametric Method

- Parametric
  - Assume that data are drawn from a specific form of function up to unknown parameters
    - Linear regression, logistic regression

- Nonparametric
  - Assume that data are drawn from a certain unspecified function
  - Unlike parametric methods, there is no single global model
  - Learn to find patterns from training set and interpolate
  - Heavier computational cost than parametric ones

# Distance Measure

- Distance is a numerical description of how far apart objects are
  - Euclidean distance, one of distance measures, is common
    - Euclidean distance of two-dimensional data points, $(x_1, y_1), (x_2, y_2)$

$$d = \sqrt{(x_1 - x_2)^2 + (y_1 - y_2)^2}$$

    - In general, Euclidean distance of two data points, $(x_1, x_2, , \dots, x_n), (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$

$$d = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2 + \cdots + (x_n - y_n)^2}$$

- Other distance measures
  - 1-norm distance(Manhattan distance)

$$\sum_i^n |x_i - y_i|$$

  - $p$-norm distance (when $p$=2→ Euclidean distance)

$$\left(\sum_i^n (x_i - y_i)^p\right)^{1/p}$$

# Distance Measure

- Distance measure should hold the following
  - $d(x, y) \geq 0$
    - Non-negativity
  - $d(x, y) = 0 \Leftrightarrow x = y$
    - Identity of indiscernibles
  - $d(x, y) = d(y, x)$
    - *symmetry*
  - $d(x, z) \leq d(x, y) + d(y, z)$
    - Subaddivity or triangle inequality

# Distance Measure

- What if variables are not numerical
  - Other metrics are required for categorical variables

- Metrics for categorical variables
  - Hamming distance

$$d(x, y) = \frac{\sum_i I(x_i \neq y_i)}{\dim(x)}$$

    - $I(x_i \neq y_i)$ is 1 if and only if $x_i \neq y_i$
    - $\dim(x)$ is the dimension of $x$

# Distance Measure

- Metrics for categorical variables
  - Jaccard distance
    - Used to calculate the distance between binary vectors

| | | $y$ | |
|---|---|---|---|
| | | 0 | 1 |
| $x$ | 0 | $a$ | $b$ |
| | 1 | $c$ | $d$ |

- $a$: the total number of attributes where $x$ and $y$ both have a value of 0
- $b$: the total number of attributes where the attribute of $x$ is 0 and the attribute of $y$ is 1
- $c$: the total number of attributes where the attribute of $x$ is 1 and the attribute of $y$ is 0
- $d$: the total number of attributes where $x$ and $y$ both have a value of 1

$$d(x, y) = \frac{b + c}{b + c + d}$$

# Question

- Find $k$-nearest neighbors based on given data points

1) Find $k$-nearest neighbors of 5th objects when $k$=3 using Euclidean distance

2) Find $k$-nearest neighbors of 5th objects when $k$=3 using Manhattan distance

| index | $x$ | $y$ |
|---|---|---|
| 1 | 1 | 1 |
| 2 | 2 | 3 |
| 3 | 4 | 6 |
| 4 | 3 | 1 |
| 5 | 2 | 4 |
| 6 | 4 | 0 |
| 7 | 7 | 5 |
| 8 | 6 | 2 |

# Feature Scaling

- Scale of variable affects on determination of nearest neighbors

- Which sample is the nearest neighbor of data sample 1?

| $i$ | $x_1$ | $x_2$ | $x_3$ | $x_4$ | $y$ |
|---|---|---|---|---|---|
| 1 | 9 | 30 | 100 | 0.5 | 1 |
| 2 | 9 | 25 | 250 | 0.1 | 0 |
| 3 | 9 | 44 | 220 | 0.7 | 0 |
| 4 | 7.5 | 75 | 170 | 1.2 | 1 |
| | … | … | … | … | … |

| $i$ | Distance from $p_1$ |
|---|---|
| 1 | - |
| 2 | 150.0838 |
| 3 | 120.8141 |
| 4 | 83.23305 |
| | … |

  - Scale of variable $x_3$ dominates over other variables
  - The nearest neighbor is strongly dependent on $x_3$

**It is unfair!**

# Normalization

- Normalization is to adjust values of variables with different scales to common scale
  - There are several different ways for normalization

- Commonly used normalization method

$$x \rightarrow \frac{x-\mu}{\sigma}$$

  - $\mu$=mean value of the variable
  - $\sigma$=standard deviation of the variable
  - $\mu$ and $\sigma$ are computed by sample data points

$$x \rightarrow \frac{x-x_{min}}{x_{max}-x_{min}}$$

  - $x_{max}$ is the maximum value of variable $x$ and $x_{min}$ is the minimum value of variable $x$
  - Normalized value is within [0, 1]

# Mahalanobis Distance

- Normalization based on normal distribution($x \rightarrow \frac{x-\mu}{\sigma}$) assumes that the sample points are distributed about the center of mass in a spherical manner
  - In real data, variables are correlated with other variables

**Need to consider scale (level of spread along axis) and correlation to measure distance**

$\Downarrow$

**Mahalonobis distance**

# Mahalanobis Distance

- Mahalanobis distance

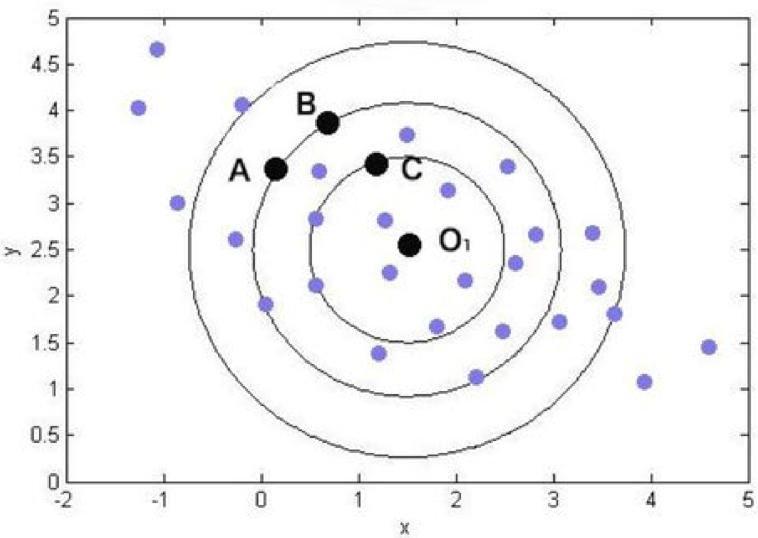
$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{(\mathbf{x}_1 - \mathbf{x}_2)^T S^{-1} (\mathbf{x}_1 - \mathbf{x}_2)}$$

  - $S$ is sample covariance matrix
  - If covariance matrix is diagonal(no correlation), the resulting distance measure is as the same as the standardized distance

$$d(\mathbf{x}_1, \mathbf{x}_2) = \sqrt{\sum_{i=1}^{p} \frac{(x_{1i} - x_{2i})^2}{s_i^2}}$$

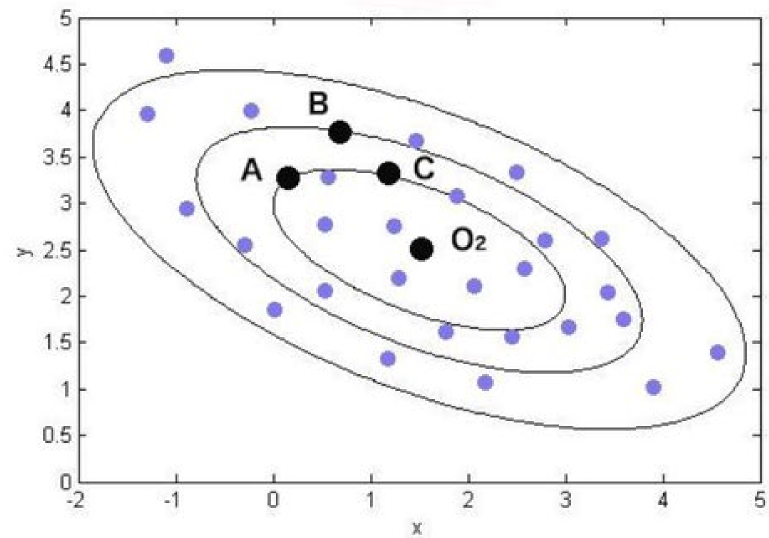
# Mahalanobis Distance

- Comparison between Euclidean distance and Mahalanobis distance




**(a)**

Euclidean distance




**(b)**

Mahalanobis  distance

# Procedure of $k$NN

Decide the number of nearest neighbors $k$ and distance measure

↓

For all data point in test set, find $k$ nearest neighbors

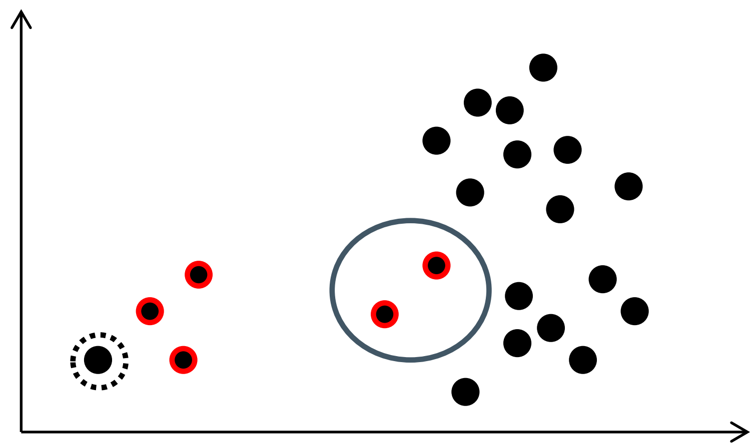
↓

Obtain output value based on output values of neighbors

# Fixed-radius Near Neighbors

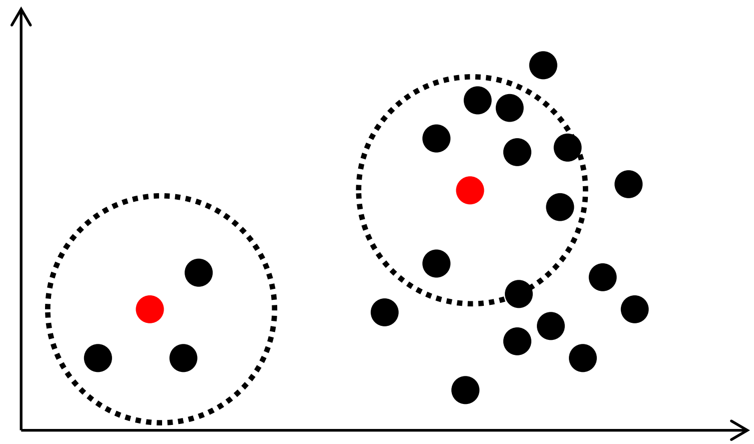
# Problem of Fixed-Number of Nearest Neighbors

- When distribution of data set is not homogenous, samples not similar to data point $x$ can be obtained in the nearest neighbors
  - $k = 5$



# Fixed-Radius Near Neighbors

- Fixed-radius near neighbors are neighbors within fixed range from data point $x$
  - Because of that, the number of neighbors may be different depending on the location



# Fixed-Radius Near Neighbors Methods

- The only difference of fixed-radius NN from $k$NN is the method to find the nearest neighbors
  - Remained steps of classification and regression are the same

Decide radius of range from data point and distance measure

↓

For all data point in test set, find fixed-radius near neighbors

↓

Obtain output value based on output values of neighbors