# k-NEAREST NEIGHBOR ALGORITHM

Pronalaženje skrivenog znanja

Bojan Furlan

# SUPERVISED VERSUS UNSUPERVISED METHODS

- In unsupervised methods, no target variable is identified (e.g. clustering).
- Most data mining methods are supervised methods meaning that:
  - there is a particular prespecified target variable
  - the algorithm is given many examples
  - where the value of the target variable is provided
  - learning which values of the target variable are associated with which values of the predictor variables.
- Supervised methods are including decision trees, neural networks and k-nearest neighbors.

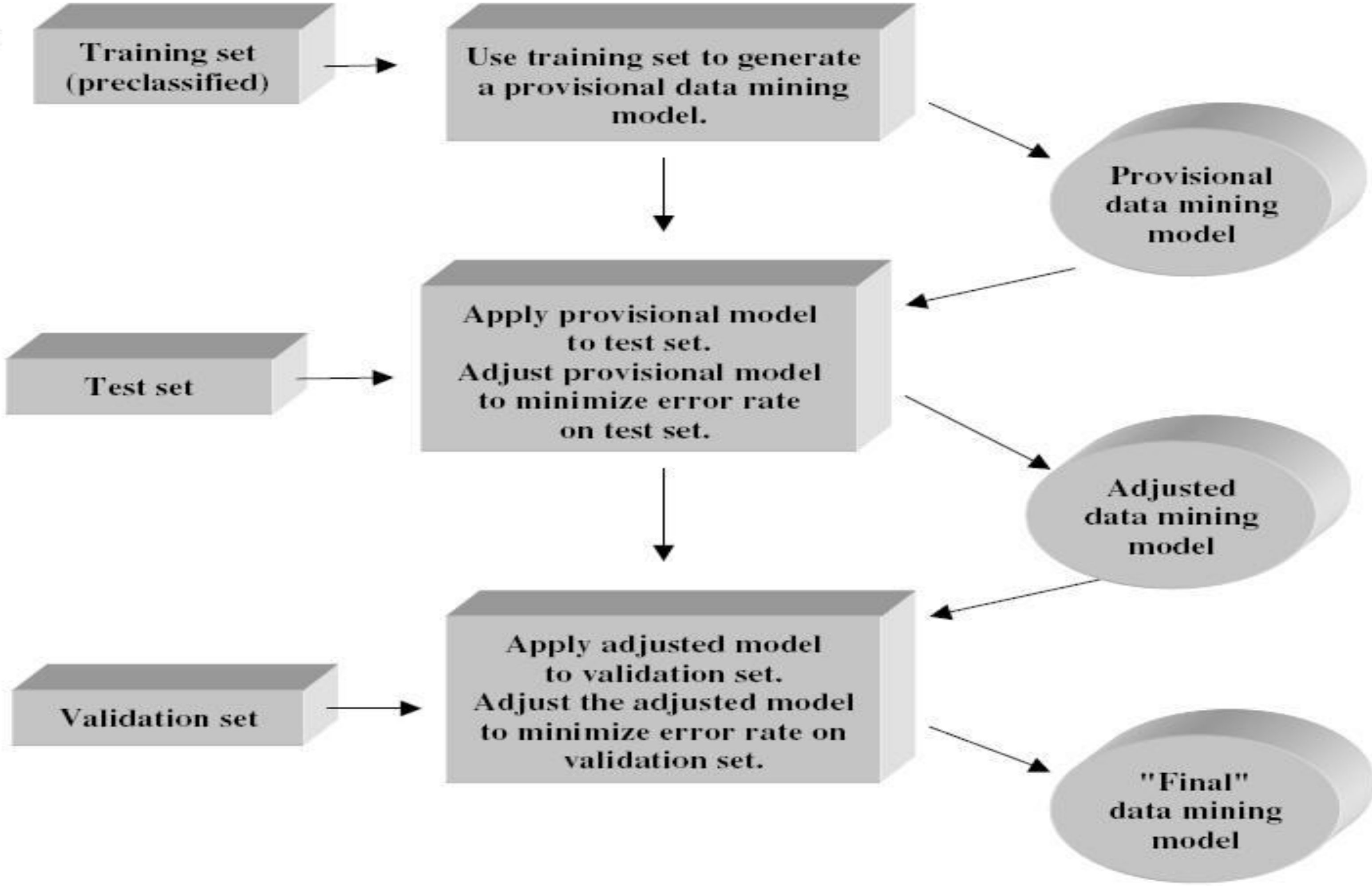
# CLASSIFICATION TASK

| Subject | Age | Gender | Occupation | Income Bracket |
|---|---|---|---|---|
| 001 | 47 | F | Software engineer | High |
| 002 | 28 | M | Marketing consultant | Middle |
| 003 | 35 | M | Unemployed | Low |
| ⋮ | | | | |

Excerpt from Data Set for Classifying Income

- Classifying the income bracket of persons not currently in the database
  - based on the other characteristics associated with that person such as age, gender, and occupation.

# METHODOLOGY FOR SUPERVISED MODELING

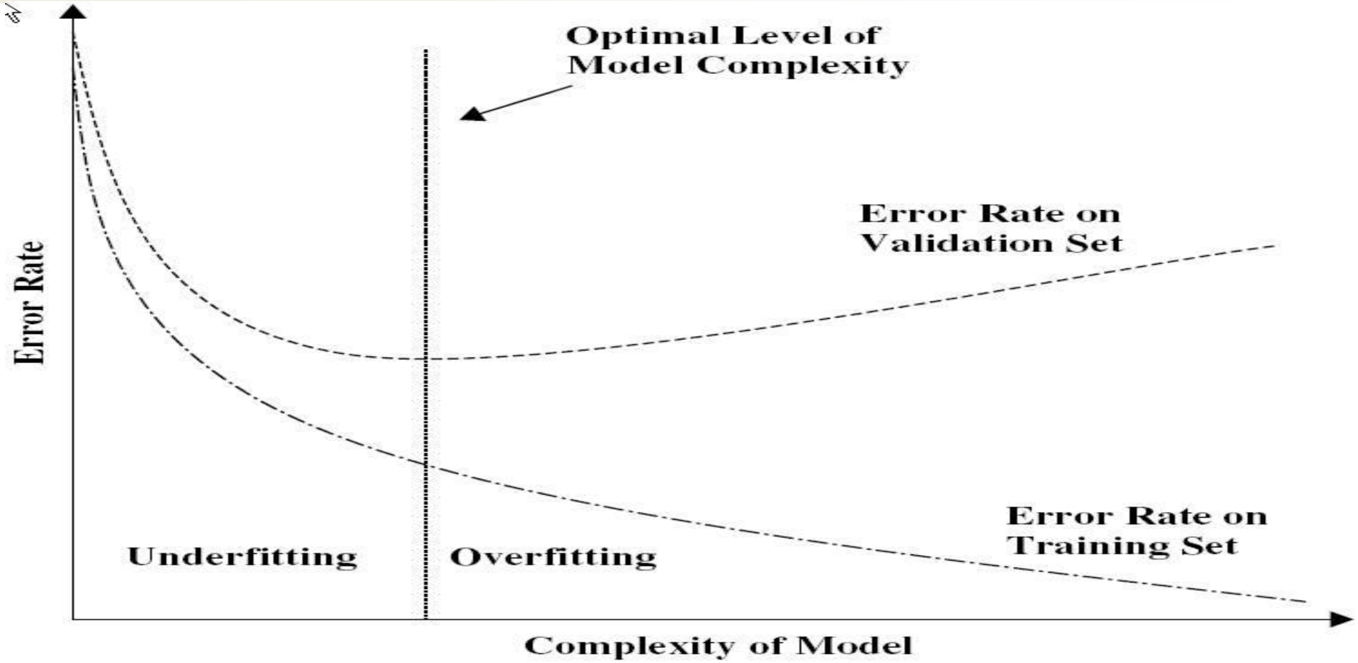



Methodology for supervised modeling

# METHODOLOGY FOR SUPERVISED MODELING

- Training set is necessarily incomplete, it doesn't include future data that we want to classify.
- Algorithm needs to guard against "memorizing" the training set
  - E.g. all customers named "David" in a training set may be in the high-income bracket.
  - => "If the customer's first name is David, the customer has a high income."

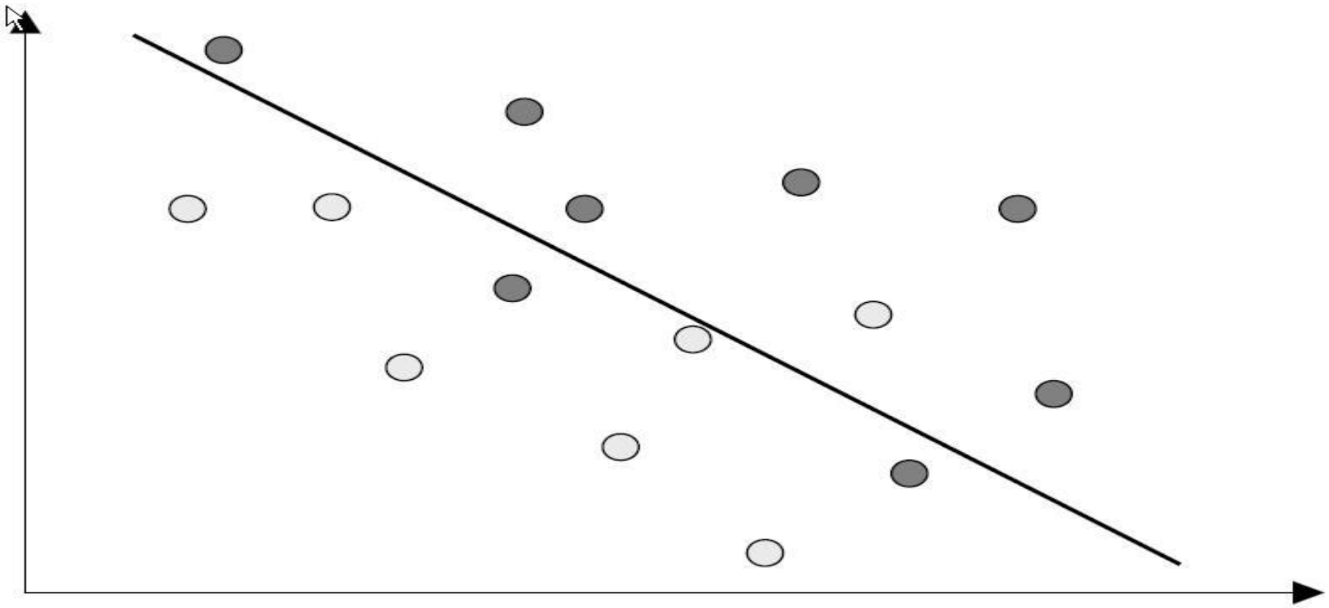
# COMPLEXITY OF MODEL




The optimal level of model complexity is at the minimum error rate on the validation set

- Overfitting results when the model tries to account for every possible trend or structure in the training set
  - "David" example.
- Increasing the complexity of the model leads to a degradation in the <u>generalizability</u>.

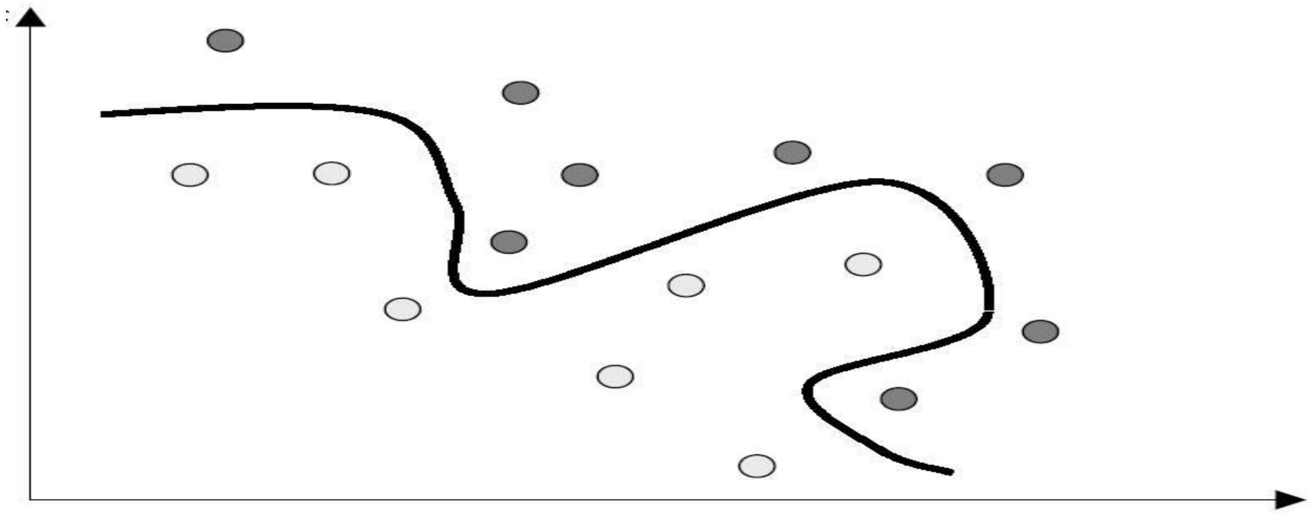
# COMPLEXITY TRADE-OFF Example



Low-complexity separator with high error rate

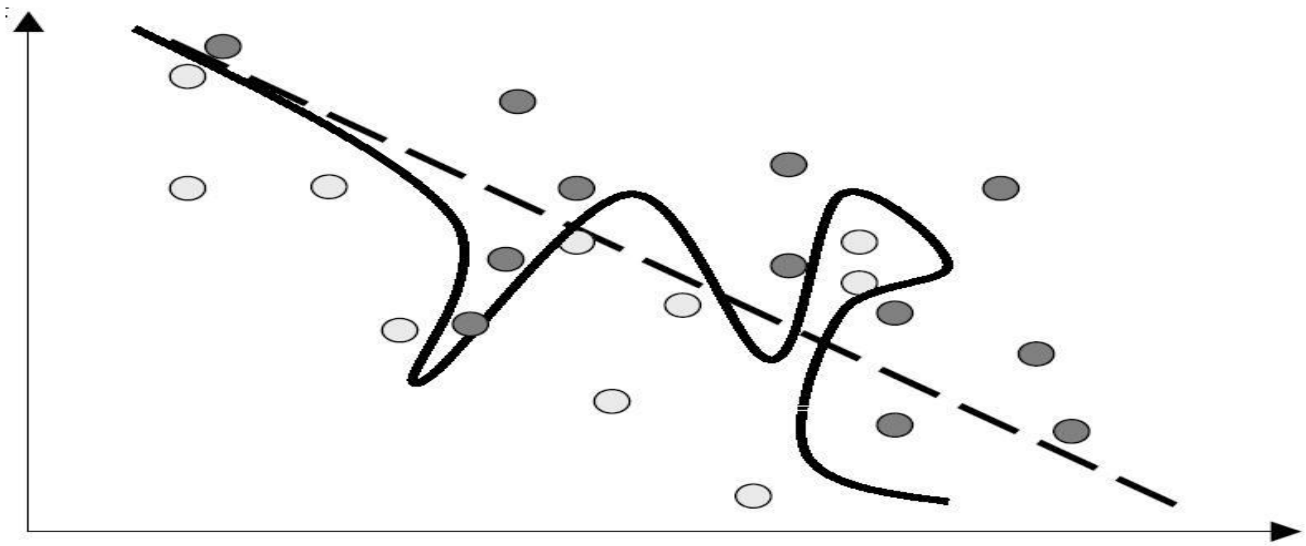
- Constructing the optimal curve (or straight line) that will separate the dark gray points from the light gray points.

# COMPLEXITY TRADE-OFF Example



High-complexity separator with low error rate

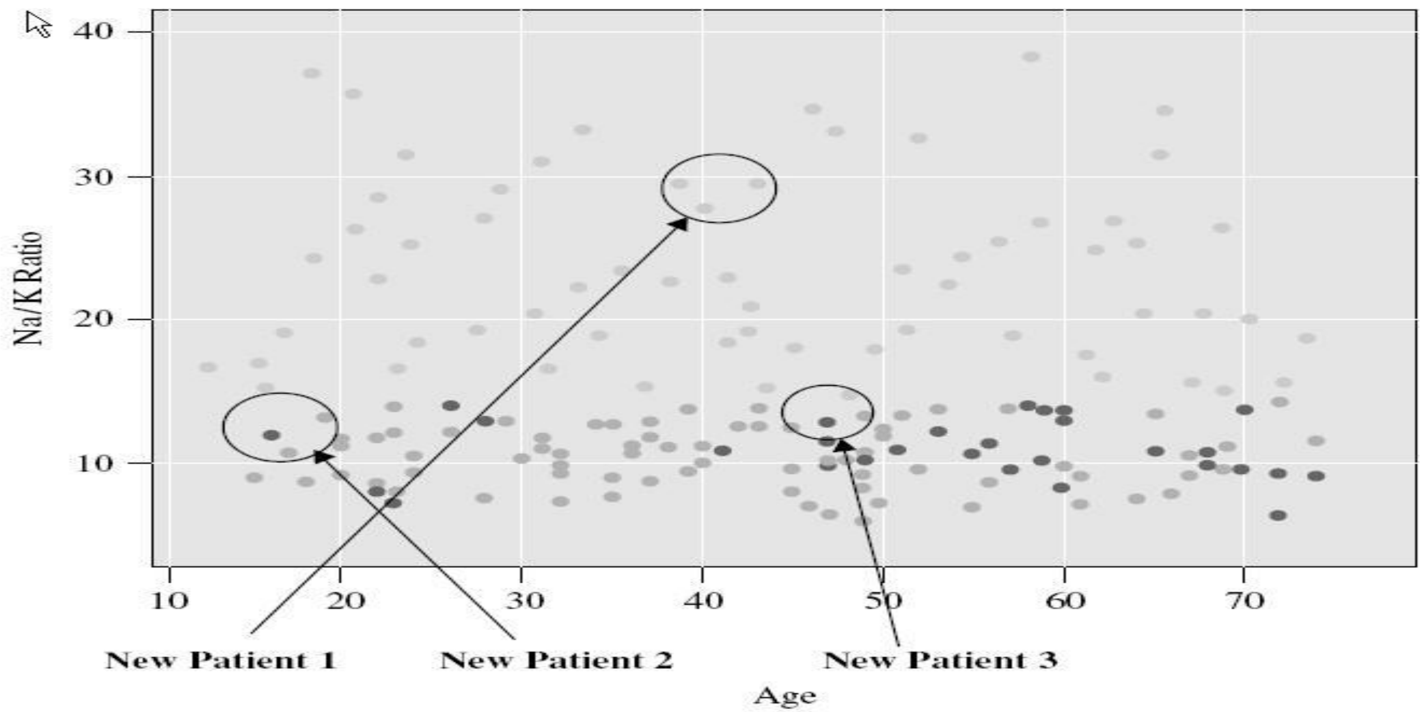
# COMPLEXITY TRADE-OFF Example



With more data:
Low-complexity separator doesn't need to change much;
High-complexity separator needs much revision

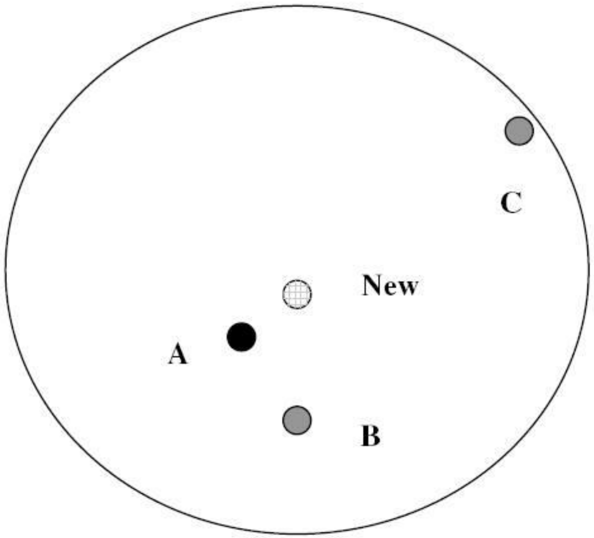
# k-NEAREST NEIGHBOR ALGORITHM

- k -Nearest neighbor is an example of instance-based learning
- classification for a new unclassified record is done by comparing it to the most similar records in the training set. (S kim si, takav si ☺ )




Plot of Na/K ratio against age, with drug overlay

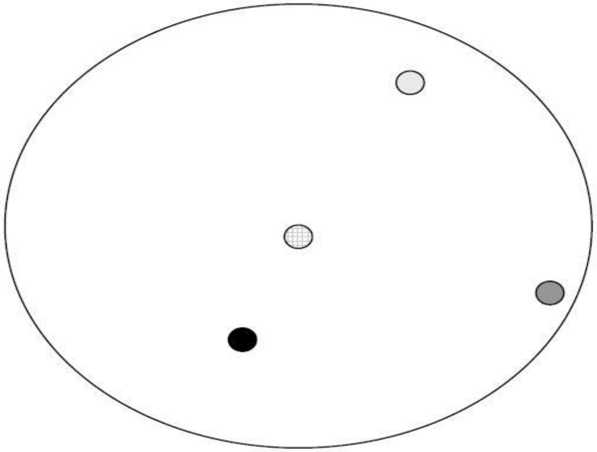
# k-NEAREST NEIGHBOR ALGORITHM




Close-up of three nearest neighbors to new patient 2

- Patient 2 who is 17 years old with a Na/K ratio of 12.5.
- Let k = 1,2,3 for our k-nearest neighbor algorithm
- Classification assigned for new patient 2 differed based on which value we chose for k .

# k-NEAREST NEIGHBOR ALGORITHM



Close-up of three nearest neighbors to new patient 3

- Patient 3, who is 47 years old and has a Na/K ratio of 13.5.
- For k = 2 voting would not help.
- Voting would not help for k = 3 in this case either, since the three nearest neighbors to new patient 3 are of three different classifications.

# k-NEAREST NEIGHBOR ALGORITHM

- some of the issues involved in building a classifier:
  - How do we measure distance?
  - How do we combine the information from more than one observation?
  - Should all points be weighted equally, or should some points have more influence than others?
  - How many neighbors should we consider? That is, what is k ?

# DISTANCE FUNCTION

- For a new record, the k-nearest neighbor algorithm assigns the classification of the most similar record or records.
- How do we define similar?
- The most common distance function is Euclidean distance:

$$d_{\text{Euclidean}}(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_i (x_i - y_i)^2}$$

where $x = x_1, x_2, \ldots, x_m$, and $y = y_1, y_2, \ldots, y_m$ represent the m attribute values of two records.

# DISTANCE FUNCTION – Problems

- Attributes with large values can overwhelm the influence of attributes measured on a smaller scale:
  - e.g. income vs. age
- To avoid this normalize the attribute values.
- For continuous variables, the min-max normalization or Z-score standardization.
- For categorical variables, Euclidean distance metric is not appropriate.
- we define a function, "different from," used to compare the i^th attribute values of a pair of records:

$$\text{different}(x_i, y_i) = \begin{cases} 0 & \text{if } x_i = y_i \\ 1 & \text{otherwise} \end{cases}$$

where $x_i$ and $y_i$ are categorical values.

# DISTANCE FUNCTION - Example

| Patient | Age | $Age_{MMN}$ | $Age_{Zscore}$ | Gender |
|---|---|---|---|---|
| A | 50 | $\frac{50-10}{50}=0.8$ | $\frac{50-45}{15}=0.33$ | Male |
| B | 20 | $\frac{20-10}{50}=0.2$ | $\frac{20-45}{15}=-1.67$ | Male |
| C | 50 | $\frac{50-10}{50}=0.8$ | $\frac{50-45}{15}=0.33$ | Female |

Variable Values for Age and Gender

- Which patient is more similar to a 50-year-old male: a 20-year-old male or a 50-year-old female?
- The original variable values, with the min-max normalization ($age_{MMN}$) and Z -score standardization ($age_{Zscore}$) in the table.

# DISTANCE FUNCTION - Example

- What happens when we forget to normalize the age variable? Distance between patients A and B is:

$$d(A,B) = \sqrt{(50-20)^2 + 0^2} = 30$$

Distance between patients A and C is:

$$d(A,C) = \sqrt{(50-50)^2 + 1^2} = 1$$

- The problem is that the age variable is measured on a larger scale

# DISTANCE FUNCTION - Example

- min-max normalization:

$$d_{MMN}(A,B) = \sqrt{(0.8-0.2)^2 + 0^2} = 0.6$$

$$d_{MMN}(A,C) = \sqrt{(0.8-0.8)^2 + 1^2} = 1.0$$

Means that patient B is now more similar to patient A.

- Z -score standardization:

$$d_{Zscore}(A,B) = \sqrt{[0.33-(-1.67)]^2 + 0^2} = 2.0$$

$$d_{Zscore}(A,C) = \sqrt{(0.33-0.33)^2 + 1^2} = 1.0$$

Means that patient C is again closer.

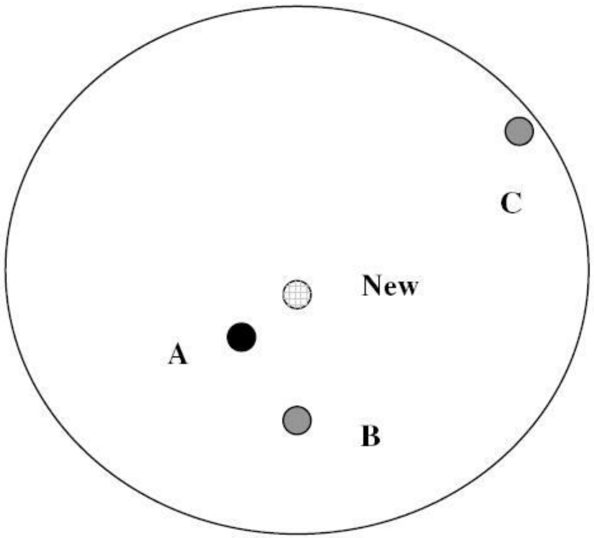
# DISTANCE FUNCTION – Conclusion

- The min-max normalization will almost always lie between zero and 1.
- Z -score standardization, takes values $-3 < z < 3$, representing a wider scale than that of the min-max normalization.
- When mixing categorical and continuous variables, the min-max normalization may be preferred.
  - "different from" function is in range [0,1].
- Other distance measures:

$$d_{Cosine}(x,y) = 1 - \frac{\sum_i^n (x_i * y_i)}{\sqrt{\sum_i^n (x_i)^2} * \sqrt{\sum_i^n (y_i)^2}}$$

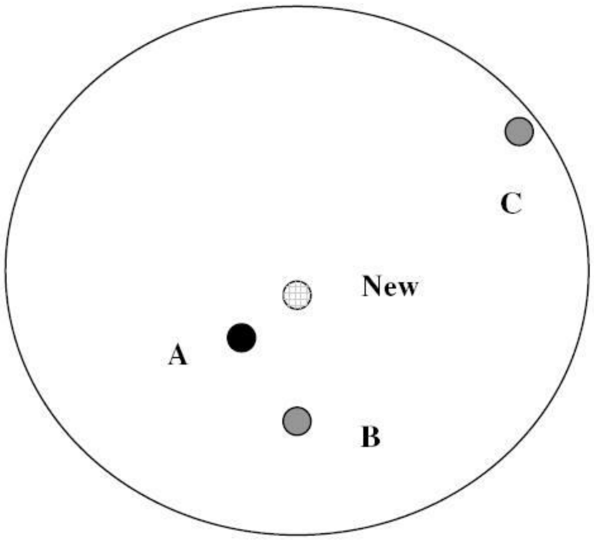
# COMBINATION FUNCTION

- How to combine similar records to provide a classification decision for the new record?
- **Simple Unweighted Voting**
- The classification would be made with confidence 2/3 = 66.67%,
- where the confidence level represents the count of records, with the winning classification divided by k .



# COMBINATION FUNCTION

- **Weighted Voting**
- Neighbors that are closer or more similar to the new record are weighted more heavily than more distant neighbors.
  - Eg A is closer than C



# COMBINATION FUNCTION – Example

| Record | Age | Na/K | $Age_{MMN}$ | $Na/K_{MMN}$ |
|---|---|---|---|---|
| New | 17 | 12.5 | 0.05 | 0.25 |
| A (dark gray) | 16.8 | 12.4 | 0.0467 | 0.2471 |
| B (medium gray) | 17.2 | 10.5 | 0.0533 | 0.1912 |
| C (medium gray) | 19.5 | 13.5 | 0.0917 | 0.2794 |

Age and Na/K Ratios for Records from Figure

- Example, drug classification for a new record, using the k = 3 nearest neighbors.

$$d(\text{new},A) = \sqrt{(0.05 - 0.0467)^2 + (0.25 - 0.2471)^2} = 0.004393$$

$$d(\text{new},B) = \sqrt{(0.05 - 0.0533)^2 + (0.25 - 0.1912)^2} = 0.58893$$

$$d(\text{new},C) = \sqrt{(0.05 - 0.0917)^2 + (0.25 - 0.2794)^2} = 0.051022$$

# COMBINATION FUNCTION – Example

- The votes of these records are then weighted according to the inverse square of their distances.
- One record (A)

$$\text{votes (dark gray)} = \frac{1}{d(\text{new}, A)^2} = \frac{1}{0.004393^2} \simeq 51817.63$$

- Two records (B and C)

$$\text{votes (medium gray)} = \frac{1}{d(\text{new}, B)^2} + \frac{1}{d(\text{new}, C)^2} = \frac{1}{0.058893^2} + \frac{1}{0.051022^2} \simeq 672$$

- Dark gray – winner !!!

# QUANTIFYING ATTRIBUTE RELEVANCE: STRETCHING THE AXES

- Not all attributes may be relevant to the classification.
- By default in the k -nearest neighbor algorithm, the distances are calculated on all the attributes.
- Determining which fields are more or less important is equivalent to finding a coefficient $z_j$ by which to multiply the $j^{th}$ axis
  - larger values of $z_j$ => more important variable axes.

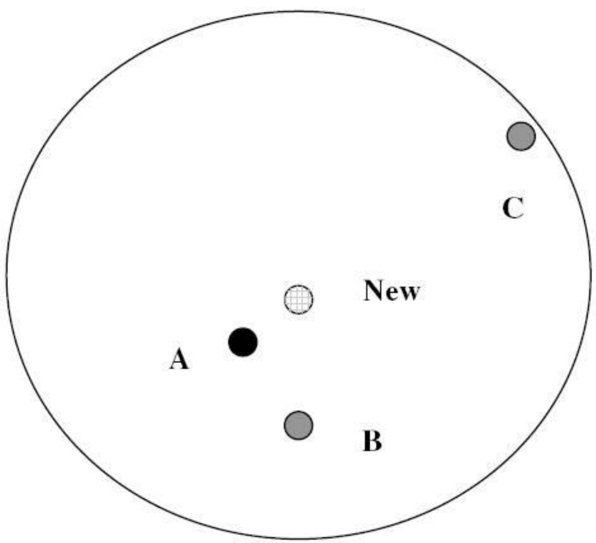
# STRETCHING THE AXES - Example

- From experts: zNa/K  = 3 and zAge  = 1.

$$d(\text{new}, A) = \sqrt{(0.05 - 0.0467)^2 + [3(0.25 - 0.2471)]^2} = 0.009305$$

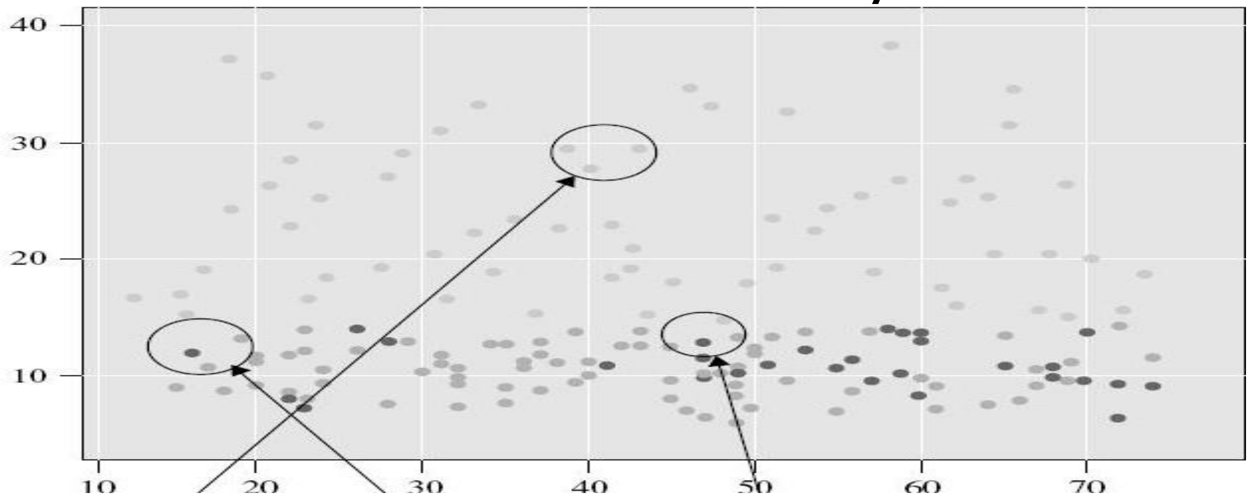
$$d(\text{new}, B) = \sqrt{(0.05 - 0.0533)^2 + [3(0.25 - 0.1912)]^2} = 0.17643$$

$$d(\text{new}, C) = \sqrt{(0.05 - 0.0917)^2 + [3(0.25 - 0.2794)]^2} = 0.09756$$



# DATABASE CONSIDERATIONS

- Rare classifications must be represented sufficiently
  - so algorithm does not only predict common classifications.
- On the other hand, big amount of data can take significant time to process -> Because of distance calculation
- Optimization – One method to is to reduce the proportion of records with more common classifications.
- E.g. New records with Na/K ratio > 19 could be omitted, because all can be classified similarly.



# k-NEAREST NEIGHBOR ALGORITHM FOR ESTIMATION AND PREDICTION

**Locally weighted averaging** method

$$y_{\text{new}} = \frac{\sum_i w_i y_i}{\sum_i w_i}$$

- weighted average for the *k nearest neighbors*
- Wi – inverse square of the distances from i-th neighbor
- Yi – value of the i-th neighbor for estimated attribute

# k-NEAREST NEIGHBOR ESTIMATION - Example

| Record | Age | Na/K | BP | $Age_{MMN}$ | $Na/K_{MMN}$ |
|---|---|---|---|---|---|
| New | 17 | 12.5 | ? | 0.05 | 0.25 |
| A | 16.8 | 12.4 | 120 | 0.0467 | 0.2471 |
| B | 17.2 | 10.5 | 122 | 0.0533 | 0.1912 |
| C | 19.5 | 13.5 | 130 | 0.0917 | 0.2794 |

*k* = 3 Nearest Neighbors of the New Record

- Estimating Blood Presure - from range [80,160]
- $z_{Na/K}$ = three-axis-stretching to reflect greater importance of the Na/K ratio.

$$\hat{y}_{\text{new}} = \frac{\sum_i w_i y_i}{\sum_i w_i} = \frac{\frac{120}{0.009305^2} + \frac{122}{0.17643^2} + \frac{130}{0.09756^2}}{\frac{1}{0.009305^2} + \frac{1}{0.17643^2} + \frac{1}{0.09756^2}} = 120.0954.$$

# CHOOSING k

- Small k (e.g., k = 1), the algorithm will return the target value of the nearest observation
  - the algorithm may reach overfitting
  - tending to memorize the training data set => lack of generalizability.
- Choosing a value of k that is too large, locally interesting behavior will be overlooked.
- Choose the value of k that can be decidable (3,5...)
- Choose the value of k that minimizes the classification or estimation error on validation set.