IMAG/e

TU/e Technische Universiteit Eindhoven University of Technology

Image Analysis Project 8QA01

Part 4 – Evaluating your method

Dr. Veronika Cheplygina

Goal: Evaluate if your method is suitable for assessing skin lesions

- You developed some features and classifiers, now what?
- Evaluate features and/or features + classifier

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

True label
0
1
0
1

Predicted label
0
1
0
0

Evaluating features

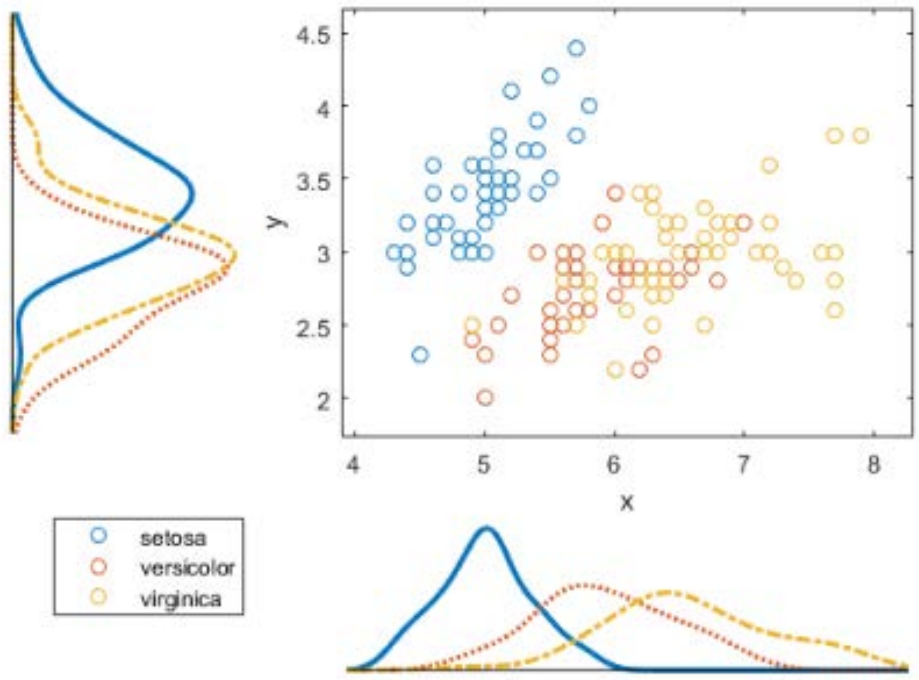
- Calculate features for artificial images (e.g. a circle) – are the values logical?
- Compare features to measuring properties by hand
  - Correlation, inter-observer agreement, …

Asymmetry (algorithm)
0
1
2
100

Asymmetry (observer 1)
0
1
4
3

Evaluating features

- Look at plots of your features per class – are there clusters?



https://stackoverflow.com/questions/57267150/scatterplot-with-marginal-kde-plots-and-multiple-categories-in-matplotlib

Evaluating classifiers

- Compare predicted label to true label
- Accuracy = correctly classified images / total images * 100 (here 75%)

True label	Predicted label
0	0
1	1
0	0
1	0

Evaluating classifiers

- If classifier predicts a probability, convert to a label by **thresholding**
- The threshold is a hyperparameter
  - Use 0.5 or select a different one based on training/validation sets

True label
0
1
0
1

Predicted probability
0.2
0.9
0.3
0.4

> 0.5 ? →

Predicted label
0
1
0
0

Evaluating classifiers

- Accuracy is not always a suitable metric
- If we have only 1% cancer in the data, predicting all images as healthy is 99% accurate!
- Consider other metrics, such as precision/recall, AUC (area under the receiver-operating curve) and others

Evaluating classifiers

- Our method has 75% performance on the test set
- Now what?

Table 1: Performance of our method
75%

Evaluating classifiers - wrongly classified examples

- Which images are incorrectly classified?
  - Confusion matrix

Predicted → True ↓	NC	C
Non-Cancer	70	30
Cancer	5	95

Evaluating your method - wrongly classified examples

- Inspect the images visually
  - If your classifier outputs probabilities, look at “most incorrect” ones
- Are there any patterns?
  - Images too light/dark
  - Etc

True label	P(class 1)
0	**0.9**
0	0.6
1	**0.1**
1	0.4

Evaluating classifiers - different versions of your method

- What is the most important part of your method?
- Repeat your experiment, but remove 1 feature or parameter at a time
- Also referred to as “ablation experiments”

Evaluating classifiers – multiple runs

- Classifier A has 75% and classifier B has 76% – is B better?
- Repeat the experiment 5 times, shuffling the data, to get mean + standard deviation
  - 75 ± 0.1% and 76 ± 0.1% vs
  - 75 ± 5% and 76 ± 5%
- You can shuffle the data, or use cross-validation

Evaluating algorithms

- We discuss evaluation from an algorithm's point of view
- In practice, other factors should be considered
  - Do people trust this technology?
  - What are the rewards and risks to society?
  - Etc

Summary

- Goal of project
- Extracting features
- Training a classifier
- Evaluating your method

- Next: requirements for assignment