Pattern Recognition

Rajmund Kożuszek, MSc., room 205
Office hours: Tuesday 12.30-14.00
r.kozuszek@ii.pw.edu.pl

All information concerning lectures and laboratory will be published on official faculty course page:

https://studia3.elka.pw.edu.pl/

(you'll need an account on a faculty server to access the page).

Lectures:	Tuesday 8.15-10.00 r. 121
Laboratory:	Thursday (ODD) 16.15-20.00 r. 139 (group A/101)
	Thursday (EVEN) 16.15-20.00 r. 139 (group B/102)
	Monday (ODD) 16.15-20.00 r. 139 (group C/103)

Grading

Overall 100 points are divided into:

- Laboratory: 2 laboratory assignments graded in the 0-6 scale; 2 mini-projects graded in the 0-12 scale
  => maximum of 36 points
  To pass successfully laboratory you have to collect at least 18 points.
- Theoretical part – 64 points (test I: max. 31 and test II max. 33 points. Tests are written with notes (*own notes*). Important dates:
  November 19th: test I (lectures 1-6)
  January 21st: test II (lectures 8-13)
  To pass theoretical part you have to collect at least 32 points.

Laboratory

Subject matter:

1. Nearest neighbour classification (startup exercise – not graded).
2. Bayes classification with normal distribution and probability density estimation with Parzen window. (6 points)
3. Linear classification. (12 points)
4. Artificial neural networks – back-propagation algorithm. (12 points)
5. Recognition quality enhancement. (6 points)

Tools:

Octave (www.octave.org)

Laboratory

Rules:

1. All exercises have defined due date. Solutions turned in on this date can earn maximum grade.
2. Exercises can be carried out outside laboratory, but the solution must be original student's work.
3. To get the grade student should present: source code of the solution, report (in electronic form) and discuss solution and results with person leading exercise.
4. For each **started** week of delay in returning exercise's solution maximum grade is lowered by 1. Maximum delay is 3 weeks.

	1	2	3		4		5
Group A (USOS 101)	17.10	31.10	14.11	28.11	12.12	16.01	30.01
Group B (USOS 102)	10.10	24.10	07.11	21.11	05.12	19.12	23.01
Group C (USOS 103)	14.10	28.10	13.11	25.11	09.12	13.01	27.01

Bibliography

Duda R.O., Hart P.E., Stork D.G., *Pattern Classification*, Wiley-Interscience, 2000

Bishop C.M., *Pattern Recognition and Machine Learning*, Springer 2006

Goodfellow I., Bengio Y., Courville A., *Deep Learning*, MIT Press, 2016 (http://www.deeplearningbook.org/)

Jain A. K., *Fundamentals of Digital Image Processing*, Prentice-Hall International Editions, Engelwood Hills, 1989

Press W. H., *Numerical Recipes in C*, Cambridge University Press, Cambridge 1992 (http://www.nr.com)

Learning outcomes

Knowledge:

1. Student knows basic pattern classification methods
2. Student knows initial data analysis and clustering methods
3. Student knows basic methods of designing classifiers' committees

Skills:

1. Student can analyze training set, design simple classifier and assess its quality
2. Basing on the analysis of the training set student is able to select proper classification method and determine its parameters
3. Student is able to assess a pattern classification solution and propose its enhancements

Pattern recognition

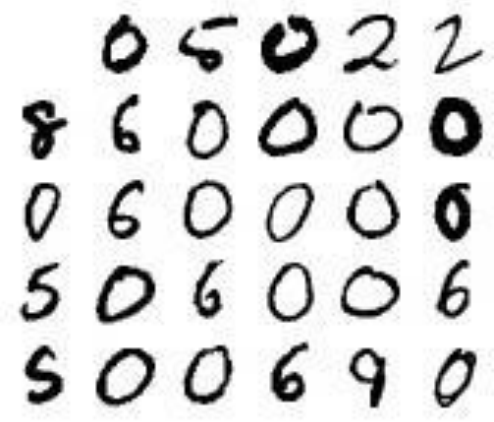
= *pol.* Rozpoznawanie obrazów

Pattern

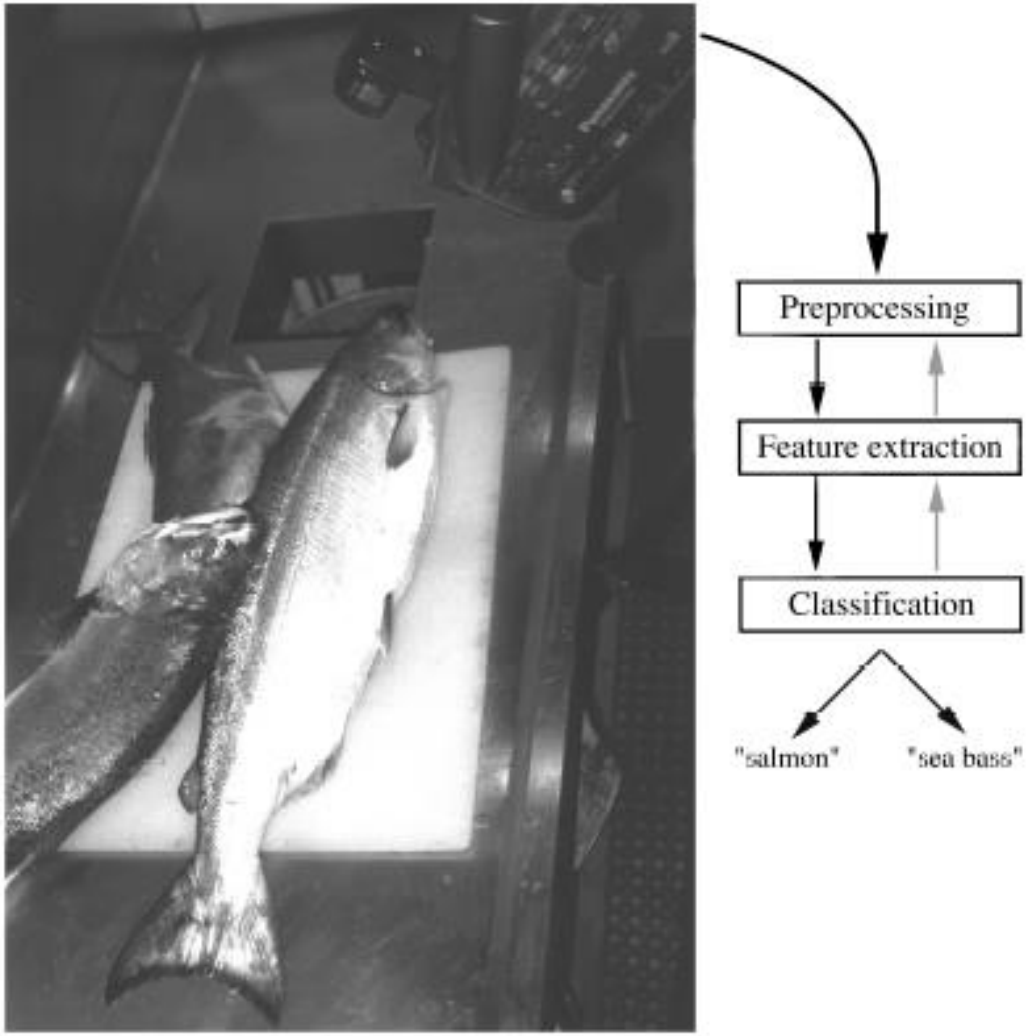
1. a repeated decorative design, e.g. on fabric or china
2. a natural or chance arrangement or sequence
   a frost pattern
   the pattern of events
3. a design, model, or set of instructions for making things
   a dress pattern
4. a model for making a mould into which molten metal is poured to form a casting
5. a form or model proposed for imitation; an example
6. a specimen or sample, e.g. of wallpaper

Patterns

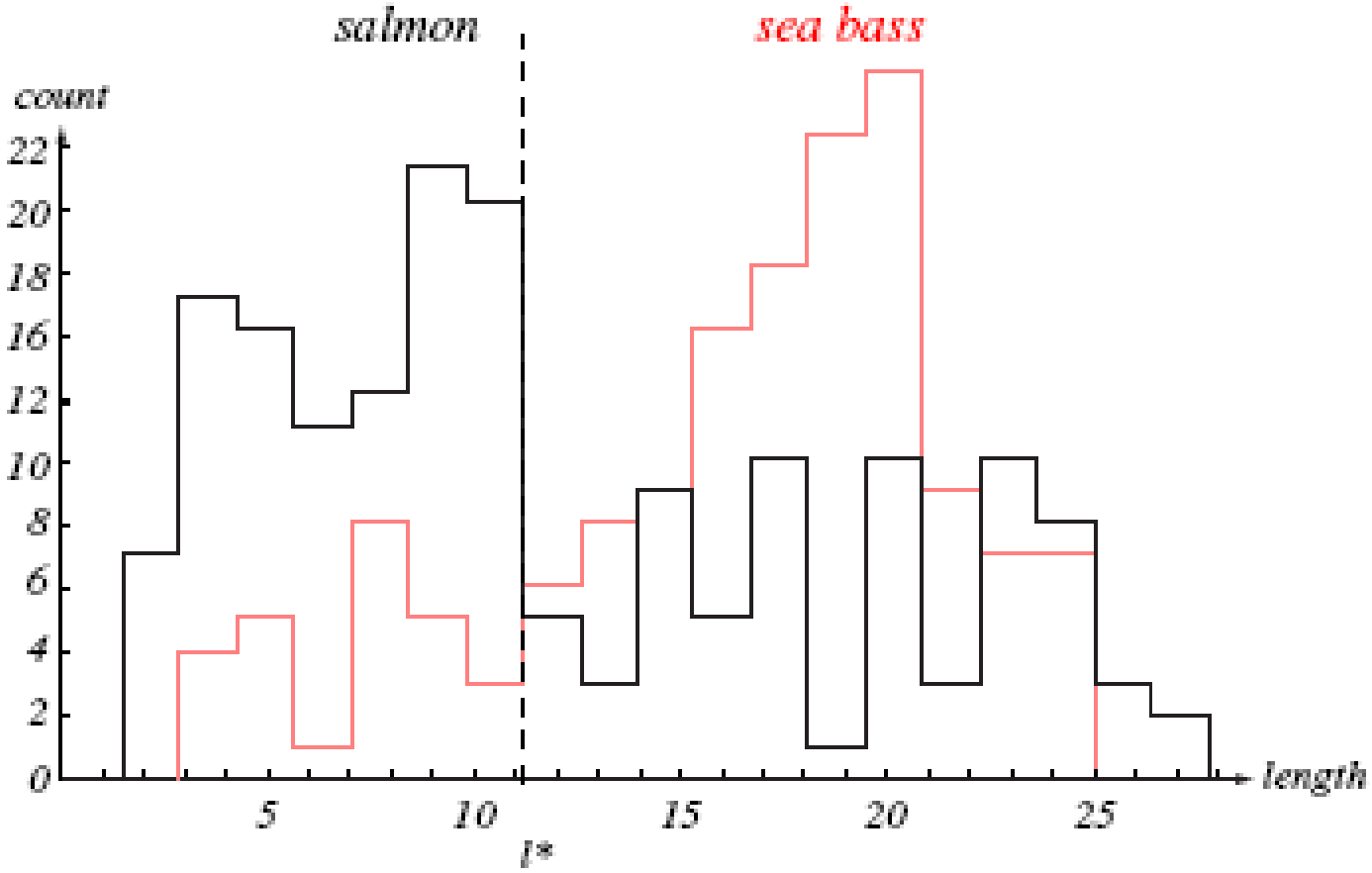
```
0.11586,0.86382,4.2998,0.88029
0.031722,-0.97325,-0.18586,1.154
0.049696,0.39263,-6.17,0.47397
-0.017119,0.096331,-4.3949,0.68433
0.22962,-0.085538,-1.7988,-1.2181
-0.11064,-1.0803,2.1374,0.21424
-0.13247,-1.4722,4.8275,0.84576
0.1299,-0.12251,-0.66075,-1.1466
0.096487,-0.40473,0.91897,1.052
0.036473,-0.33991,3.959,1.2906
```



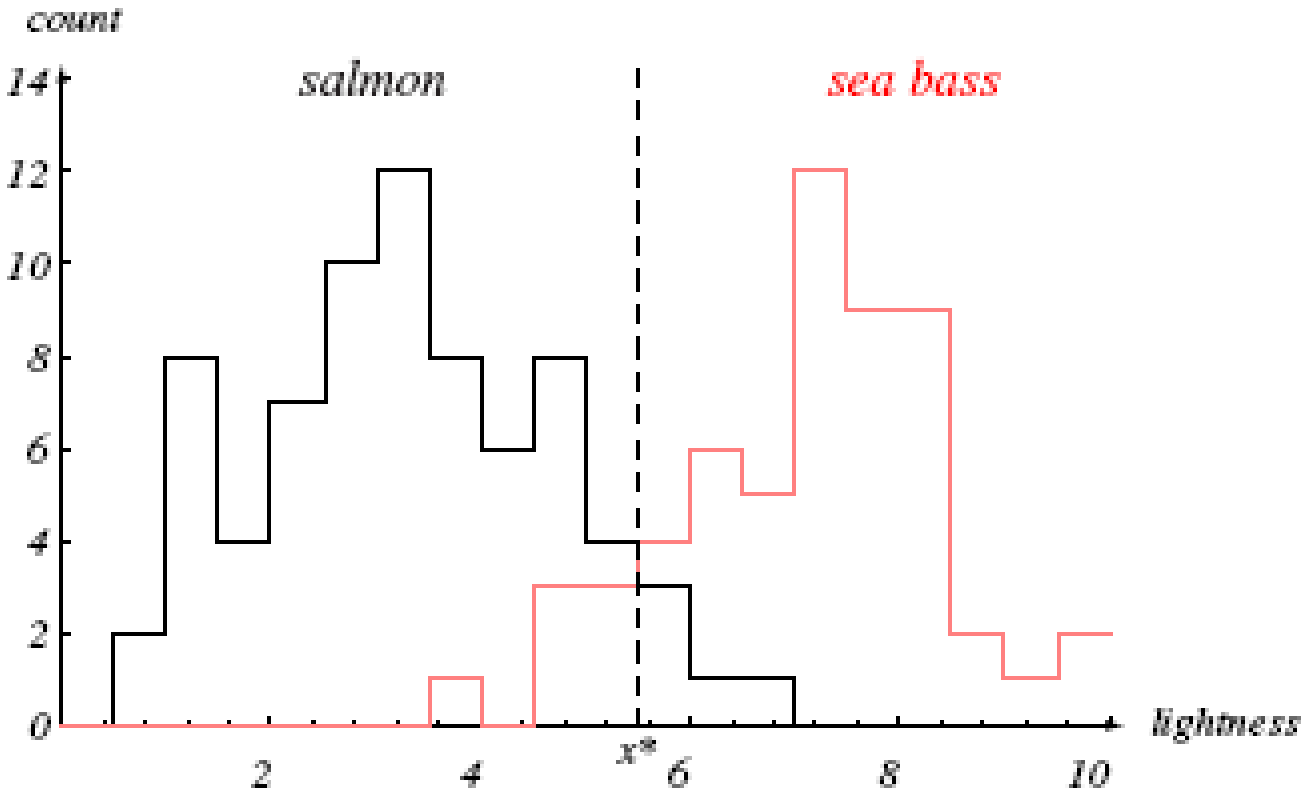
Example – fish sorting



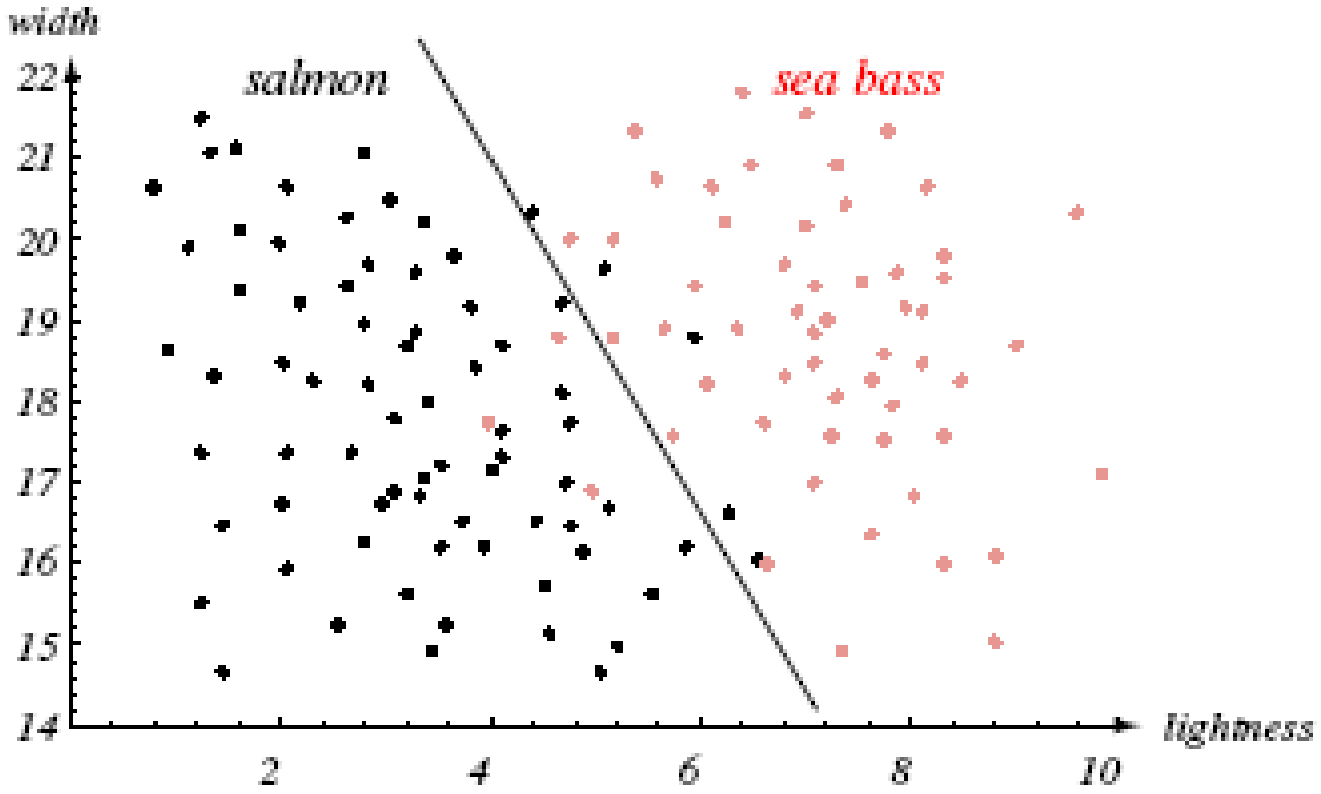
Feature selection



Feature selection

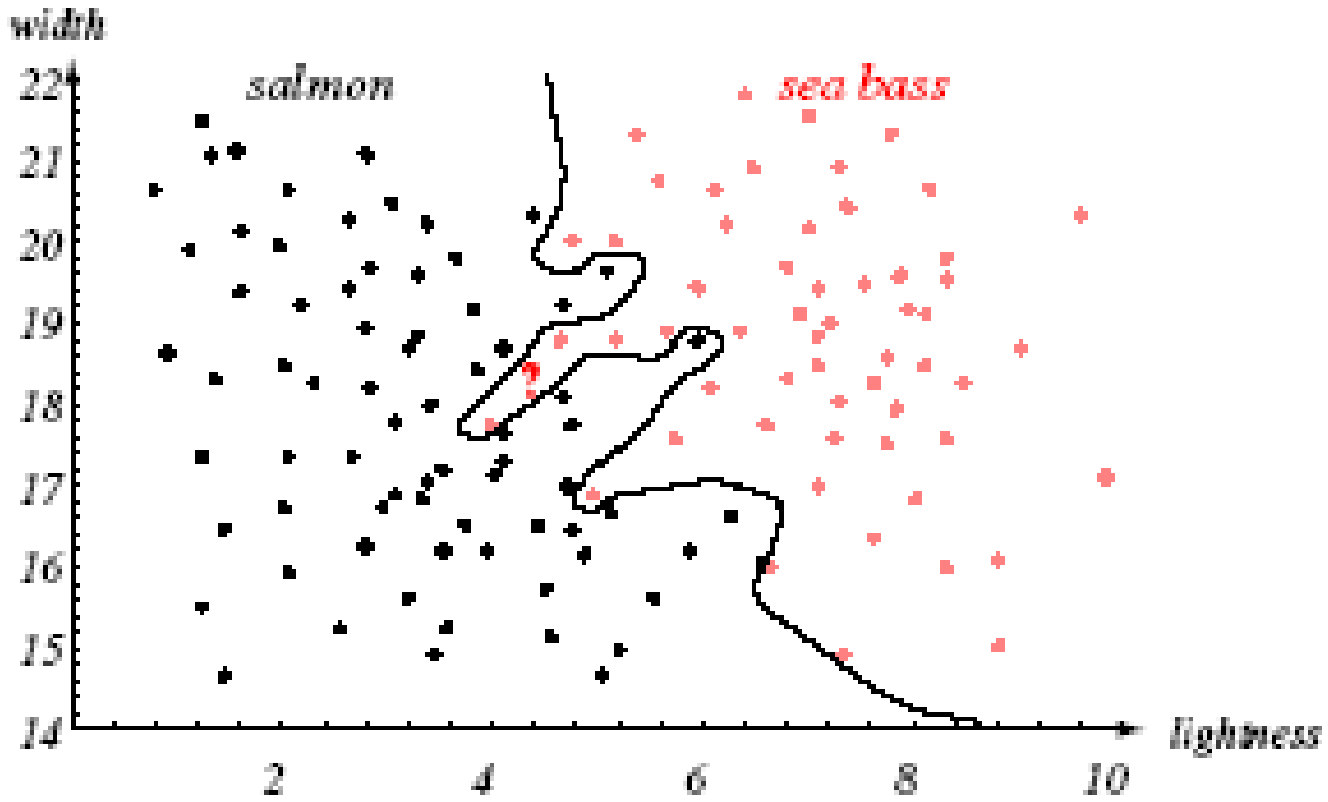


Feature selection



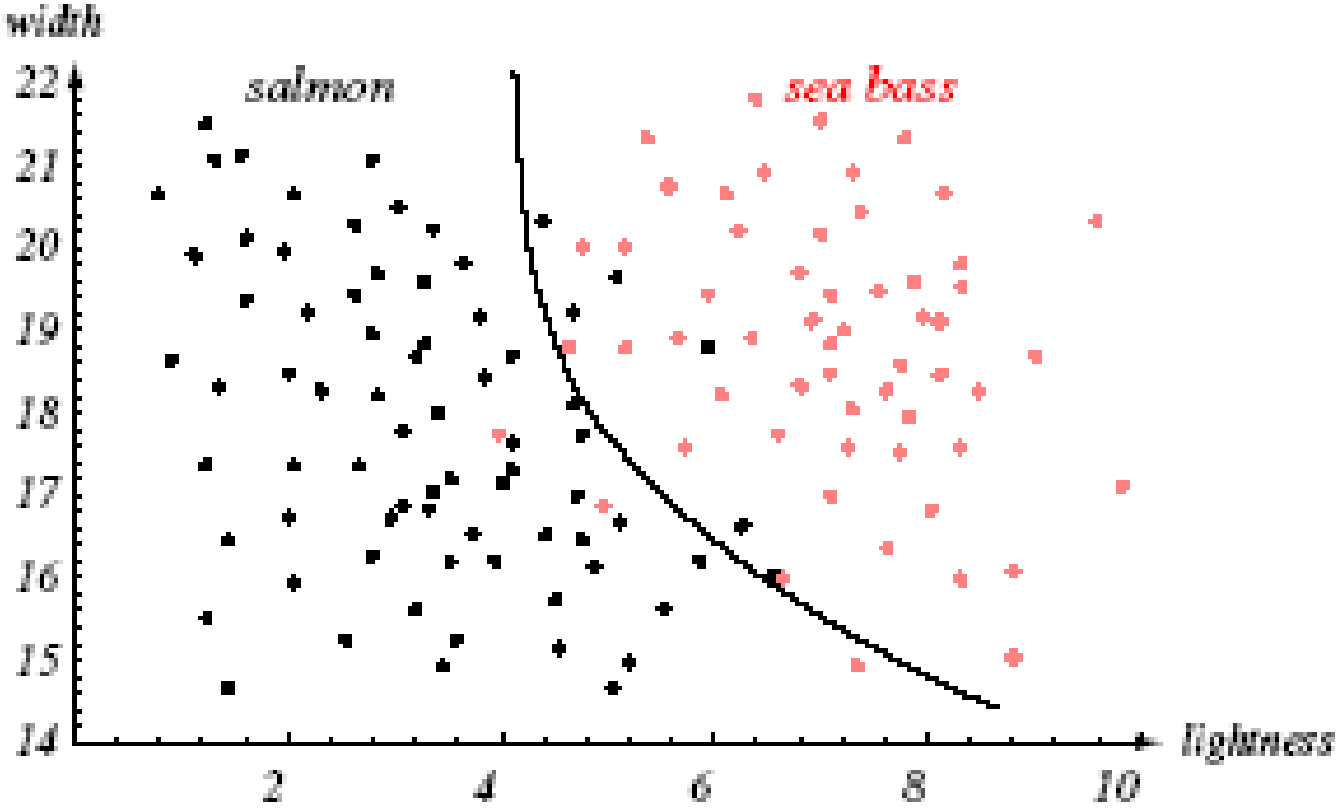
We can use the simplest decision boundary i.e. a line in two dimensional case.

Classifier training



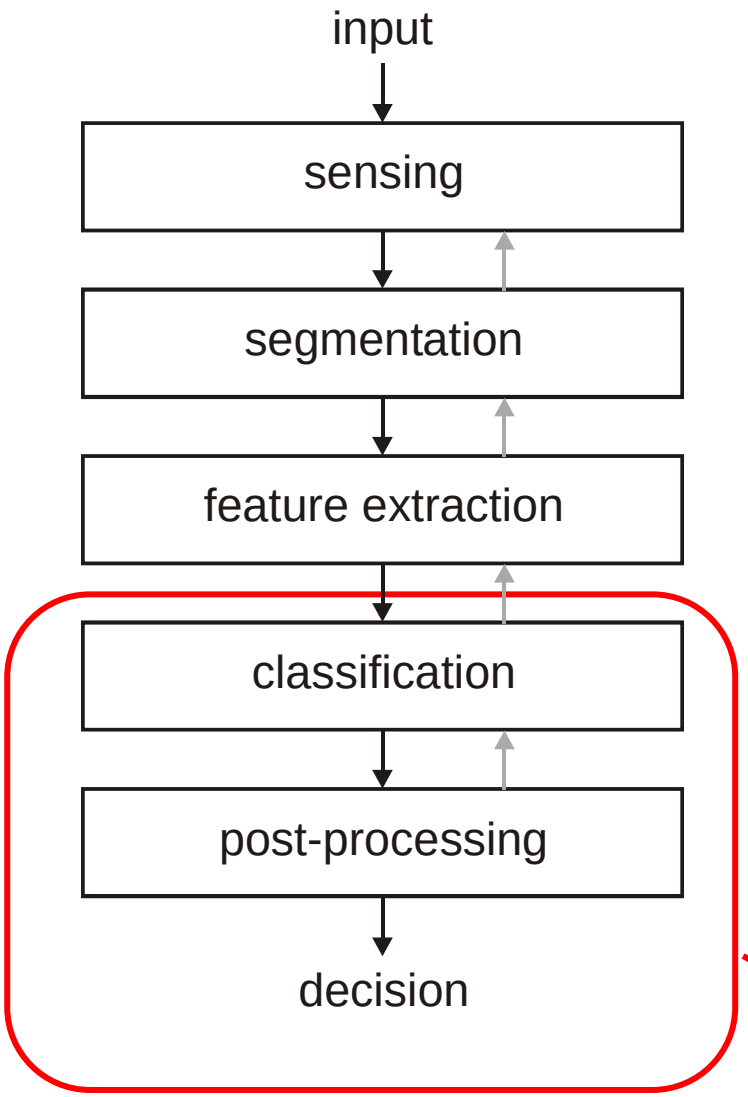
Decision boundary visible on the diagram is too complicated – this classifier will perform poorly on novel data. This phenomenon is called *overfitting of classifier*.

Classifier training



Although this classifier is not perfect on the training set, we can hope that its performance on real data will be better than that of the previous classifier.

Pattern recognition system



Very important step, often decides about recognition quality.

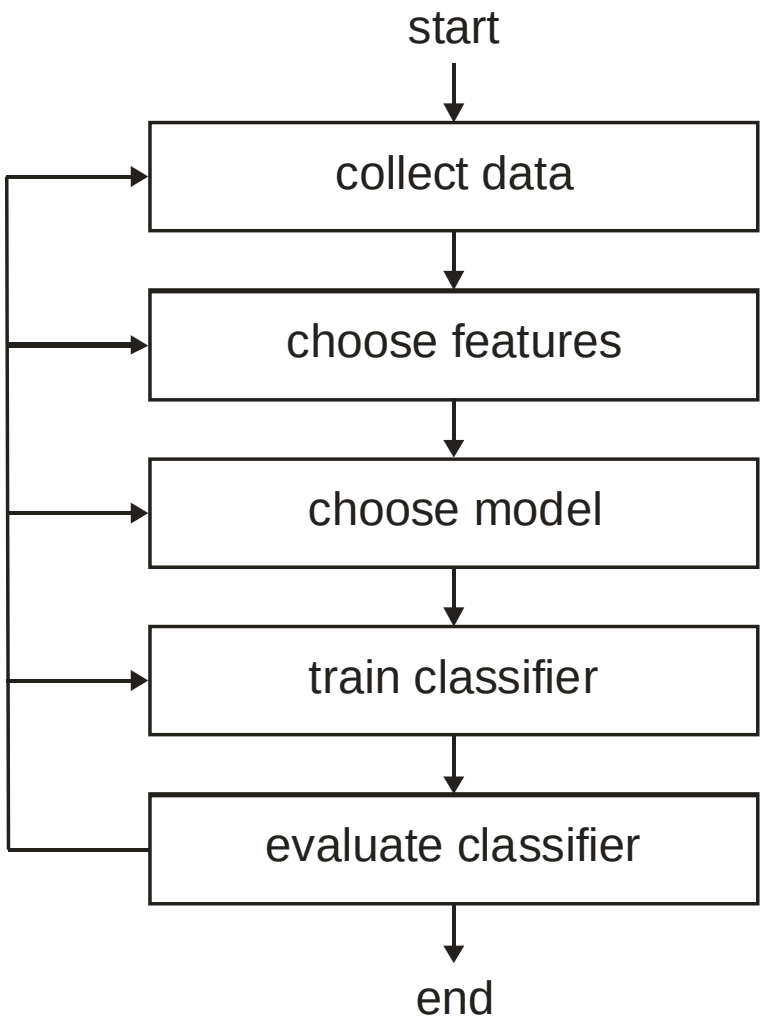
Feature invariance!

Missing features,
use of context information

Use of context information,
risk estimation

EPART interest domain

Design cycle



Prior domain knowledge.

Model invariance with respect to the expected distortions.

Possibility to fit too much to the training data (*overfitting*).

Subject matter

1. Introduction
2. Nearest neighbourhood methods
3. Optimal Bayes classification
4. Linear classification
5. Dimensionality reduction
6. Clustering
7. Test I

Subject matter

8. Neural networks I
9. Neural networks II
10. Markov models
    Exact string matching
11. Approximate string matching
12. Decision trees
13. Recognition quality enhancement I
14. Test II
15. Recognition quality enhancement II

Top 10 Algorithms in Data Mining discussed on EPART

+ C4.5 and beyond (decision trees)
+ The k-means algorithm
+ Support vector machines
- The Apriori algorithm
- The EM algorithm
- PageRank
+ AdaBoost
+ kNN – k-nearest neighbour classification
+ Naïve Bayes
+ CART

Wu, Xindong et al, Top 10 algorithms in data mining, Journal of Knowledge and Information Systems, Volume 14 Issue 1, December 2007, Pages 1-37