Machine learning basics

www.jussitohka.net

14th August 2017

Jussi Tohka

UEF // University of Eastern Finland

About me

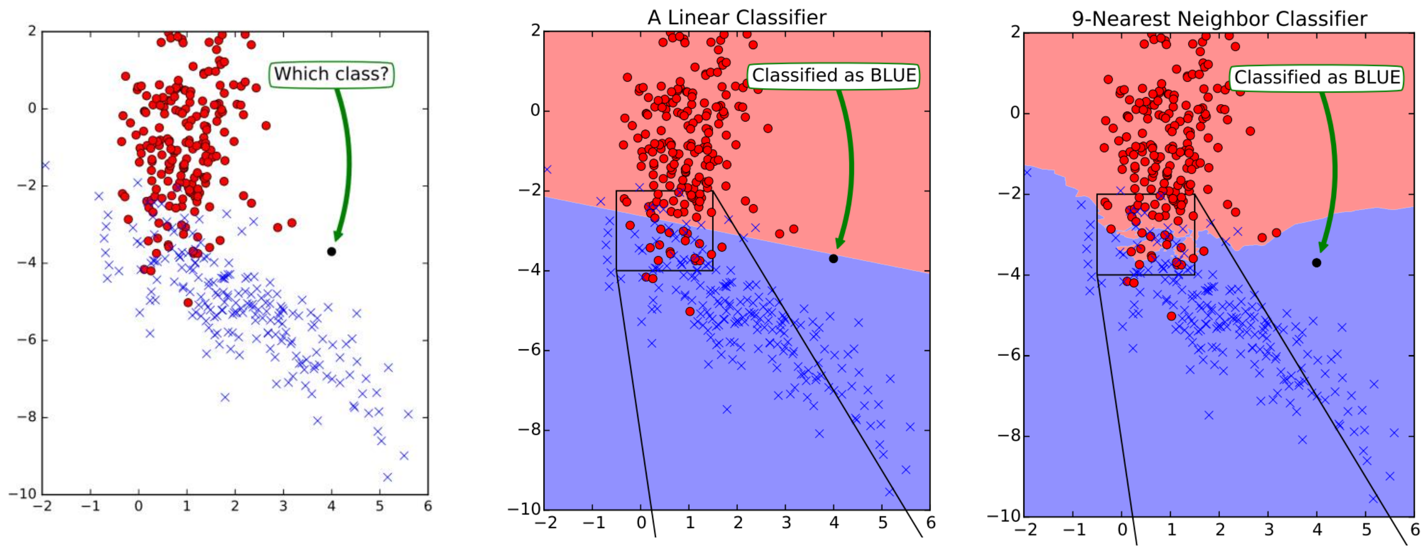




Currently, Associate professor at AI Virtanen Institute, University of Eastern Finland
2015 – 2016, CONEX professor at Universidad Carlos III de Madrid, Spain
2009 – 2014, Academy research fellow, team leader, Tampere University of Technology
2005 – 2009, Senior researcher (Academy post-doc, Scientific coordinator of STATCORE research cluster of excellence) Tampere University of Technology
2004 – 2005, Post-doc, Laboratory of Neuro Imaging, UCLA, USA
1999 – 2003, PhD in signal processing, Tampere, including the first of several visits to Montreal Neurological Institute

Machine learning (supervised)

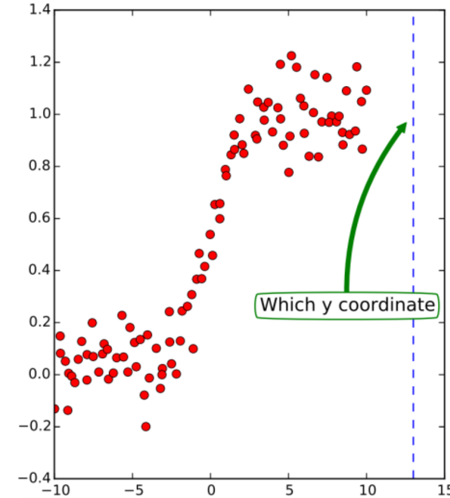
- Learn a function *f* from a training set $(X,Y) = \{(x\_i,y\_i), i = 1,\ldots,N\}$
- Then given **any** *x_unseen* we can find *f*(*x_unseen*) = *y_unseen*n
- The aim is to **generalize well** to unseen test data

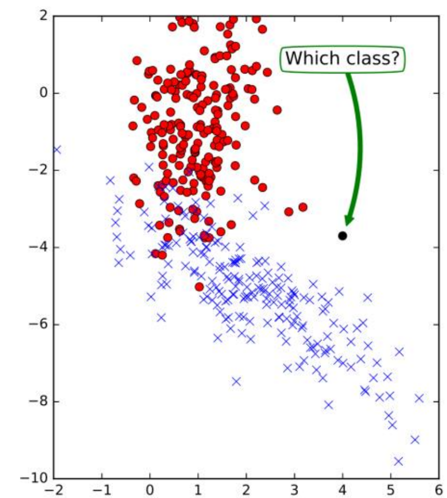


Figures by Heikki Huttunen TUT

Machine learning (supervised)

- **Features:** *x_i, x_unseen* usually d-dimensional vectors (reals,integers,binary)
- **Targets:** *y_i* real-valued then **regression**
- **Targets:** *y_i* categorical ({red, blue} or {1,2,3,…,k}) then **classification**
- But these are not the only possibilities





Figures by Heikki Huttunen TUT

Where to get features?

- Example: Bag of words for spam filtering

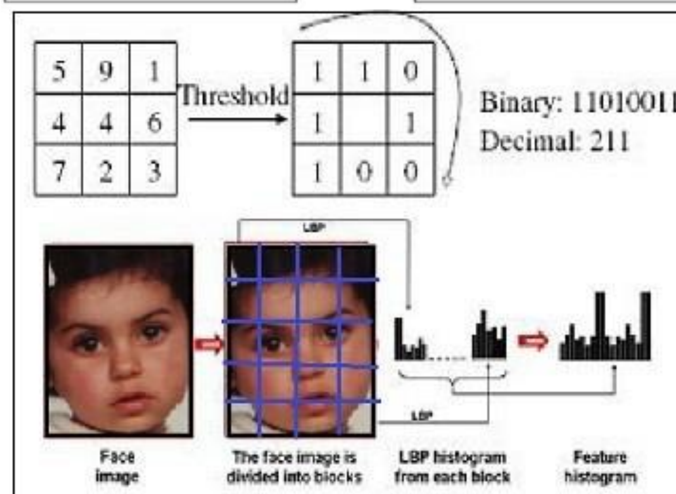
P(word | spam)

word	
the :	0.0156
to :	0.0153
and :	0.0115
of :	0.0095
you :	0.0093
a :	0.0086
with:	0.0080
from:	0.0075
...	

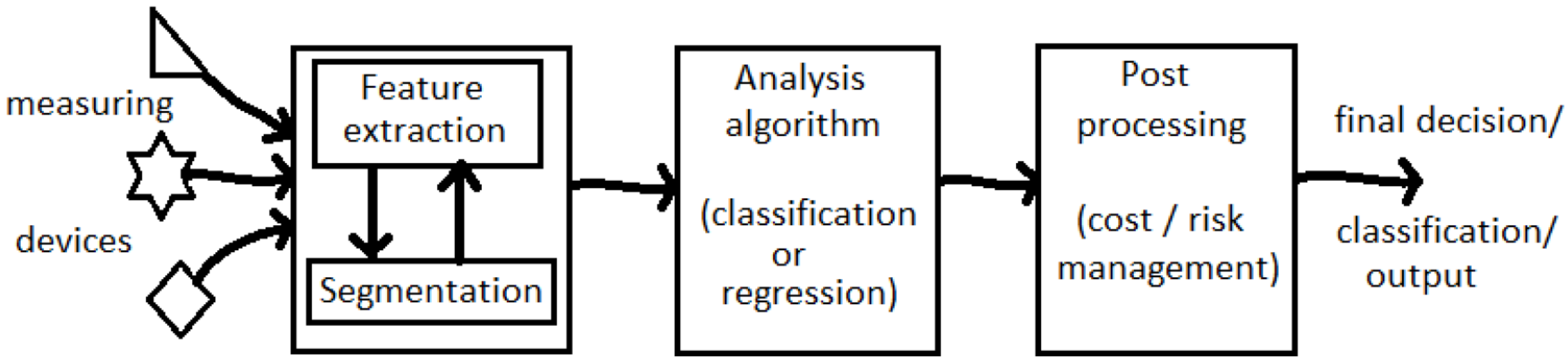
P(word | ¬spam)

word	
the :	0.0210
to :	0.0133
of :	0.0119
2002:	0.0110
with:	0.0108
from:	0.0107
and :	0.0105
a :	0.0100
...	

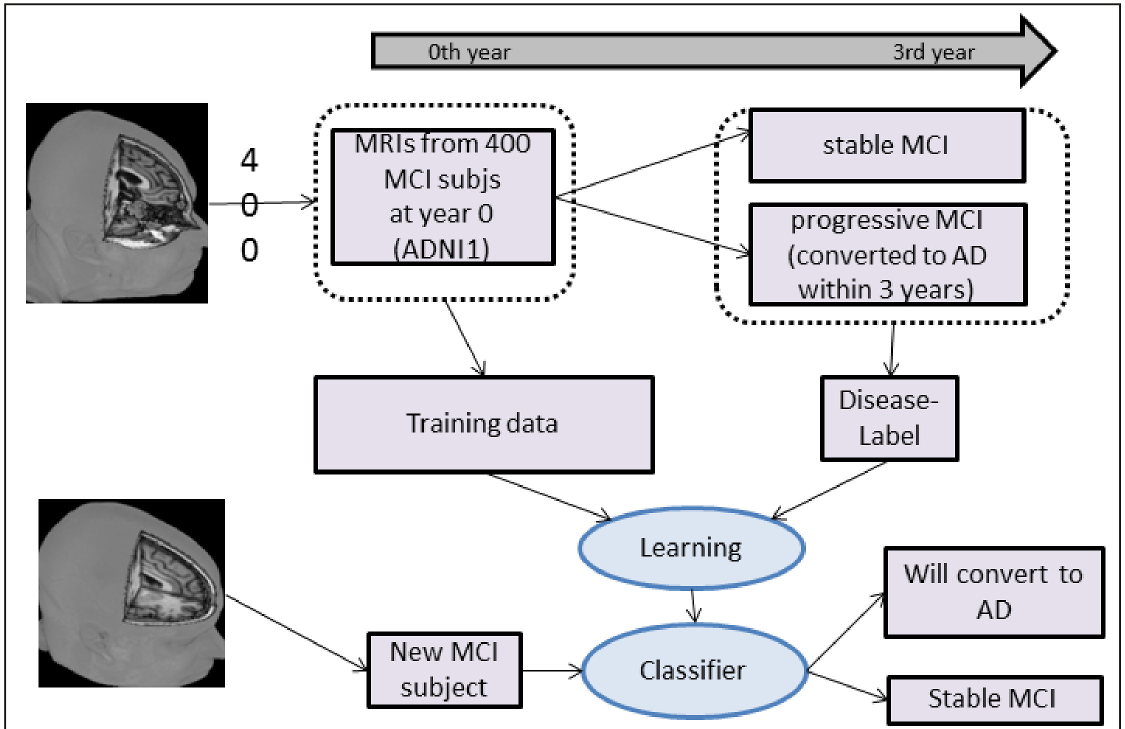
- Example 2: Local binary patterns for face recognition



Traditional view of pattern recognition system



Example: Early diagnosis of Alzheimer's Disease



Moradi, Pepe, Gaser, Huttunen, Tohka,
Neuroimage, 2015

Linear models for regression (least squares loss)

- Assume $y_i = b_0 + b_1 x_{i1} + \dots b_p x_{ip} + e_i$
- To make it simpler: $y_i = b_0 x_0 + b_1 x_{i1} + \dots b_p x_{ip} + e_i$
- e_i are i.i.d. (identically and independently distributed) Gaussian
- To solve b_i minimize $\sum_i (y_i - \sum_j b_j x_{ij})^2$
- Define $X = \begin{pmatrix} x_{10} & \cdots & x_{1p} \\ \vdots & \ddots & \vdots \\ x_{n0} & \cdots & x_{np} \end{pmatrix}$, $\mathbf{y} = [y_1, \dots, y_n]^T$,
  $\mathbf{b} = [b_0, b_1, \dots, b_p]^T$
- $\mathbf{b} = (X^T X)^{-1} X^T \mathbf{y}$
- Then for any $\mathbf{x}_{unseen}$,

$$y_{unseen} = \mathbf{b}^T \mathbf{x}_{unseen}$$

Bayes optimal classifiers

- We want to place $\mathbf{x}$ into one of the classes (categories) $1, \ldots, c$.
- For the optimal class *opt*

$$p(opt|\mathbf{x}) \geq p(j|\mathbf{x}), \quad \forall j \in \{1, \ldots, c\},$$

  where $p(i|\mathbf{x}) = \frac{p(\mathbf{x}|i)p(i)}{p(\mathbf{x})}$, i.e., *posterior* probability equals *class-conditional pdf* times *prior* per *evidence*.
- This is *Bayes classifier* which minimizes the *classification error* over the whole feature space
- We can forget the evidence: *opt* is the class for which

$$p(\mathbf{x}|i)p(i) \geq p(\mathbf{x}|j)p(j), \quad \forall j \in \{1, \ldots, c\}.$$

Plug-in classifiers

- To use the Bayes decision rule, we must know:
  1. Priors $p(i)$ for each class
  2. Class conditional pdfs $p(\mathbf{x}|i)$ for each class - note that we must know the value $p(\mathbf{x}|i)$ for all $\mathbf{x}$.
- Plug-in classifiers: fix the parametric form of the pdfs and estimate parameters based on training data
- Linear Gaussian classifiers; naive Bayes (NB); Naive Gaussian Bayes (NGB)
- naive Bayes: $p(i)p(\mathbf{x}|i) = \prod_{k=1}^{d} p(x_k|i)p(i)$

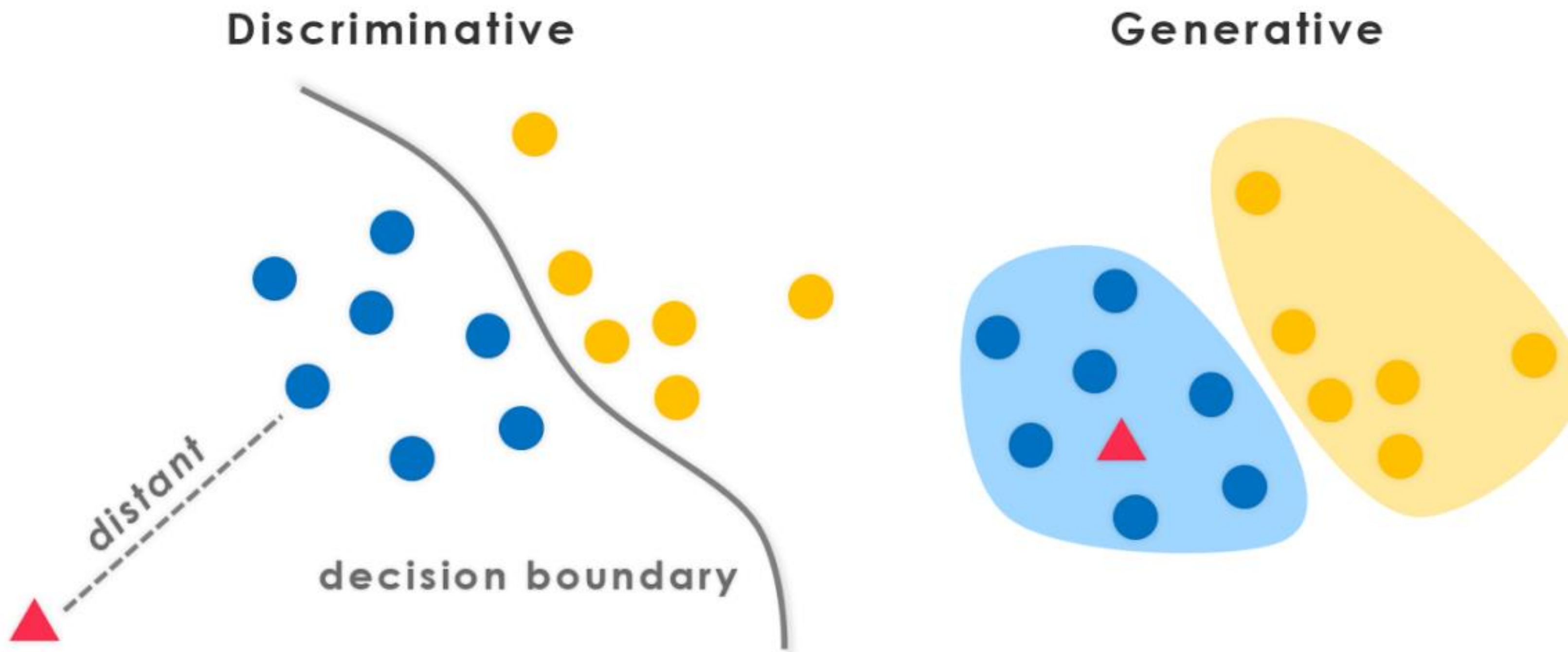
P(word | spam)

the	:	0.0156
to	:	0.0153
and	:	0.0115
of	:	0.0095
you	:	0.0093
a	:	0.0086
with	:	0.0080
from	:	0.0075
...		

P(word | ¬spam)

the	:	0.0210
to	:	0.0133
of	:	0.0119
2002	:	0.0110
with	:	0.0108
from	:	0.0107
and	:	0.0105
a	:	0.0100
...		

Discriminative classifiers



Most classifiers are discriminative; However, the theory in previous slides is important also to them

Figure: http://www.evolvingai.org/fooling

Linear classifiers

- Classes are separated by a linear boundary. To classify **x**, compute

$$F(\mathbf{x}) = \begin{cases} \text{Class 1}, & \text{if } \mathbf{w}^T\mathbf{x} < b \\ \text{Class 2}, & \text{if } \mathbf{w}^T\mathbf{x} \geq b \end{cases}$$

- Many flavours: 1) Fisher's LDA; 2) Support vector machines; 3) logistic regression

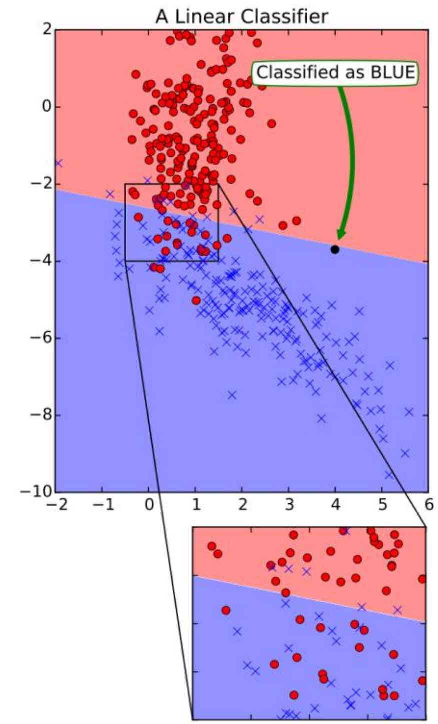
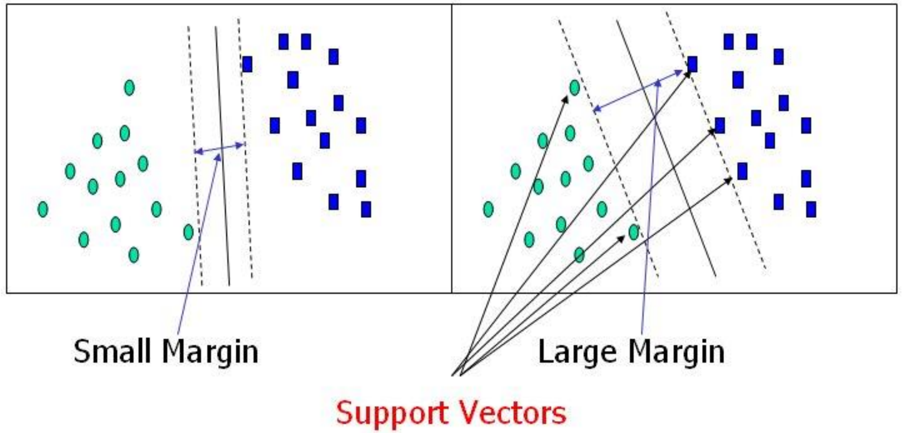
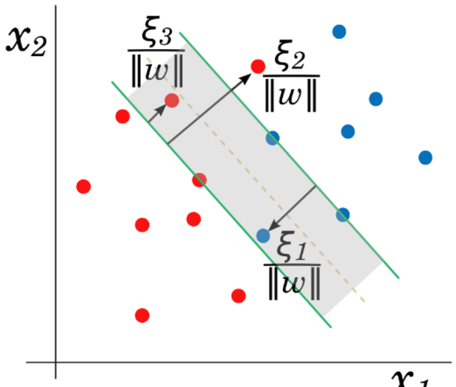


Figure by Heikki Huttunen TUT

Support vector machines

- Idea: maximize the margin between two classes
- Soft margins via using slack variables
- In practice, training consists of solving optimization problem
- Nonlinearity via kernels
- Robust for high dimensional data





K-nearest neighbours

- Find k nearest training samples to x and classify x based on majority vote among these k samples

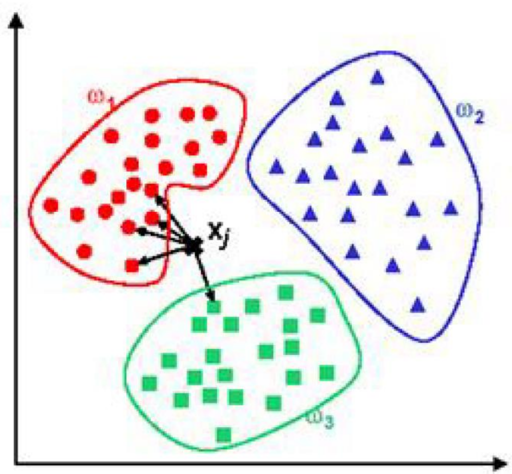


Figure Celebi: Neural Networks and Pattern Recognition Using MATLAB from `http://www.byclb.com/TR/Tutorials/neural_networks/`

Unsupervised machine learning

- When we don't know y_i
- Clustering
- Principal component analysis (PCA)
- Independent component analysis (ICA)
- For segmentation, feature extraction, data summarizaion…

Things to consider

- Training set is finite
  - Overfitting: too flexible model + too little data
  - Underfitting: a too rigid model
  - Generalization performance: how well the trained model works on unseen data?
- Optimality
  - Optimal in what sense? (classification accuracy, least-squares error,…)
  - Possibly different criteria in training and assessment
- Representative data: is the training data representative of the task?

Exercise for learning diary: Practical meaning of the prior

- Bayes formula (posterior): $p(\omega|x) = \frac{p(x|\omega)p(\omega)}{p(x)}$
- Example: The occurrence rate of a cancer C in a certain population P is 1%. A medical screening test T works with the following accuracy: the false negative rate is 5% and the false positive rate is 10%. Assume that subject A belongs to P and is tested with T which says that he has C (positive result). Given this information what is the probability that A truly has C; discuss?