IMAG/e

TU/e Technische Universiteit Eindhoven University of Technology

Image Analysis Project 8QA01

Part 3 – Training classifiers

Dr. Veronika Cheplygina

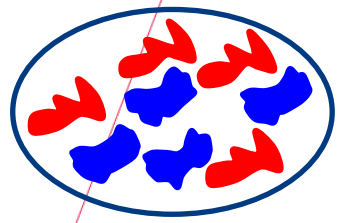
Goal: Transform features to a score

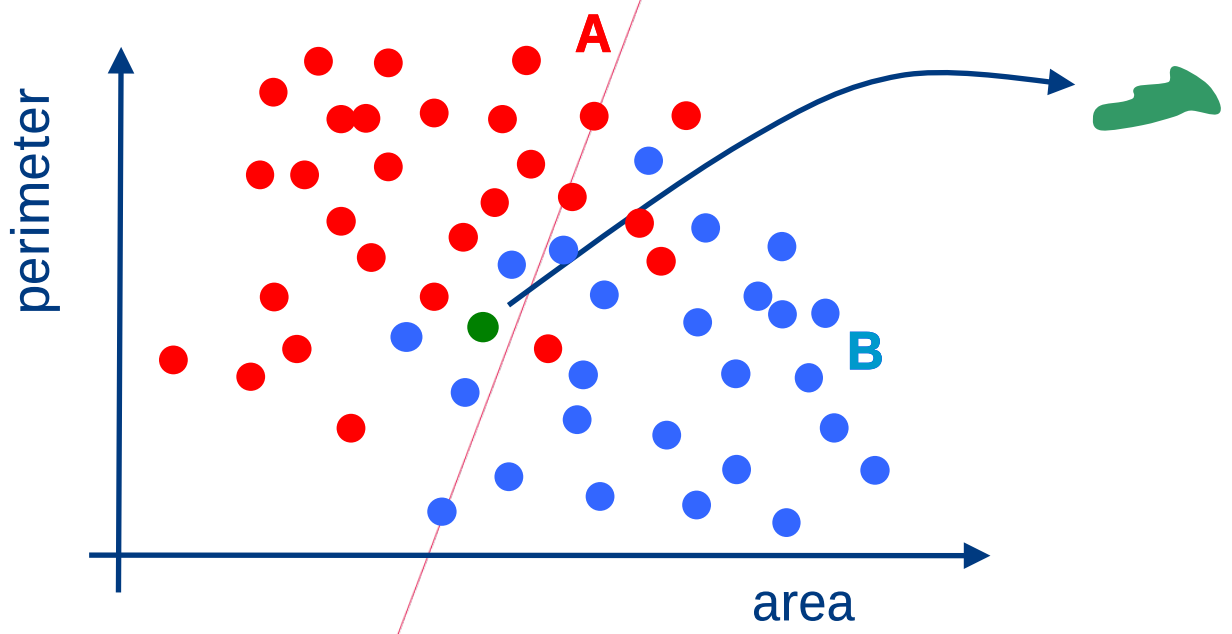
- Binary label (yes/no)
- Continuous score (probability of cancer)

Q1: Age	…	Q7: Color	New feature
20		1	1.7
25		1	2.5
40		2	1.3
70		3	0.1

P(cancer)
0
1
0
1

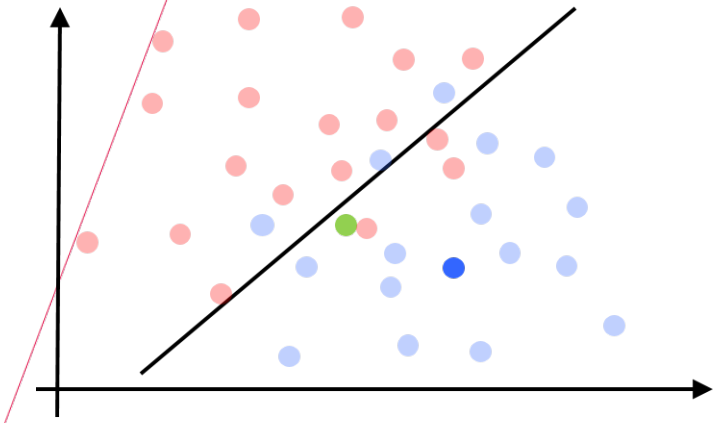
- Use the features of existing data to train a **classifier**





Example classifier – Nearest mean classifier

- Represent each class by its average point
- Measure distances to “average red” and “average blue”



Example with 5 training points using 3 features

Class 1:

[3 4 4]

[5 5 2]

[1 6 3]

Class 2:

[2 2 4]

[2 4 6]

Test point:

[5 3 4]

Pause video to practice example!

Nearest mean classifier - Steps

- Model each class **by its mean**
  - [3 5 3] for class 1, [2 3 5] for class 2
- For the test point, calculate its distances to the class means
  - $\sqrt{9}$ to class 1, $\sqrt{10}$ to class 2
- Find class with the smallest distance
  - Answer is **class 1**

Other classifiers:

- Nearest neighbor
- Decision tree
- Logistic regression
- …

Neural networks are often used in practice, but for this project we will only look at using a few features, with a classifier that is easy to interpret

Normalize features before training your classifier

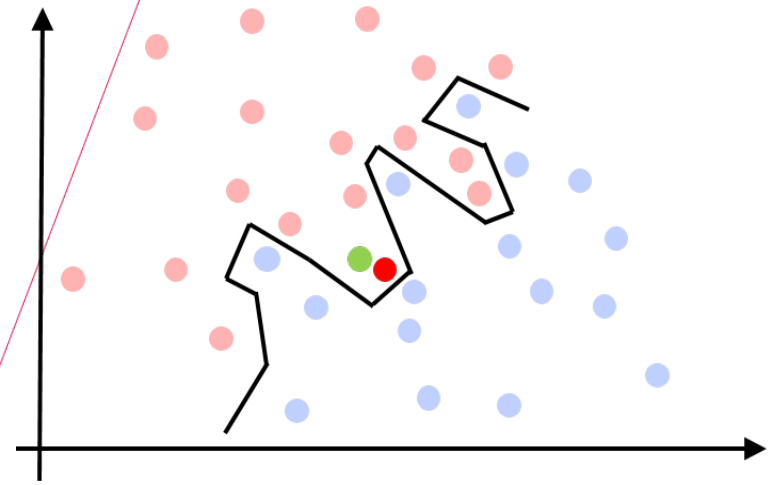
- Different features might have different scales, for example
  - Color between 1 and 5
  - Area between 1000 and 5000

- Features with larger scales will have more influence on the classifier

- Normalize features (for example to [0,1] interval]) to ensure each feature is equally important

What is a good classifier?

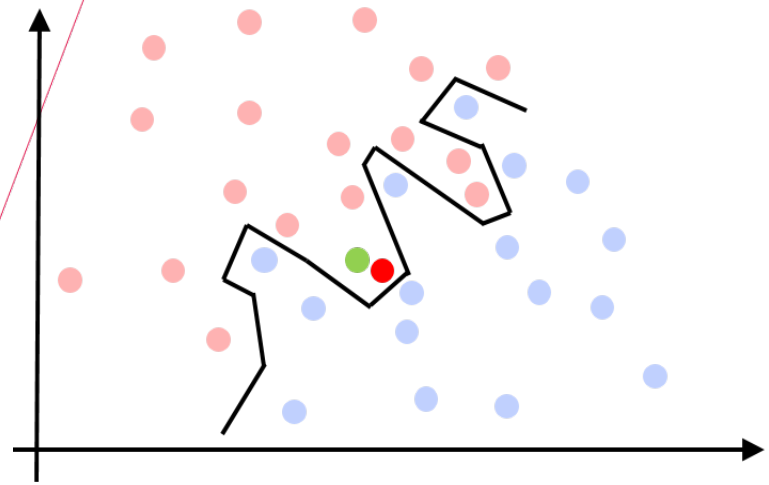
- We want to have good separation of the categories (low class overlap)
- It is always possible to design a (complex) classifier that perfectly separates your training data
- But, that classifier might **overfit** = not **generalize** to future data



What is a good classifier?

- Use different subsets of data for
  - Training
  - Validation (practice test set)

- Goal: similar performances training & validation sets
  - Bad: 100% training, 50% validation
  - Better: 75% training, 75% validation



What is a good classifier?

- When training & validation performances are similar, we expect similar performances on “true” test data

- Use a third subset, a.k.a. test set
  - External / future data
  - Only for reporting, do not change your classifier after