*** Applied Machine Learning Fundamentals ***
k-Nearest Neighbors

Daniel Wehner, M.Sc.

SAP SE / DHBW Mannheim

Winter term 2023/2024

Find all slides on GitHub (DaWe1992/Applied_ML_Fundamentals)

Lecture Overview

Unit I	Machine Learning Introduction
Unit II	Mathematical Foundations
Unit III	Bayesian Decision Theory
Unit IV	Regression
Unit V	**Classification I**
Unit VI	Evaluation
Unit VII	Classification II
Unit VIII	Clustering
Unit IX	Dimensionality Reduction

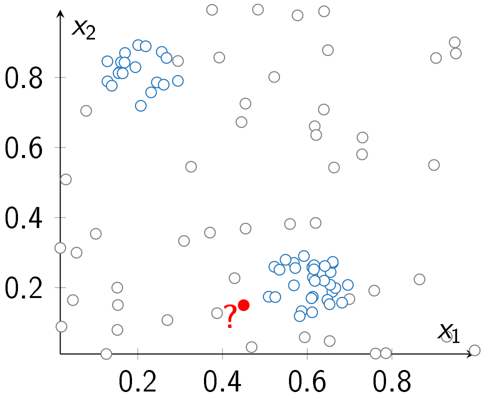
Agenda for this Unit

Section:
Introduction

Introduction

- **Basic idea:** Predict the class label based on nearby known examples
- Instance-based learning, a. k. a. **lazy learning**

We do not learn any model, the data speaks for itself!



Derivation of the Algorithm)

- Unconditional density:

$$p(x) = \frac{k}{n \cdot v}$$

- Class priors:

$$p(\mathcal{C}_j) = \frac{n_j}{n}$$

Combine them using Bayes' theorem:

$$p(\mathcal{C}_j|x) = \frac{p(x|\mathcal{C}_j) \cdot p(\mathcal{C}_j)}{p(x)} = \frac{\frac{k_j}{n_j \cdot v} \cdot \frac{n_j}{n}}{\frac{k}{n \cdot v}} = \frac{k_j}{k} \tag{1}$$

Section:

Distance Metrics

Properties of Distance Metrics
Minkowski, Manhattan, Euclidean
Cosine Similarity

Distance Metrics

- How to measure the distance between two data points u and v?
  $\Rightarrow$ **Distance metrics**
- Let d be a function $d : (u, v) \mapsto \mathbb{R}^+$ (including 0)
- This function has the following properties:

1. $d(u, v) = d(v, u)$ **(commutativity)**
2. $d(u, v) = 0 \Rightarrow u = v$
3. $d(u, k) \leqslant d(u, v) + d(v, k)$ **(triangle inequality)**

Distance Metrics (Ctd.)

Minkowski distance:

$$d_p(u, v) = \left(\sum_{j=1}^{m} |x_j^{(u)} - x_j^{(v)}|^p \right)^{1/p} \tag{2}$$

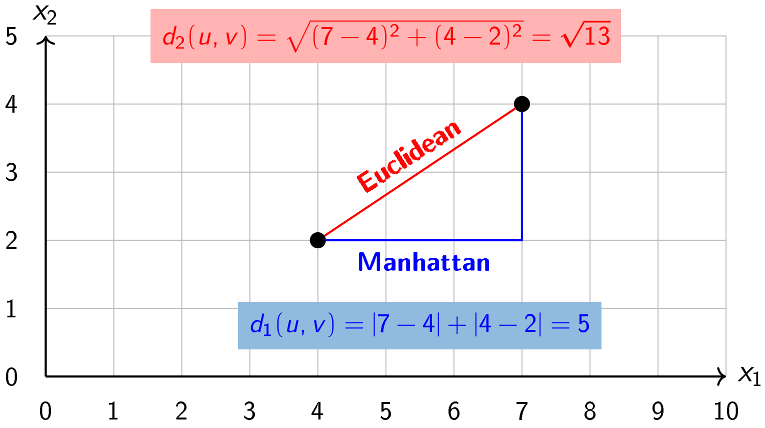
Manhattan distance: $(p = 1)$

$$d_1(u, v) = \sum_{j=1}^{m} |x_j^{(u)} - x_j^{(v)}|$$

Euclidean distance: $(p = 2)$

$$d_2(u, v) = \sqrt{\sum_{j=1}^{m} |x_j^{(u)} - x_j^{(v)}|^2}$$

Distance Metrics (Ctd.)



Cosine Similarity

- **Similarity metrics** are an alternative to distance metrics
- The **cosine similarity** of two vectors $\boldsymbol{a}$ and $\boldsymbol{b}$ is the cosine of the angle between the two vectors:

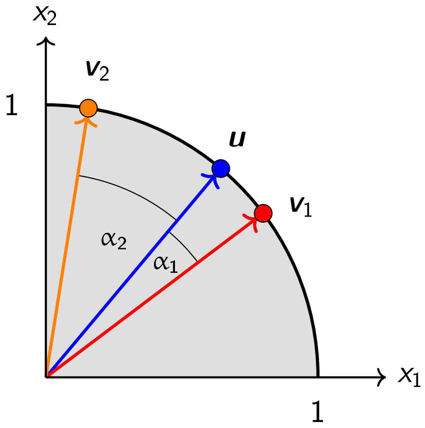
$$\cos \measuredangle(\boldsymbol{a}, \boldsymbol{b}) = \frac{\boldsymbol{a}^\intercal \boldsymbol{b}}{\|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\|} = \frac{\sum_{j=1}^{m} a_j \cdot b_j}{\sqrt{\sum_{j=1}^{m} (a_j)^2} \cdot \sqrt{\sum_{j=1}^{m} (b_j)^2}} \tag{3}$$

- The dot product is defined as (geometric interpretation):

$$\boldsymbol{a}^\intercal \boldsymbol{b} = \|\boldsymbol{a}\| \cdot \|\boldsymbol{b}\| \cdot \cos \measuredangle(\boldsymbol{a}, \boldsymbol{b}) \tag{4}$$

Cosine Similarity (Ctd.)

- $\boldsymbol{v}_1$ is closer to $\boldsymbol{u}$ than $\boldsymbol{v}_2$ because $\cos(\alpha_1) < \cos(\alpha_2)$
- Remember:
  - $\cos(0) = 1$ and
  - $\cos(90) = 0$
- **Do you see any issues?**



Section:

k-nearest Neighbors Algorithm

Predictions with k-Nearest Neighbors

k-Nearest Neighbors Algorithm:

1. Calculate the distances between the new data point and **all data points in the dataset**
2. Sort the data points by distances **in ascending order**
   (sort in descending order if similarity metrics are used)
3. Consider the first k examples and **count how often each class occurs**
4. Predict the class with **the maximum score**

❶ Calculation of Distances

v	x_1	x_2	$\mathcal{C}$	$d_2(u, v)$
1	0.66	0.24	1	0.23
2	0.25	0.79	1	0.67
3	0.16	0.81	1	0.73
4	0.57	0.21	1	0.13
5	0.21	0.72	1	0.62
6	0.66	0.27	1	0.24
7	0.27	0.11	0	0.19
8	0.39	0.13	0	0.07
9	0.39	0.86	0	0.71
10	0.44	0.67	0	0.52
11	0.31	0.33	0	0.23
12	0.03	0.51	0	0.55
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- $\boldsymbol{x}^{(u)} = (0.45, 0.15)^\intercal$
- Calculate the **Euclidean distance** between $\boldsymbol{x}^{(u)}$ and all other data points $\boldsymbol{x}^{(v)}$

Depending on the size of the dataset prediction might be expensive!

❷/❸/❹ Prediction of the Class Label

- Let k be set to 10
- Step ❷: Sort dataset by distances
  (cf. table on the right)
- Step ❸: Count class occurrences
  - **Class 0:** 3
  - **Class 1:** 7
- Step ❹: **Predict class 1!**

x_1	x_2	$\mathcal{C}$	$d_2(u, v)$
0.51	0.17	1	0.06
0.39	0.13	0	0.07
0.52	0.17	1	0.08
0.43	0.23	0	0.08
0.47	0.03	0	0.12
0.52	0.26	1	0.13
0.57	0.21	1	0.13
0.53	0.25	1	0.13
0.58	0.12	1	0.14
0.59	0.13	1	0.14
$\vdots$	$\vdots$	$\vdots$	$\vdots$

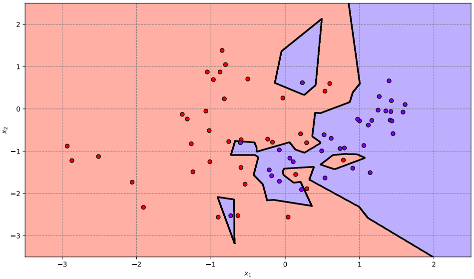
Section:

Choice of k

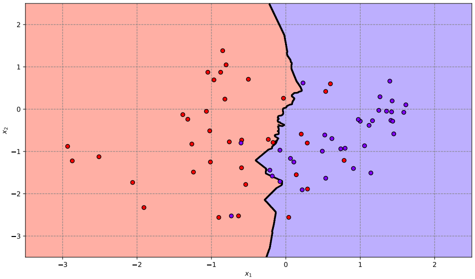
- Danger of Overfitting
- Selection Strategies

How to choose k?

The choice of k is important:



$k = 1$ (☠ **overfitting** ☠)



$k = 30$ (about right)

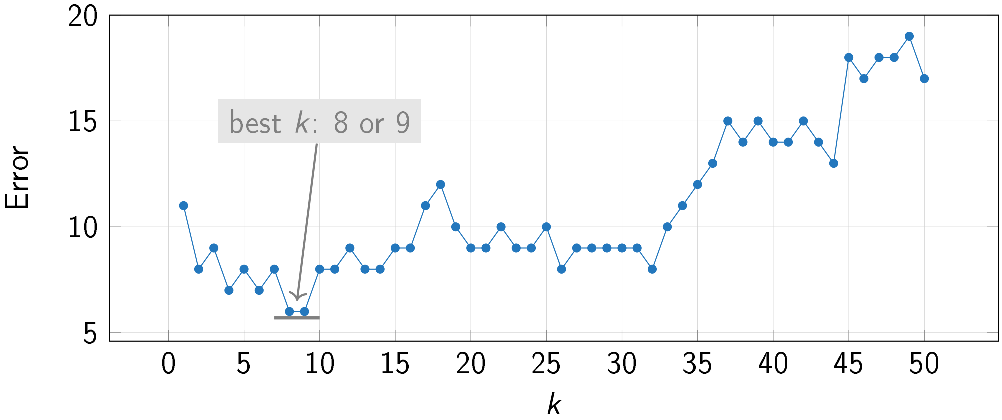
How to choose k? (Ctd.)

- First of all, it is recommended to use **odd values** for k
  (no tie-breaking necessary; at least in binary classification problems)
- Compute the value of k depending on the size of the dataset $\mathcal{D}$:

$$k = \sqrt{\frac{n}{2}} \qquad \text{or} \qquad k = \sqrt{n} \tag{5}$$

- **Usually better strategy:** Evaluate different values of k on a separate (!) development set and choose the best one (see next slide)

How to choose k? (Ctd.)



Section:

Wrap-Up

Summary
Self-Test Questions
Lecture Outlook

Summary

- The basic idea is to classify unknown instances **based on nearby examples**
- The algorithm is an example of **instance-based learning**
- **Distance metrics** allow to calculate the distance between data points:
  - Manhattan distance
  - Euclidean distance
  - Cosine similarity (as an alternative to distance metrics)
- Choose the value of k wisely:
  - Too small: **Overfitting**
  - Too large: **Underfitting**

Important

Self-Test Questions

1. Outline the k-nearest neighbors algorithm.
2. What is instance-based learning (in contrast to model-based learning)?
3. How can you compute distances? What properties do distance metrics have?
4. What is the intuition behind the triangle inequality?
5. How can you choose k?
6. Suppose you have a dataset comprising $n = 50$ examples.
   If you set $k = n$, what class does the algorithm predict?
7. What are advantages and disadvantages of the algorithm?

What's next...?

Unit I	Machine Learning Introduction
Unit II	Mathematical Foundations
Unit III	Bayesian Decision Theory
Unit IV	Regression
Unit V	**Classification I**
Unit VI	Evaluation
Unit VII	Classification II
Unit VIII	Clustering
Unit IX	Dimensionality Reduction

Thank you very much for the attention!

Topic: *** Applied Machine Learning Fundamentals *** k-Nearest Neighbors
Term: Winter term 2023/2024

Contact:

Daniel Wehner, M.Sc.
SAP SE / DHBW Mannheim
daniel.wehner@sap.com

Do you have any questions?