Implementation and evaluation of K-nearest neighbors (KNN) algorithm for handwritten digit recognition.

Project Proposal – DataScience SoSe 2021
Project 5 Group 2

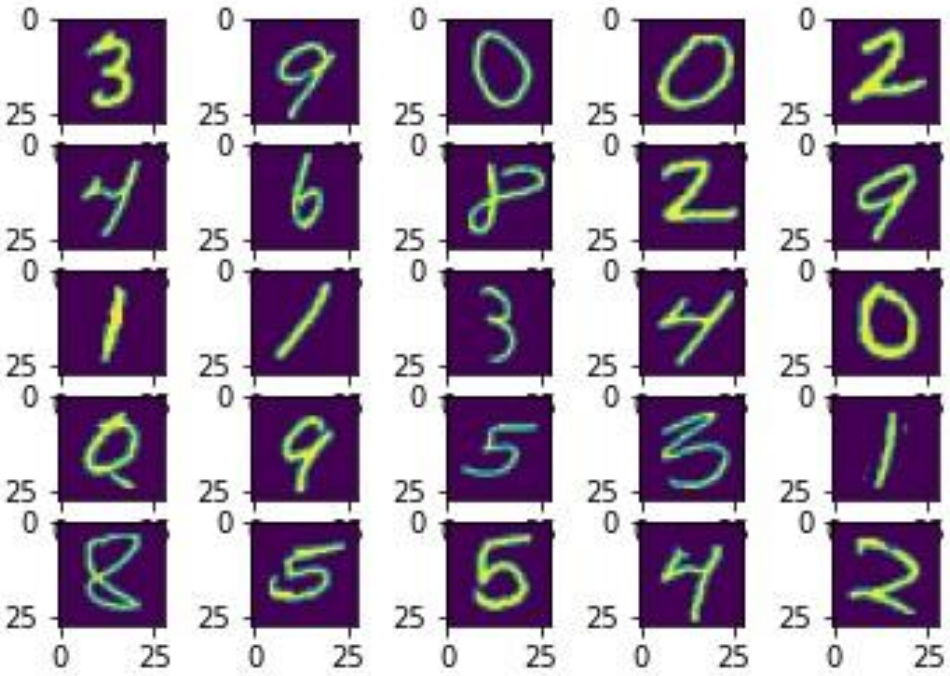
Maximilian Hingerl, Emma Kray,
Nina Gutzeit, Johannes Müller

Background image: https://unsplash.com/photos/NZ-9RqsE06I

Project Goal

To write a functional code that can recognize handwritten digits using PCA and K-nearest neighbors

Dataset



Training set - 60.000 images

Test set - 10.000 images

Size-normalized (28*28 pixels)

Centered

As .csv

- One row → one image
- First column represents label

Logical algorithm flow

Input: Image

Data normalization

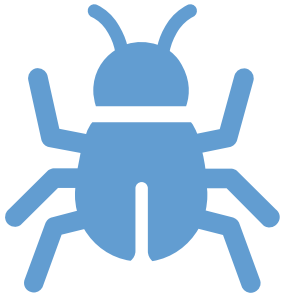
Principal component analysis

K-nearest-neighbors of training set

→Classification of test set images

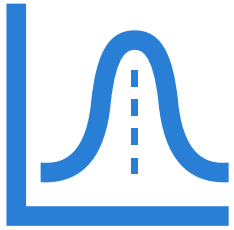
Output: class membership based on KNN

1. Milestone: implementing data normalization



Check for and resolve errors

- Duplicates?
- NA values? All values between 0-255?
- Correct labels in training dataset?
- Correct image orientation
- Identify outliers



Standardizing, optimizing for KNN

- Standardization: Z-Transformation
- or: Normalization/ Re-scaling: [0, 1]

2. Milestone: implementing PCA

Benefits of PCA

- **Reduces training time** by decreasing the dimensionality
- **Reduces noise** by reducing data set to only relevant variable

Planned analysis steps:

- Write own PCA code
- Visualize the PCA

3. Milestone: implementing a classification algorithm

Delivers: class/label of the tested data

- Should return the class (digit between 0 and 9) of the tested data

Planned analysis steps:

- Write KNN-function on our own

K-nearest neighbors algorithm

Calculate the distance between the tested data and the training data set

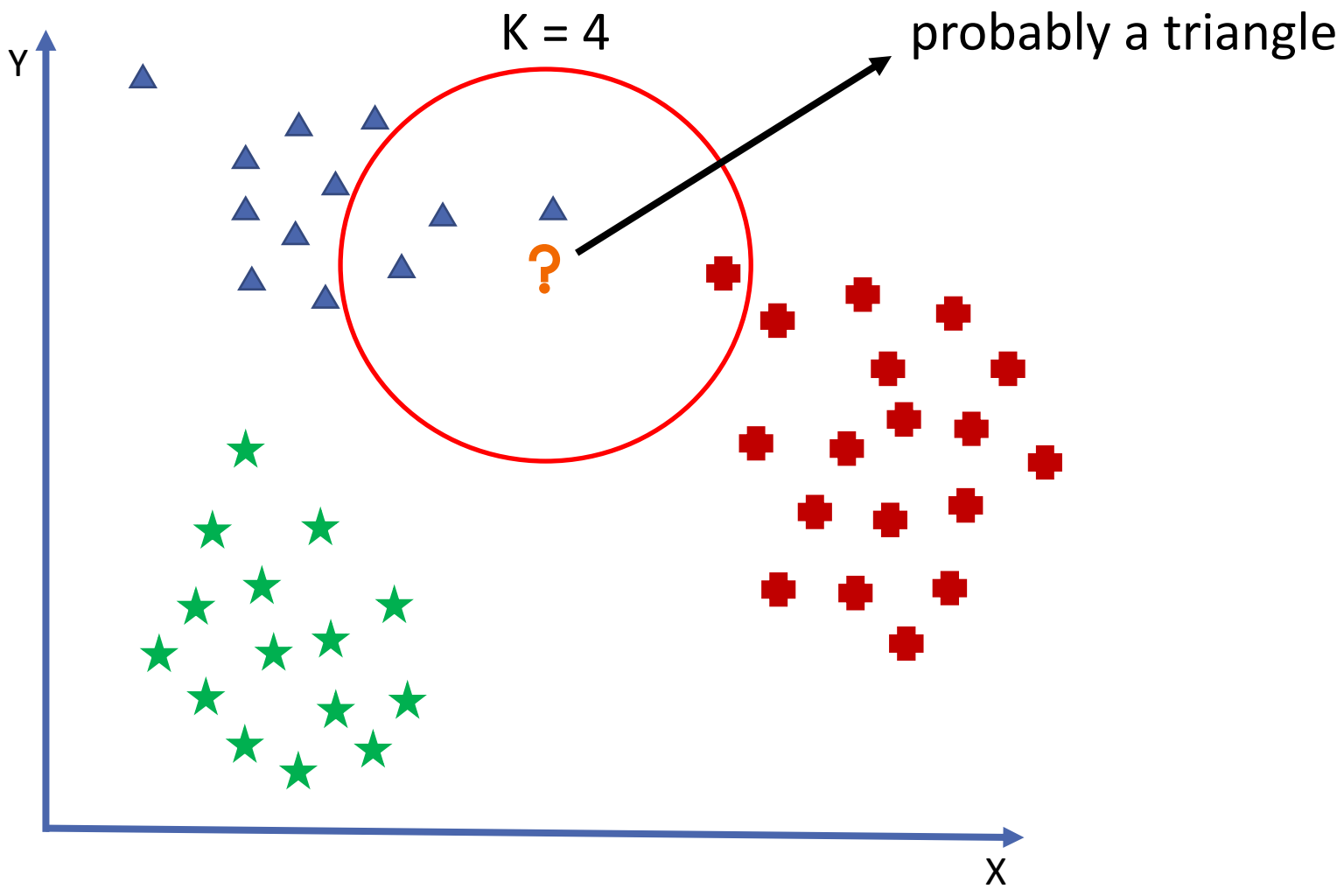
$$d_{Euclidean}(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$$

$$d_{Manhattan}(x, y) = \sum_{i=1}^{n} |x_i - y_i|$$

Find K-nearest neighbors of the tested data

Requirement: data needs to be standardized

Image source: Dr. Carl Hermann, DataAnalysis3FS_WS2021_Woche2.pdf


K = 4
probably a triangle
Y
X
?

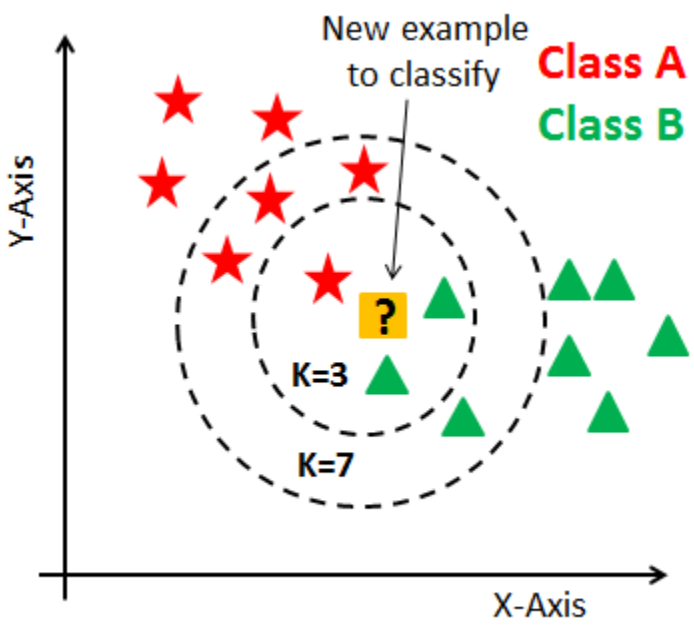
4. Milestone: testing the algorithm

Different to other machine learning algorithms

- No model to train

Instead:

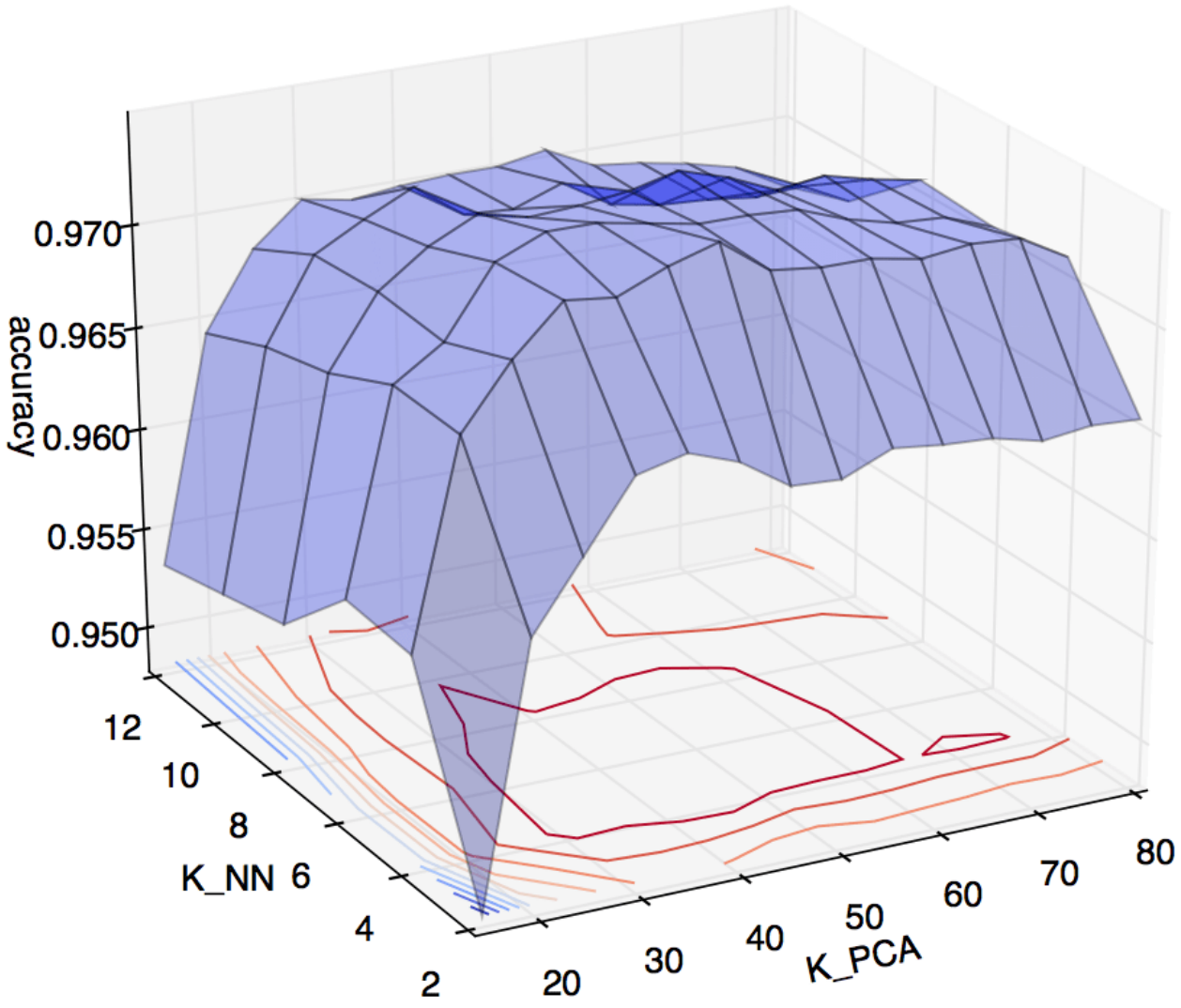
- Plotting accuracy vs. K-value vs. PCA
- Finding maximal accuracy



https://www.datacamp.com/community/tutorials/k-nearest-neighbor-classification-scikit-learn

4. Milestone: testing the algorithm

- Plotting accuracy vs. K-value vs. PCA
- Finding maximal accuracy



https://steven.codes/blog/ml/how-to-get-97-percent-on-MNIST-with-KNN/

Timeline

12.05.21 - 15.06.21

Emma:
Data Normalization

Maximilian & Nina:
PCA

Johannes:
KNN

16.06.21 - 20.06.21

Connecting our components

21.06.21 - 30.06.21

Optimization & performance evaluation

01.07.21 – 15.07.21

Optimization & performance evaluation

Possible application: Digitization phone numbers

Idea:

- digitizing handwritten phone number

Challenges:

- Number → single digits
- Centering digits
- Adjust format

→ Image preprocessing before analysis

Thank you for your attention!

Sources:

- Gerbrands, J.J. "On the relationships between SVD, KLT and PCA." Pattern Recognition (1981), vol. 14, issues 1-6, pp 375-381
- Netzer, Y. et al. "Reading Digits in Natural Imageswith Unsupervised Feature Learning." Proceedings of the Workshop on Neural Information Processing Systems (2011)
- Gareth, J. et al. "An introduction to statistical learning." Springer New York (2013), Chapter 4.4