VSB TECHNICAL UNIVERSITY OF OSTRAVA | FACULTY OF ELECTRICAL ENGINEERING AND COMPUTER SCIENCE | DEPARTMENT OF CYBERNETICS AND BIOMEDICAL ENGINEERING

Data Classification: Basic Methods, Principles, and Applications

Applied Artificial Intelligence in Medicine

doc. Eng. Karla Miriam Reyes, Ph.D.

Introduction to Data Classification

Data classification plays a crucial role in biomedical applications, enabling the categorization of data into meaningful groups based on their characteristics. This process is essential for:

Disease diagnosis

- Classifying medical images, patient data, and laboratory results to aid in accurate diagnosis and treatment.

Personalized medicine

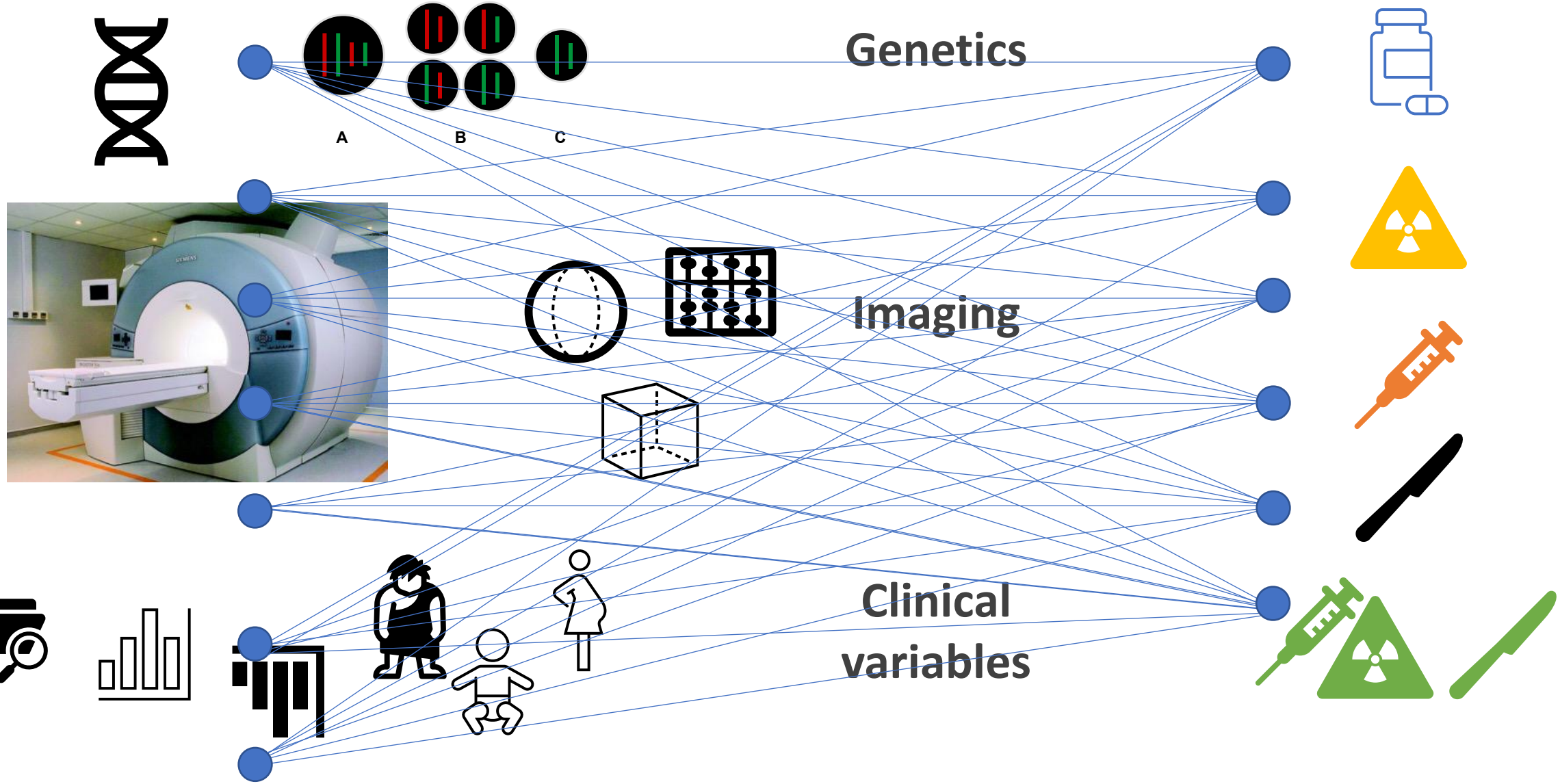
- Grouping patients based on their genetic profiles, medical history, and lifestyle to tailor treatment plans.

Clinical decision support

- Classifying patient data to provide healthcare professionals with relevant information for informed decisions.

Importance of Data Classification in Biomedical Applications

Improved diagnosis	Accurate classification of medical data enables timely and effective diagnosis.
Enhanced patient outcome	Personalized medicine and targeted treatment plans lead to better health outcomes.
Reduced healthcare costs	Data classification helps identify high-risk patients, reducing unnecessary procedures and hospitalizations.


A
B
C
Genetics
Imaging
Clinical variables

Real-World Applications in Medicine

Disease Diagnosis

1. **Diabetes**: Classifying patients into different risk groups based on their blood glucose levels, HbA1c, and other factors to determine the likelihood of developing complications.

   Example: A patient's blood glucose levels are classified as normal, pre-diabetic, or diabetic, enabling healthcare providers to recommend targeted interventions.

2. **Cancer**: Classifying tumors into different subtypes based on their genetic profiles, histology, and other characteristics to determine the most effective treatment options.

   Example: A patient's tumor is classified as breast cancer subtype A or B, guiding the selection of chemotherapy and targeted therapies.

Real-World Applications in Medicine

Patient Risk Stratification

1. **Cardiovascular disease**: Classifying patients into high, medium, or low risk groups based on their cardiovascular risk factors, such as age, blood pressure, and cholesterol levels.

   Example: A patient is classified as high risk, prompting healthcare providers to recommend lifestyle modifications and medication to reduce the risk of cardiovascular events.

2. **Sepsis**: Classifying patients into different risk groups based on their clinical presentation, laboratory results, and other factors to determine the likelihood of developing sepsis.

   Example: A patient is classified as high risk, enabling healthcare providers to implement early interventions to prevent sepsis.

Real-World Applications in Medicine

Treatment Outcome Prediction

1. **Surgical outcomes**: Classifying patients into different risk groups based on their surgical history, comorbidities, and other factors to predict the likelihood of post-operative complications.

   Example: A patient is classified as high risk, prompting healthcare providers to take additional precautions to prevent surgical site infections.

2. **Chemotherapy response**: Classifying patients into different response groups based on their genetic profiles, tumor characteristics, and other factors to predict the likelihood of chemotherapy response.

   Example: A patient is classified as non-responder, enabling healthcare providers to consider alternative treatment options

Classification Problem Description

Given a set of patient data, classify patients as either Diabetic (Yes) or Non-Diabetic (No) based on their characteristics (Binary classification)

Features:

Feature	Description	Units
Age	Patient's age	years
Blood Pressure	Patient's blood pressure	mmHg
Cholesterol	Patient's cholesterol level	mg/dL
BMI	Patient's body mass index	kg/m^2
Smoking	Patient's smoking status	Binary (Yes/No)
Family History	Patient's family history of diabetes	Binary (Yes/No)

Mathematical Representation

logistic regression model to predict the probability of a patient being diabetic based on their features.

Let p be the probability of a patient being diabetic, and x be the feature vector representing the patient's characteristics.

The logistic regression model can be represented as: $p = \frac{1}{1+e^{-z}}$

where z is the log-odds of the patient being diabetic, calculated as:

$$z = \beta_0 + \beta_1 * x_1 + \beta_2 * x_2 + \ldots + \beta_n * x_n$$

where $\beta_0, \beta_1, \ldots, \beta_n$ are the model coefficients, and $x_1, x_2, \ldots, x_n$ are the feature values.

The model coefficients can be estimated using a training dataset, and the resulting model can be used to predict the probability of a patient being diabetic based on their feature values.

Classification Problem Description

Given a set of patient data, classify patients as either Diabetic (Yes) or Non-Diabetic (No) based on their characteristics (Binary classification)

Features:

Patient ID	Age	Blood Pressure	Cholesterol	BMI	Smoking	Family History	Target
1	45	130	220	30	Yes	Yes	Diabetic (Yes)
2	25	90	150	20	No	No	Non-Diabetic (No)
3	60	140	280	35	Yes	Yes	Diabetic (Yes)
4	30	100	180	25	No	No	Non-Diabetic (No)
5	50	120	200	28	Yes	Yes	Diabetic (Yes)

Supervised vs. Unsupervised Learning

Supervised Learning

- machine learning where the algorithm is trained on labeled data
- the data is already classified or categorized.
- The goal is to learn a mapping between the input data and the corresponding output labels.

Unsupervised Learning

- A type of machine learning where the algorithm is trained on unlabeled data
- the data is not already classified or categorized.
- The goal is to discover hidden patterns or structure in the data.

Supervised Learning
Classification vs. Regression

Disease diagnosis

- Predicting the presence or absence of a disease based on patient symptoms, medical history, and test results (e.g., cancer diagnosis, diabetes diagnosis)

Patient segmentation

- Classifying patients into different groups based on their characteristics, such as age, sex, and medical history (e.g., high-risk patients, low-risk patients)

Predicting treatment outcomes

- Estimating the effectiveness of a treatment based on patient characteristics and medical history (e.g., predicting the response to chemotherapy)

Predicting disease progression

- Estimating the rate of disease progression based on patient characteristics and medical history (e.g., predicting the rate of tumor growth)

Unsupervised Learning
Clustering and Dimensionality reduction

Patient clustering

- Grouping patients with similar characteristics, such as medical history, symptoms, and treatment outcomes (e.g., grouping patients with similar cancer subtypes)

Gene expression clustering

- Grouping genes with similar expression patterns in patients with different diseases (e.g., grouping genes involved in cancer development)

Reducing high-dimensional data

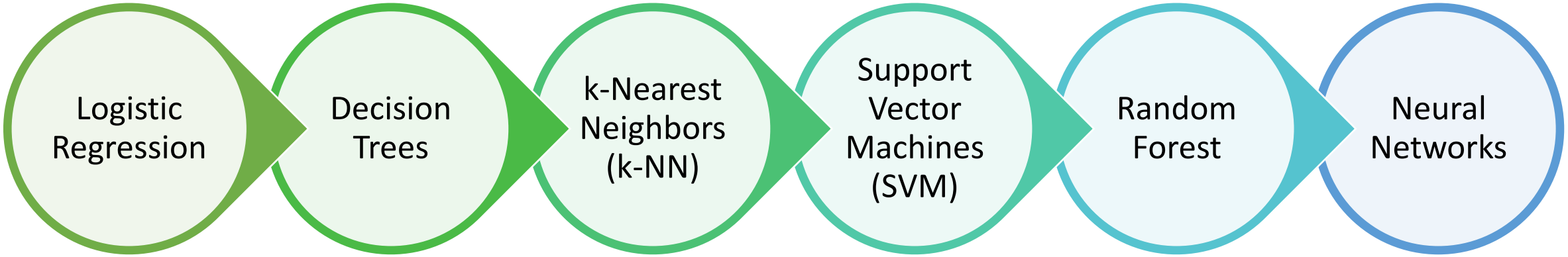
- Reducing the number of features in a dataset while preserving the most important information (e.g., reducing the number of genes in a microarray dataset)

Visualizing complex data

- Visualizing high-dimensional data in a lower-dimensional space to facilitate interpretation (e.g., visualizing the relationship between gene expression and patient outcomes)

BME BIOMEDICÍNSKÉ INŽENÝRSTVÍ

Common Classification Algorithms



Logistic Regression

is a probability-based classification model, meaning it predicts the probability of a positive outcome (e.g., disease diagnosis) given a set of input features.

Binary classification: Is a type of regression analysis used to model the probability of a binary outcome (0/1, yes/no, etc.) based on one or more predictor variables.

Assumes: Linear relationship between features and target variable

Advantages: Simple to implement, interpretable coefficients

Disadvantages: Assumes linear relationship, may not perform well with non-linear relationships

The sigmoid function is a mathematical function that maps any real-valued number to a value between 0 and 1. It is used in logistic regression to transform the linear combination of predictor variables into a probability.

Go to Notebook Example1 : Diagnosis

Decision Trees

Decision trees are a type of machine learning algorithm that work by recursively partitioning the data into smaller subsets based on the values of the input features.

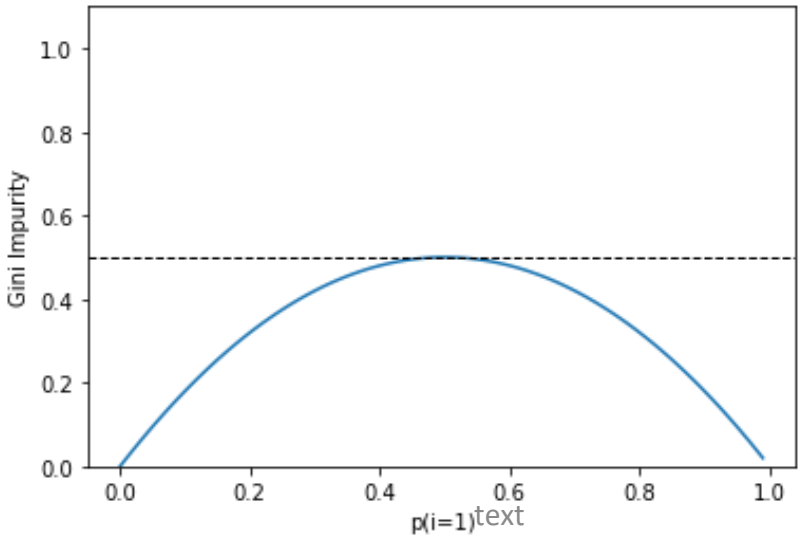
- **Classification and regression**: Builds a tree-like model to predict the target variable
- **Assumes**: No assumptions about the data distribution
- **Advantages**: Easy to interpret, handles non-linear relationships
- **Disadvantages**: Prone to overfitting, can be difficult to handle high-dimensional data

Decision Trees

Decision trees use a splitting criterion to determine which feature to split the data on at each node.
The most common splitting criteria are:

- **Gini Index**: Measures the impurity of a node, with higher values indicating more impurity. The Gini Index is calculated as:
  - $\texttt{Gini} = 1 - \sum p_i^2$, where $\texttt{p}$ is the probability of each class in the node.

When is it 0? When all instances belong to the same class (pure node)

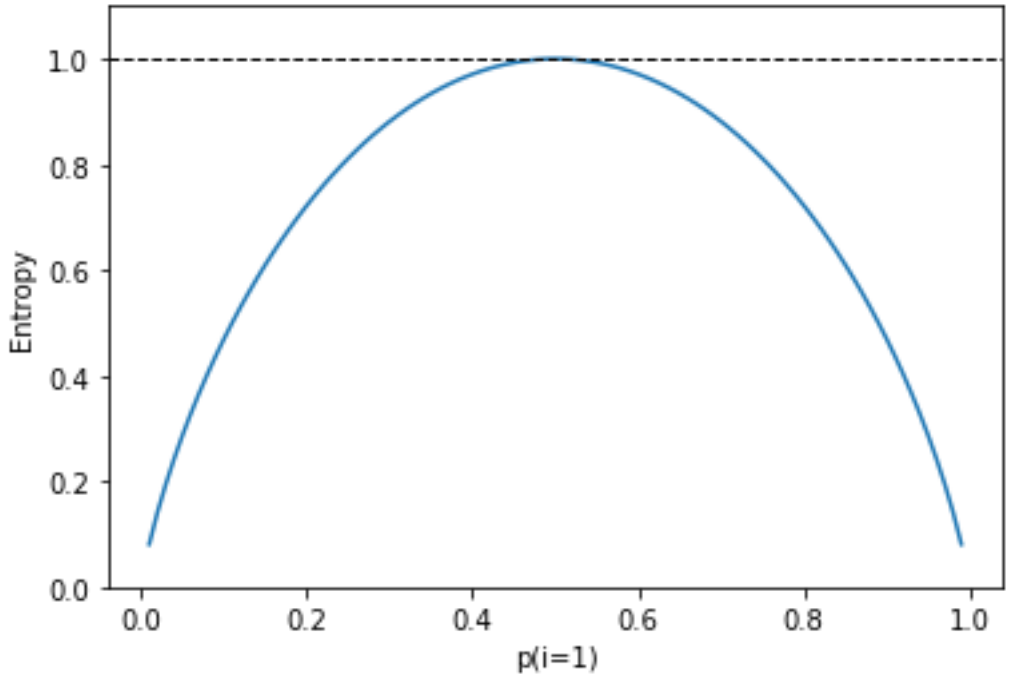


When classes are equally distributed (e.g., 50%-50%)

When is it maximum?

Decision Trees

- **Entropy**: Measures the uncertainty or randomness of a node, with higher values indicating more uncertainty. Entropy is calculated as:
  - `Entropy` $= -\sum p_i^2 \, log_2(p_i)$ where `p` is the probability of each class in the node.



Go to Notebook Example2 and Example3 :
Classifying Patients &
Prediction Accuracy by Decision Trees

- How does the logic of the decision tree work in the example?

k-Nearest Neighbors (k-NN)

KNN is an instance-based learning algorithm that makes predictions based on the similarity between a new instance and the training instances.

- **Classification and regression**: Predicts the target variable based on the k most similar observations
- **Assumes**: No assumptions about the data distribution
- **Advantages**: Simple to implement, handles non-linear relationships
- **Disadvantages**: Computationally expensive, sensitive to noise in the data

The algorithm calculates the distance to all elements of the feature space (training set).

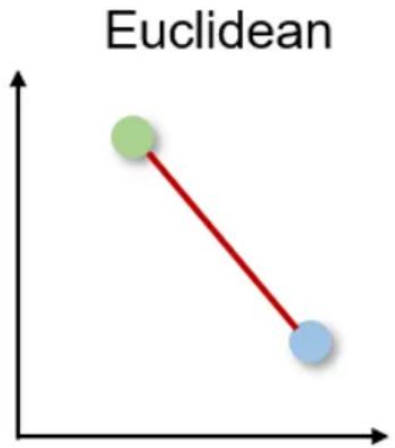
The vector of all distances is sorted in ascending order and selected to the first elements of the metric vector.

An element is classified based on **the frequency of closest occurrences**.

k-Nearest Neighbors (k-NN)

KNN uses a distance metric to measure the similarity between instances. The most common distance metrics are:

- **Euclidean Distance**: Measures the straight-line distance between two points in n-dimensional space.



$$d = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

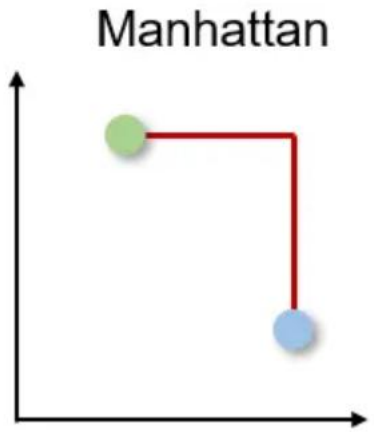
Disadvantages:
the distance measure does not work well for higher dimensional data than 2D or 3D space.

if we do not normalize and/or standardize our features, the distance might be skewed due to different units.

k-Nearest Neighbors (k-NN)

KNN uses a distance metric to measure the similarity between instances. The most common distance metrics are:

- **Manhattan Distance**: Measures the sum of the absolute differences between the corresponding coordinates of two points in n-dimensional space.



$$d = \sum_{i=1}^{n} (x_i - y_i)$$

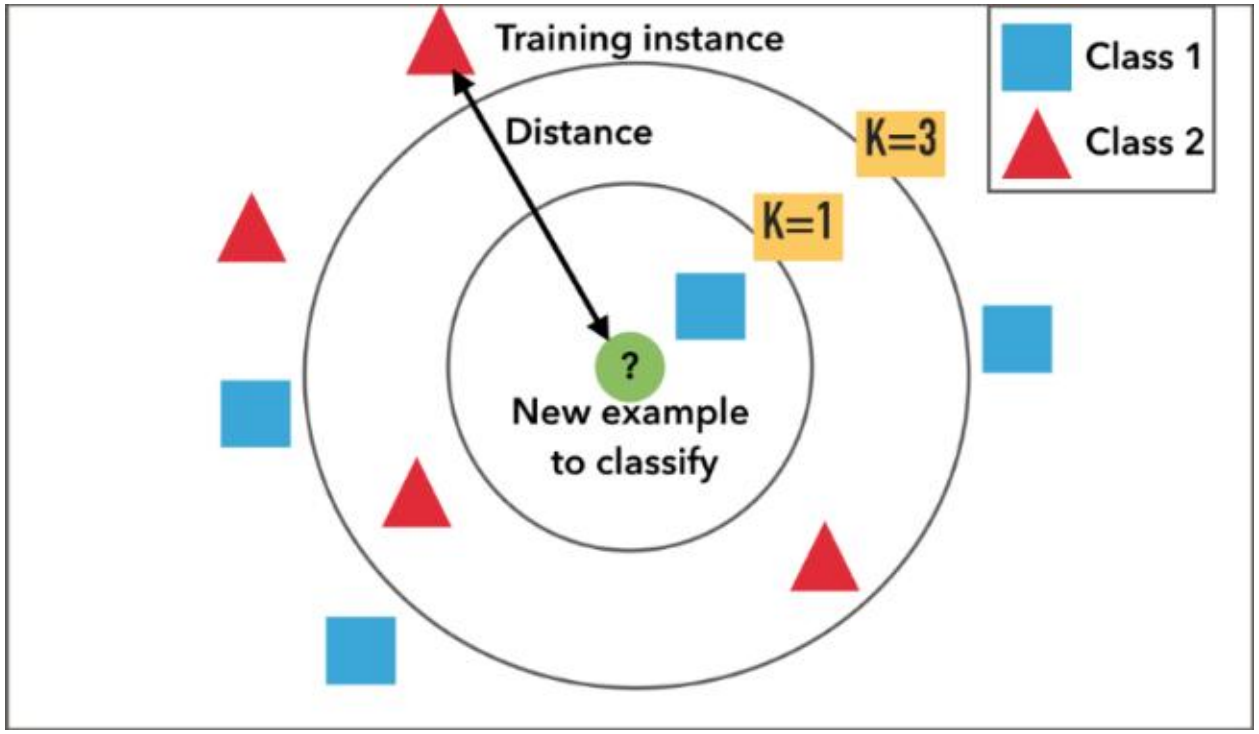
Disadvantages:

it is less intuitive than the Euclidean distance in high dimensional space

it does not show the shortest path possible. Although this might not be problematic, we should be aware of the higher distance.

K-Nearest Neighbors (K-NN) - examples

Comparison of the classifier
1 - NN and
3-NN classifier



Go to Notebook Example4:
Binary classification of patients
With cancer

BME
BIOMEDICÍNSKÉ INŽENÝRSTVÍ

How do you interpret the effect of K on the model performance?

Support Vector Machines (SVM)

- SVM is a type of supervised learning algorithm that uses a hyperplane to separate the data into different classes. The hyperplane is the decision boundary that separates the data into different regions.

- **Classification**: Finds the hyperplane that maximally separates the classes
- **Assumes**: Linear or non-linear relationship between features and target variable
- **Advantages**: Robust to noise, handles high-dimensional data
- **Disadvantages**: Computationally expensive, may not perform well with non-linear relationships

The **margin** is the distance between the hyperplane and the nearest points in the data. The goal of SVM is to find the hyperplane that maximizes the margin.

The **kernel trick** is a technique used in SVM to map the data into a higher-dimensional space where the data can be linearly separated. This is useful when the data is not linearly separable in the original space.

Go to Notebook Example5

Random Forest

Is an ensemble learning method that combines multiple decision trees to improve the accuracy and robustness of classification models.

- **Classification and regression**: Combines multiple decision trees to improve performance
- **Assumes**: No assumptions about the data distribution
- **Advantages**: Robust to overfitting, handles high-dimensional data
- **Disadvantages**: Computationally expensive, may not perform well with very large datasets

Neural Networks (Brief Overview)

Classification and regression:
- Inspired by the structure and function of the human brain

Assumes:
- No assumptions about the data distribution

Advantages:
- Can handle complex non-linear relationships, robust to noise

Disadvantages:
- Computationally expensive, may require large amounts of data to train

Architecture:
- Composed of layers of interconnected nodes (neurons)

Activation functions:
- Used to introduce non-linearity into the model

Back propagation:
- Used to train the model by minimizing the error between predictions and actual values

Types:
- Feedforward networks, convolutional neural networks (CNNs), recurrent neural networks (RNNs), etc.

Discussion Questions

- What challenges exist in medical data classification?

References

Book: DATA CLASSIFICATION: ALGORITHMS AND APPLICATIONS, Charu C. Aggarawal

Thank you for your attention

doc. Eng. Karla Miriam Reyes, Ph.D.

rey0014@vsb.cz

www.vsb.cz