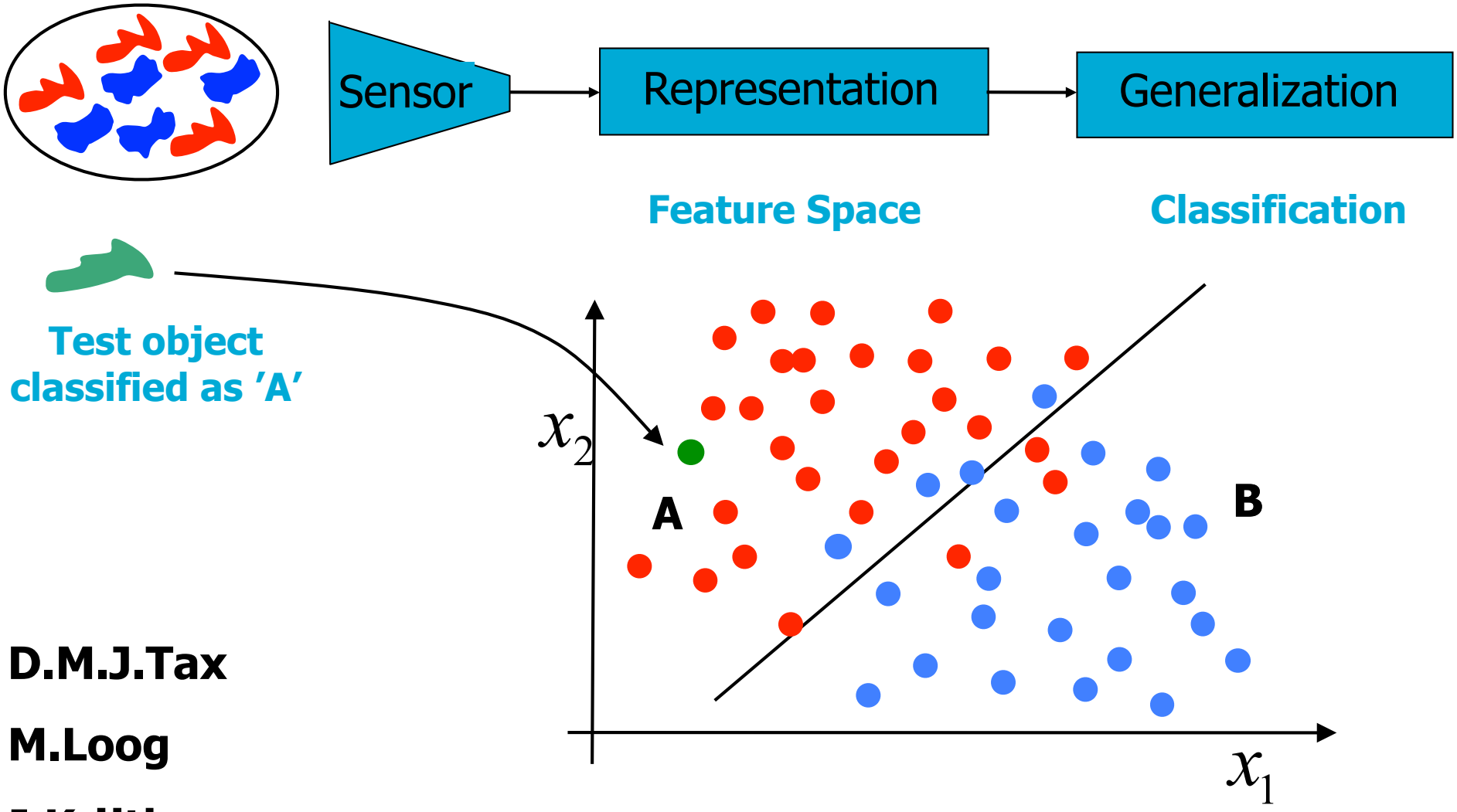
CS4220 Machine Learning 1



D.M.J.Tax

M.Loog

J.Krijthe

Pattern recognition: find the cat

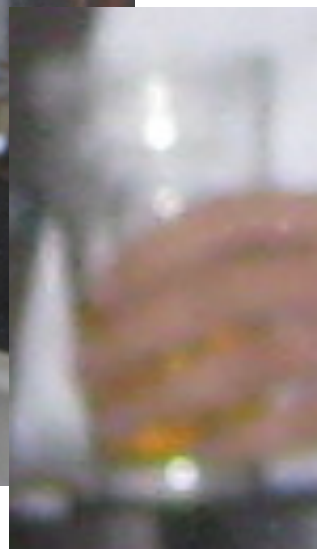
lft

Pattern recognition: find the cat

lft

Recognition problem

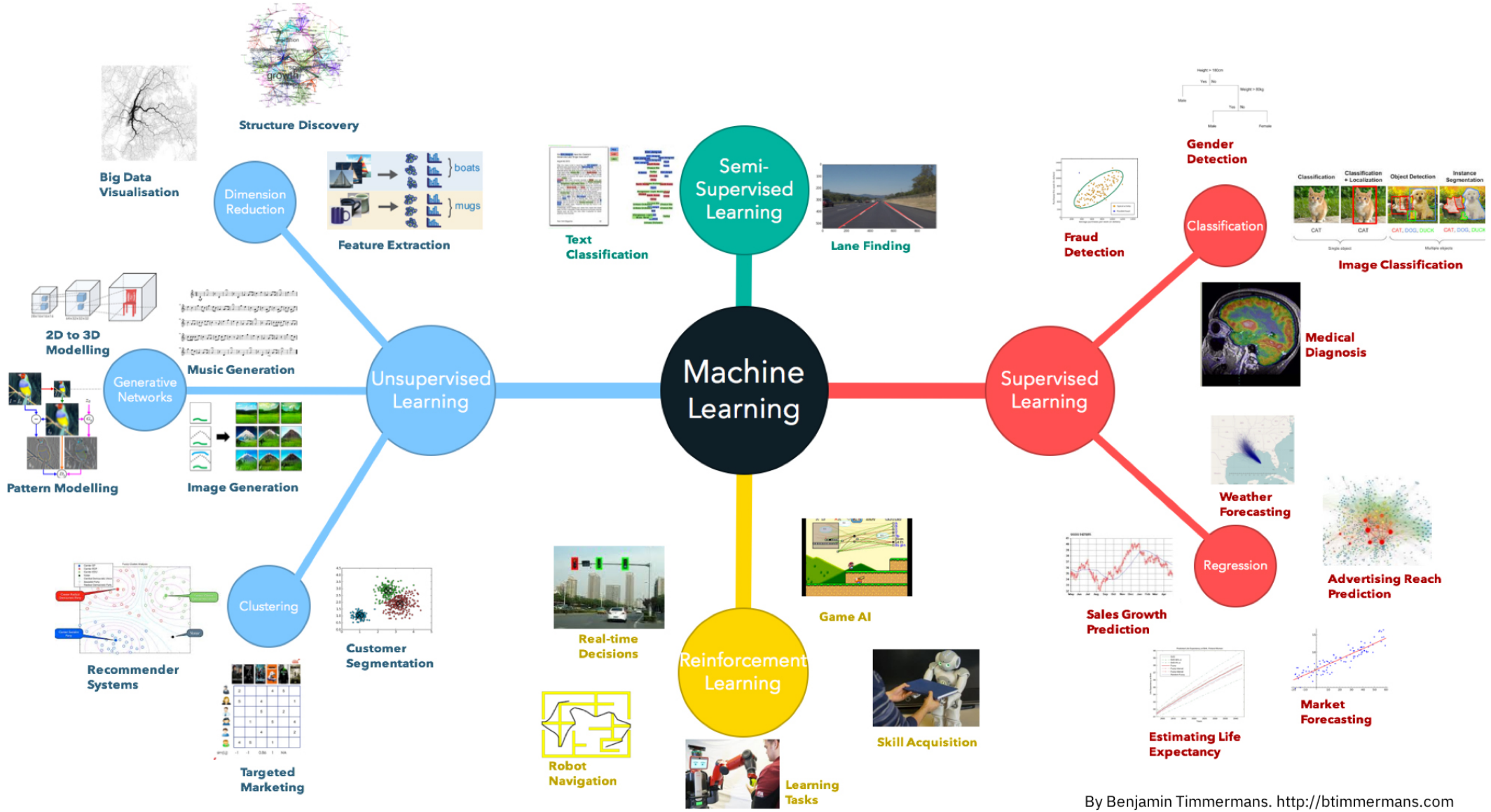
- What is happening? Where is this?
- Who is who? What is what?



Pattern recognition

Apple iPhoto'09

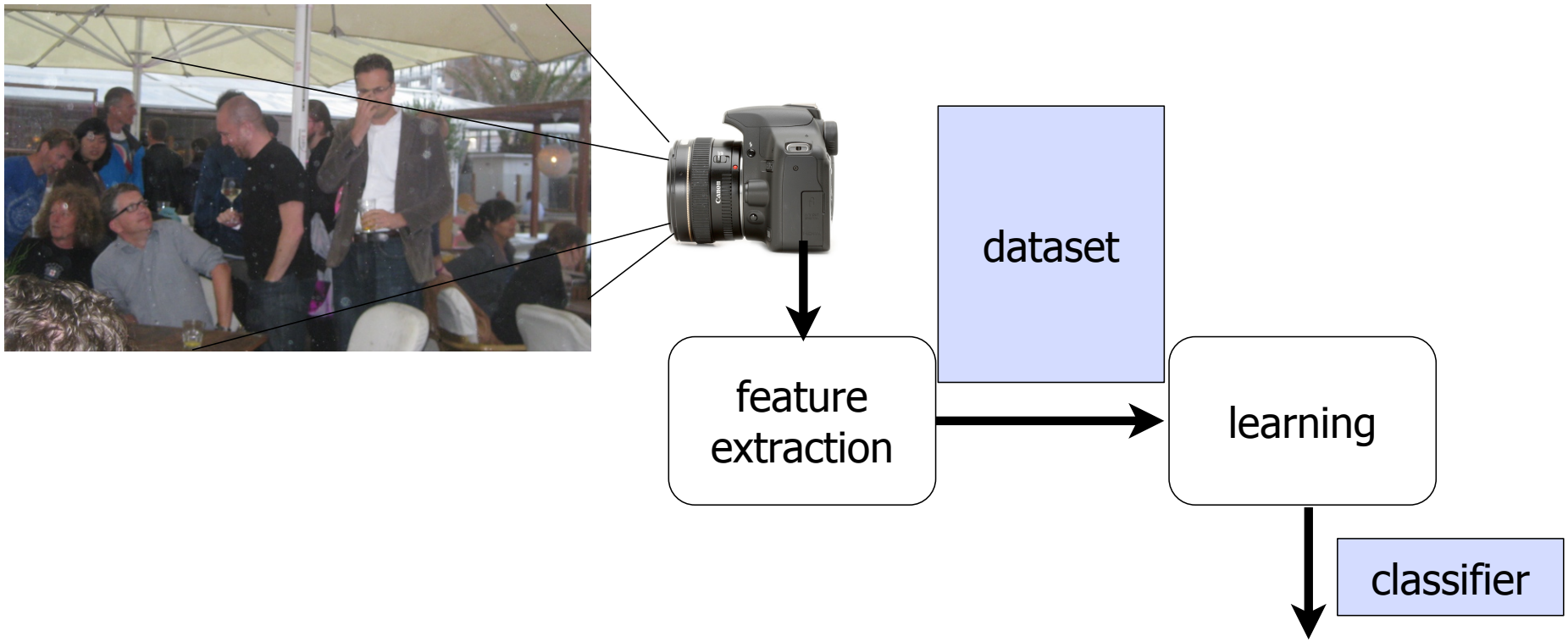
- Everyday tasks are deceptively difficult
- You have to be able to recognise places, situations, objects, people
- To do this automatically, is the task of pattern recognition: **learning from examples**



By Benjamin Timmermans. http://btimmermans.com

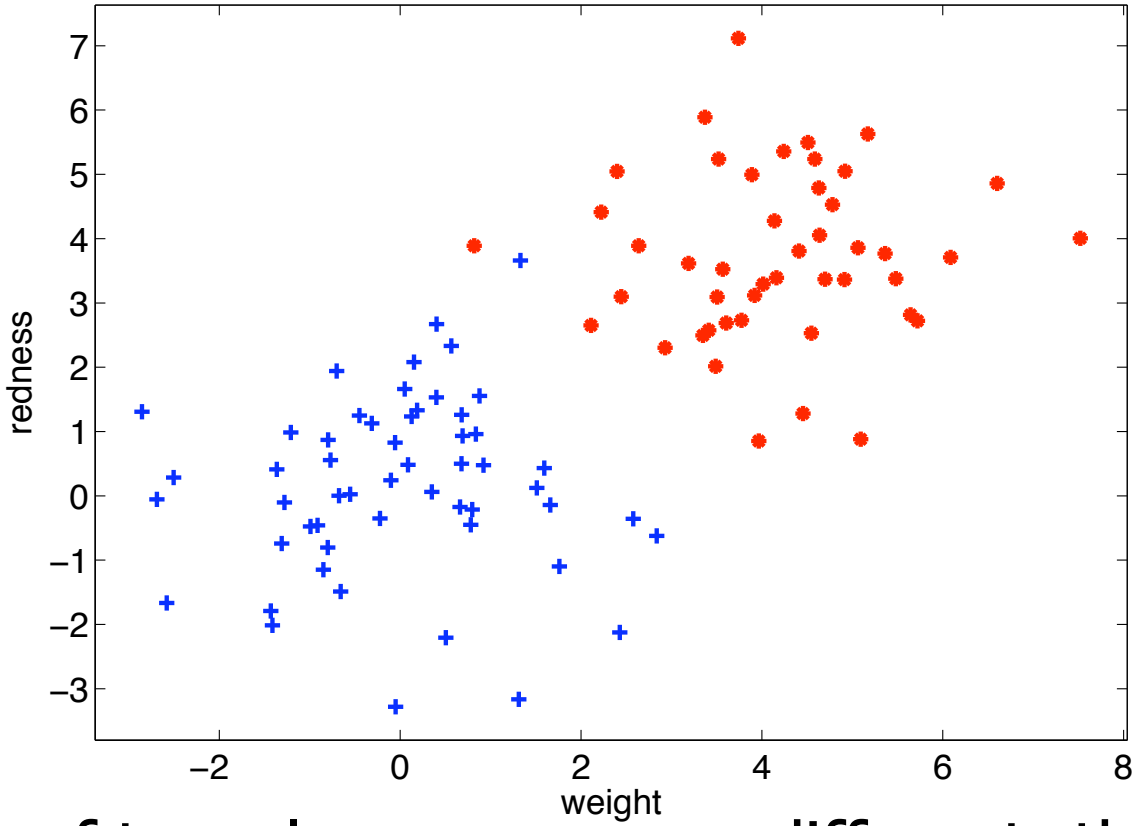
TU Delft

Pattern recognition pipeline



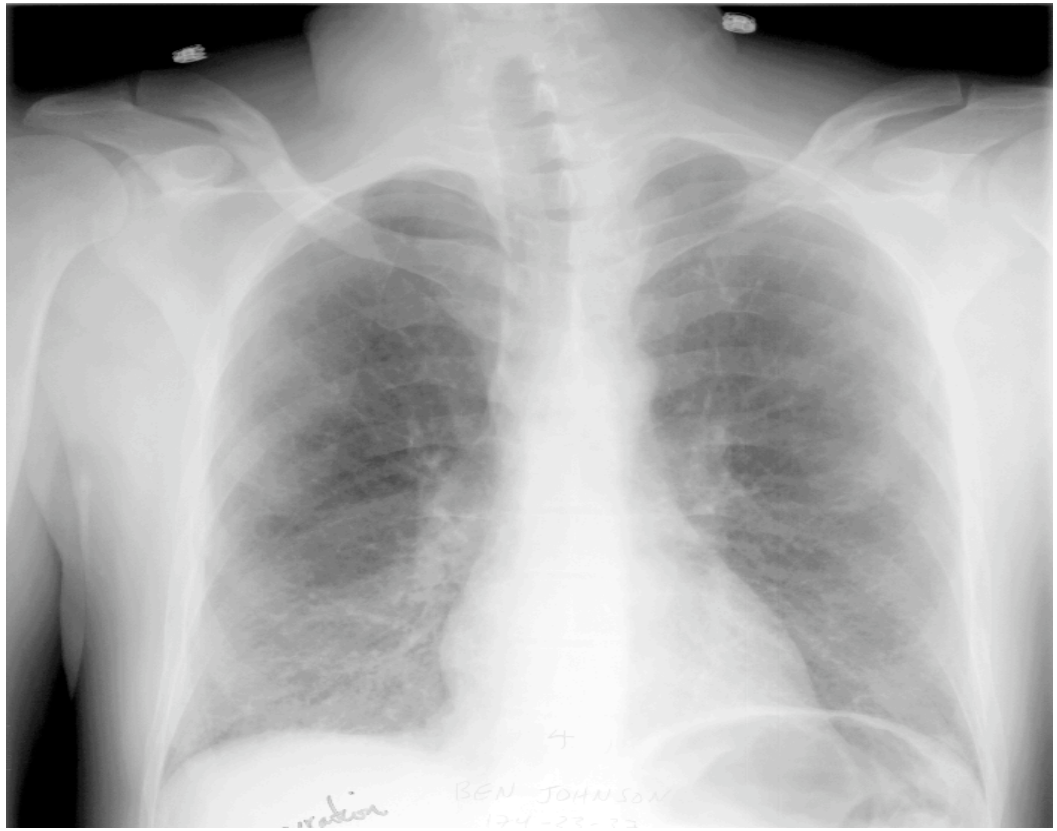
- From measurements, extract features, define labels, and train a classifier

(Good) features



- Objects of two classes are very different: there is ground truth
- Confusion by noisy measurements and poor features

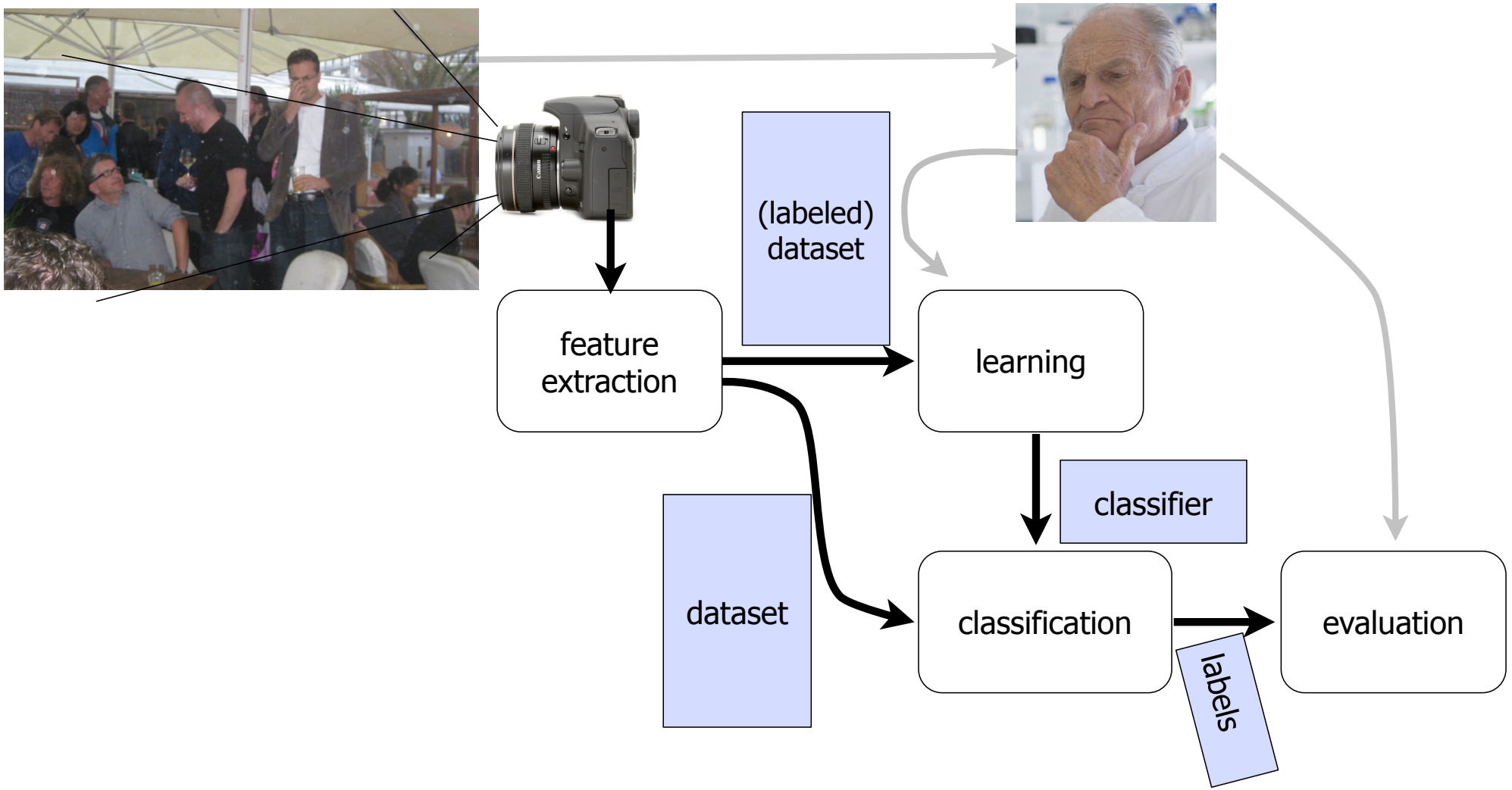
Poor features, hard problems



healthy
or
diseased?

- In many problems: experts are also not really sure.
- What features to choose is unclear

Pattern recognition pipeline



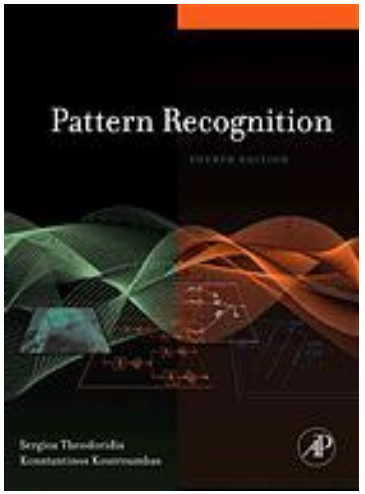
- Check with independent test data how well it works!

Machine Learning

- Prior knowledge:
  Linear algebra, basic probability theory and statistics
- Content of first week may be familiar to CS students who have followed CSE2510 Machine learning

- Material/literature:
  Different books/chapters per lecture

- This lecture: Section 2.1-2.6 of
  Pattern recognition, by Sergios Theodoridis, Konstantinos Koutroumbas (2009)



Schedule of the course

- Week 1: Basic Machine Learning: classification
- Week 2: Regression
- Week 3: Curse of dimensionality, overfitting
- Week 4: Complexity and Support Vector Machines
- Week 5: Bayesian learning, Clustering
- Week 6: Feature reduction, Combining
- Week 7: Design of ML experiments

- Exam!

Lectures, exercises and Exam

- Lectures are not mandatory
- (Non-mandatory) exercises in Python/Matlab
- Manual on Brightspace (check regularly!)
- Python library at: https://github.com/DMJTax/prtools
- Two hours per week, TAs present to answer questions (queue.tudelft.nl)

- Exams are digital.
- Midterm exam to practice and see how good your knowledge is (and if the systems are working)
- Final exam: grade!

This week

- Introduction, administrative stuff
- Learning from examples, some definitions
- Classification
- Bayes rule, Bayes error
- Misclassification costs
- Parametric classifiers: Quadratic, linear, nearest mean classifiers,
- Non-parametric classifiers: Parzen, kNN
- (Logistic?)

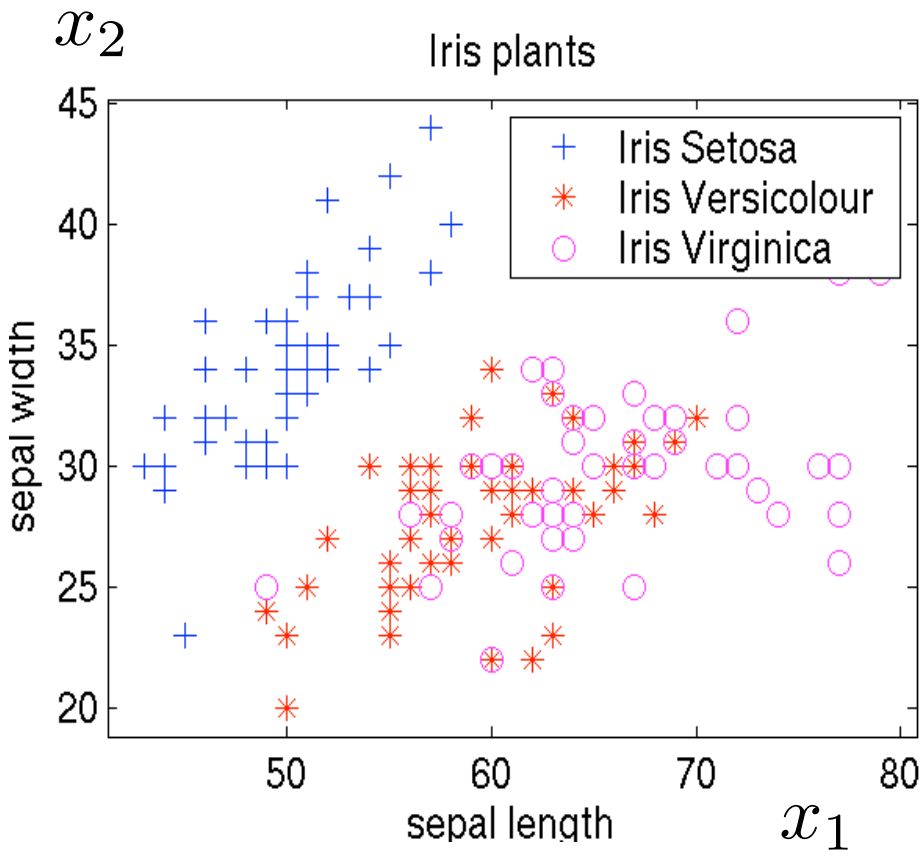
Objects in feature space

- We can interpret the measurements as a vector in a vector space:

$$\mathbf{x} = (x_1, x_2, ..., x_p)$$

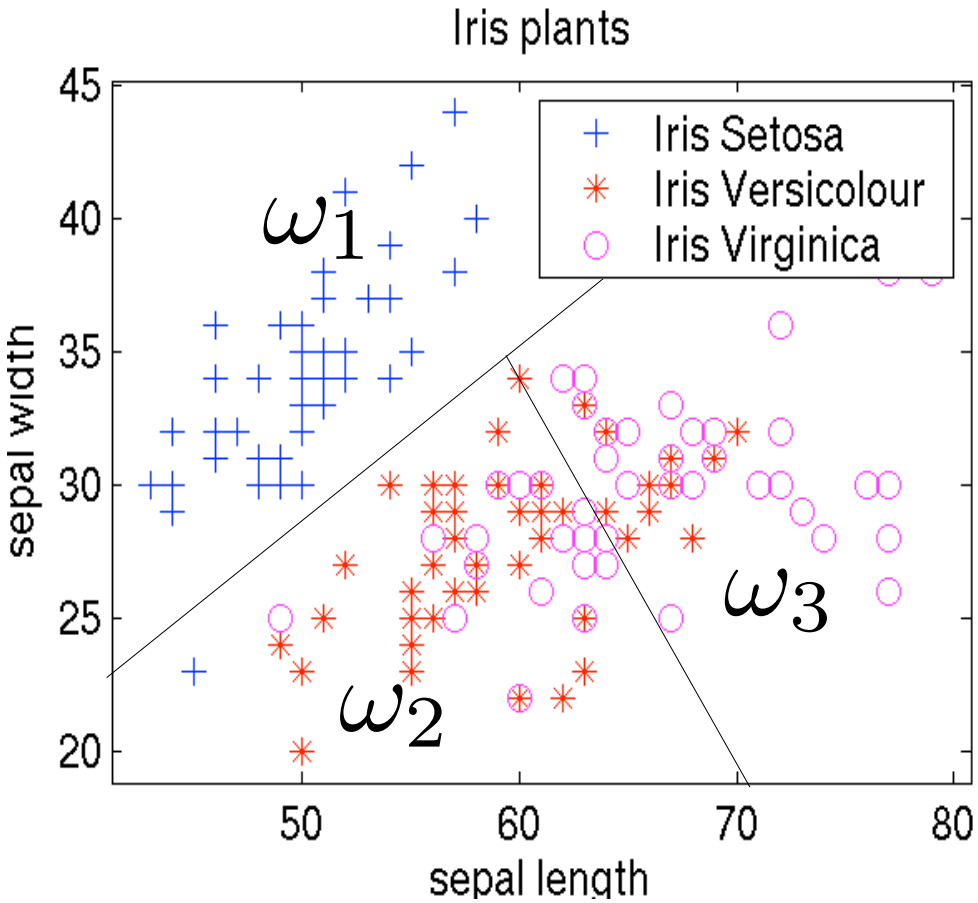
- This originates, in principle, from a probability density over the whole feature space

$$p(\mathbf{x}, y)$$

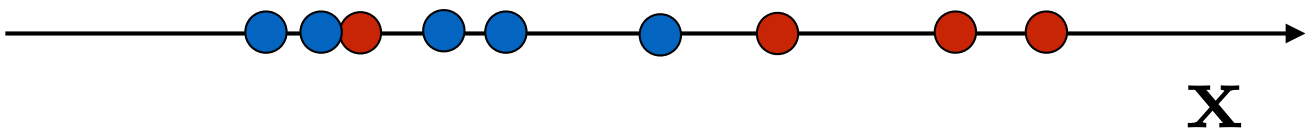


Classification

- Given labeled data: $\mathbf{x}$
- Assign to each object a class label ω
- In effect splits the feature space in separate regions

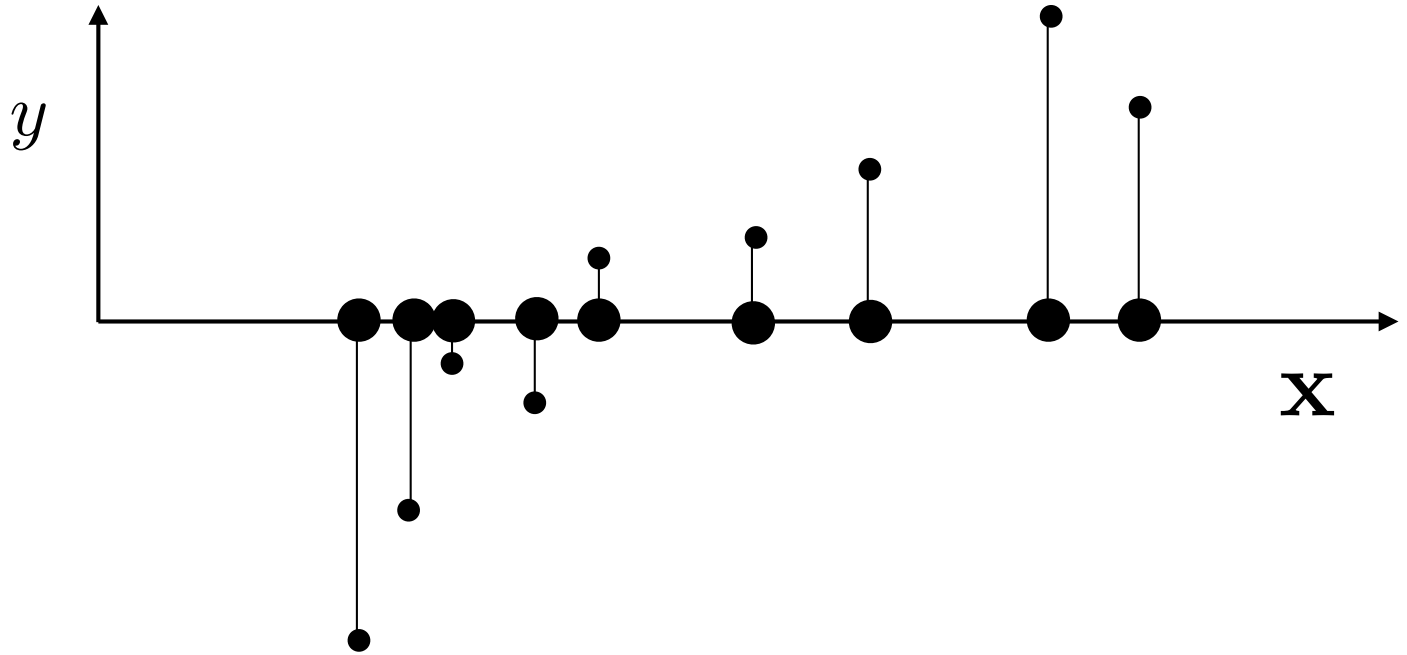


Also for regression, clustering, ...



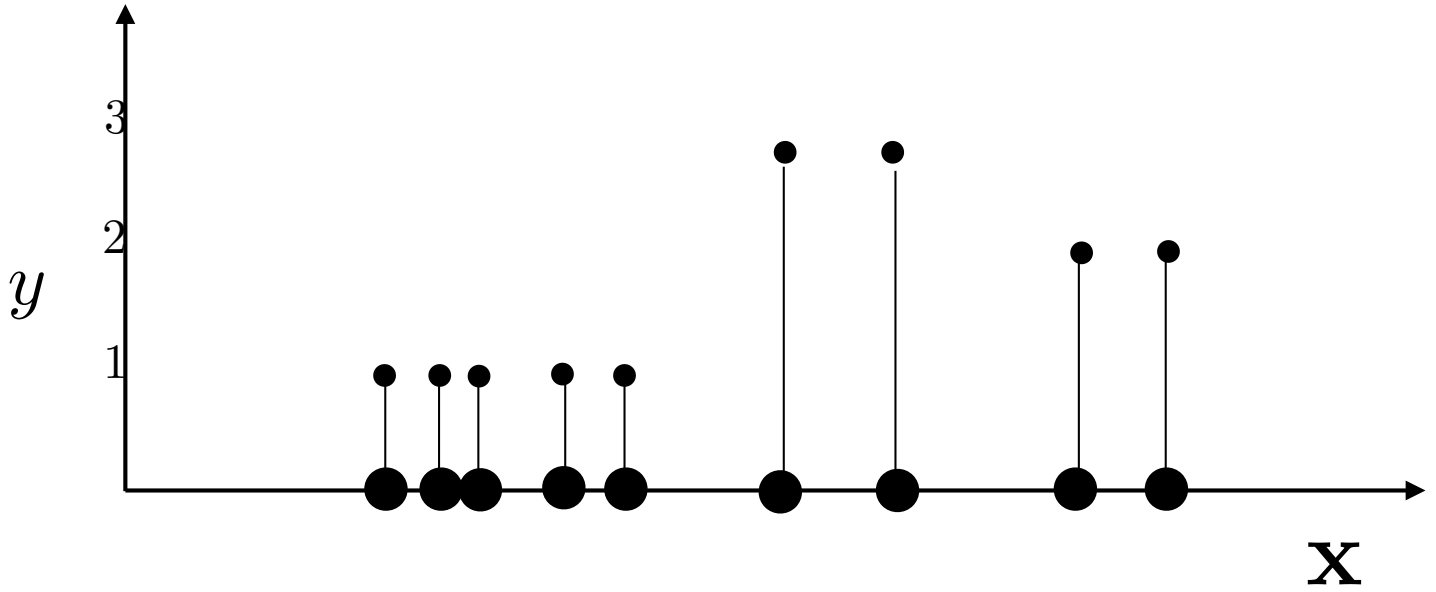
- Instead of class labels, real values: regression
- Instead of supervised labels, no labels: clustering

Also for regression, clustering, ...



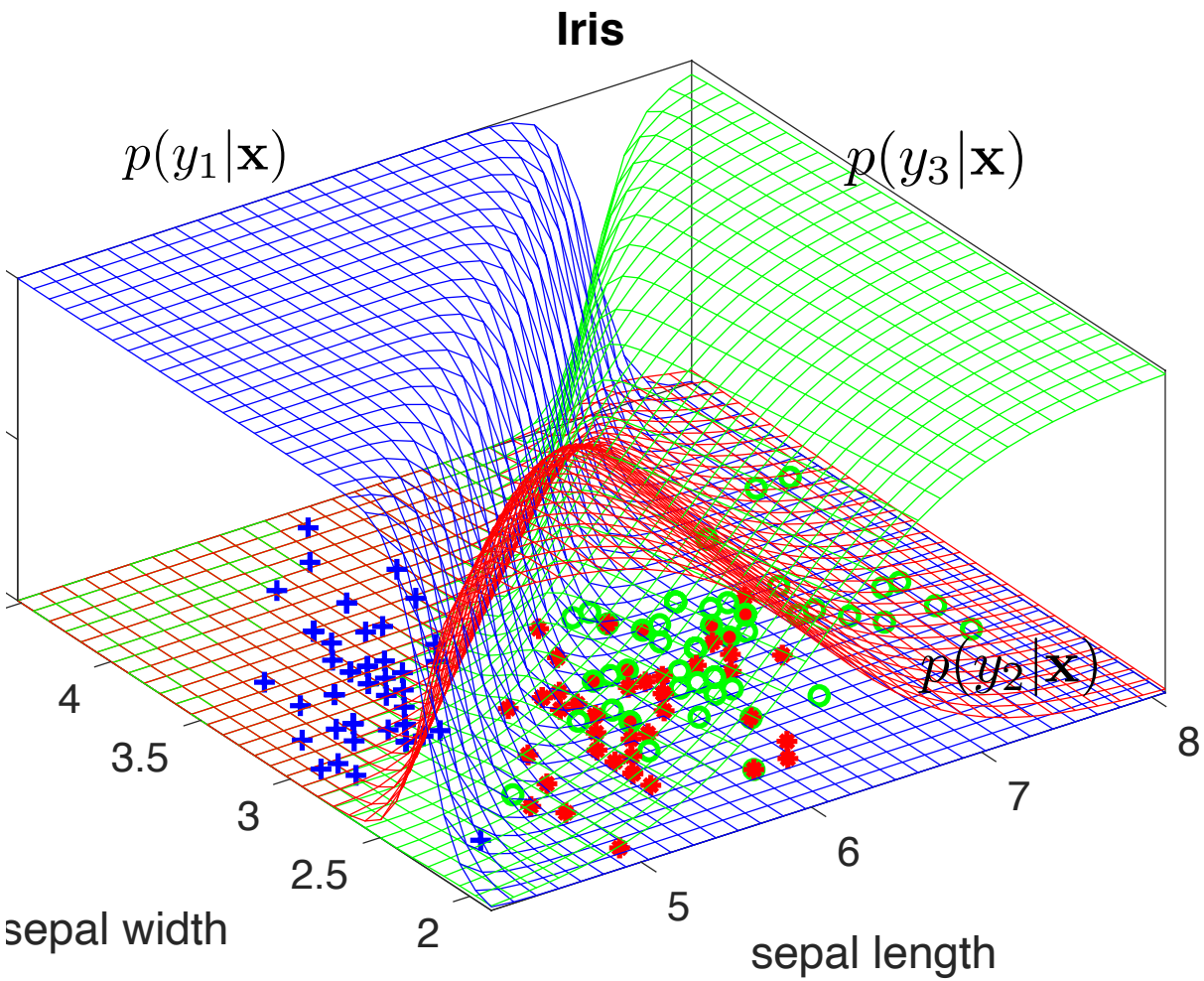
- Instead of class labels, real values: regression
- Instead of supervised labels, no labels: clustering

Also for regression, clustering, ...



- Instead of class labels, real values: regression
- Instead of supervised labels, no labels: clustering

Output of the model



- For each object in the feature space, we should find:

$$p(y|\mathbf{x})$$

- In practice, we approximate: $\hat{p}(y|\mathbf{x})$
- or we fit a function:

$$f(\mathbf{x})$$

Classification

- Classifier:

  If

$$p(y_1|\mathbf{x}) > p(y_2|\mathbf{x})$$

  assign **x** to class y_1, otherwise y_2.

Description of a classifier

There are several ways to describe a classifier:

- if $p(y_1|\mathbf{x}) > p(y_2|\mathbf{x})$ then assign to y_1
  otherwise y_2

- if $p(y_1|\mathbf{x}) - p(y_2|\mathbf{x}) > 0$ then assign to y_1

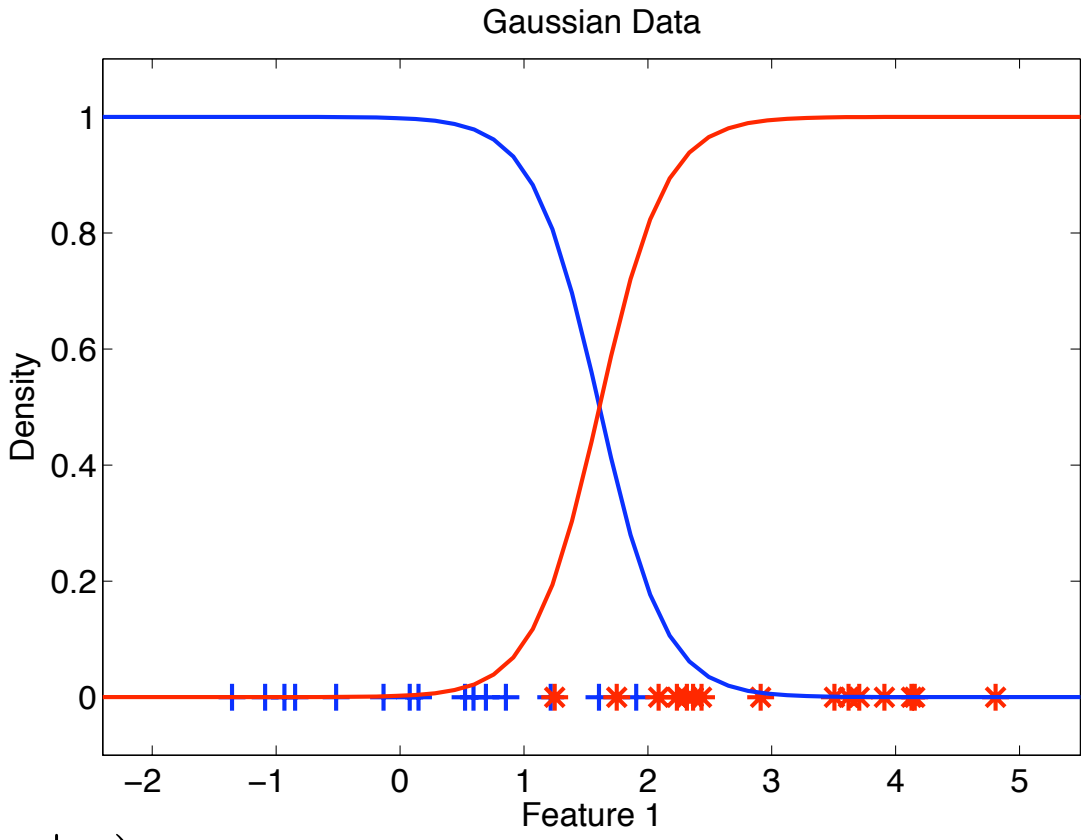
- or $\dfrac{p(y_1|\mathbf{x})}{p(y_2|\mathbf{x})} > 1$

- or $\log(p(y_1|\mathbf{x})) - \log(p(y_2|\mathbf{x})) > 0$

Class posterior probability

- For each object we have to estimate $p(\omega|\mathbf{x})$

$p(\omega|\mathbf{x})$

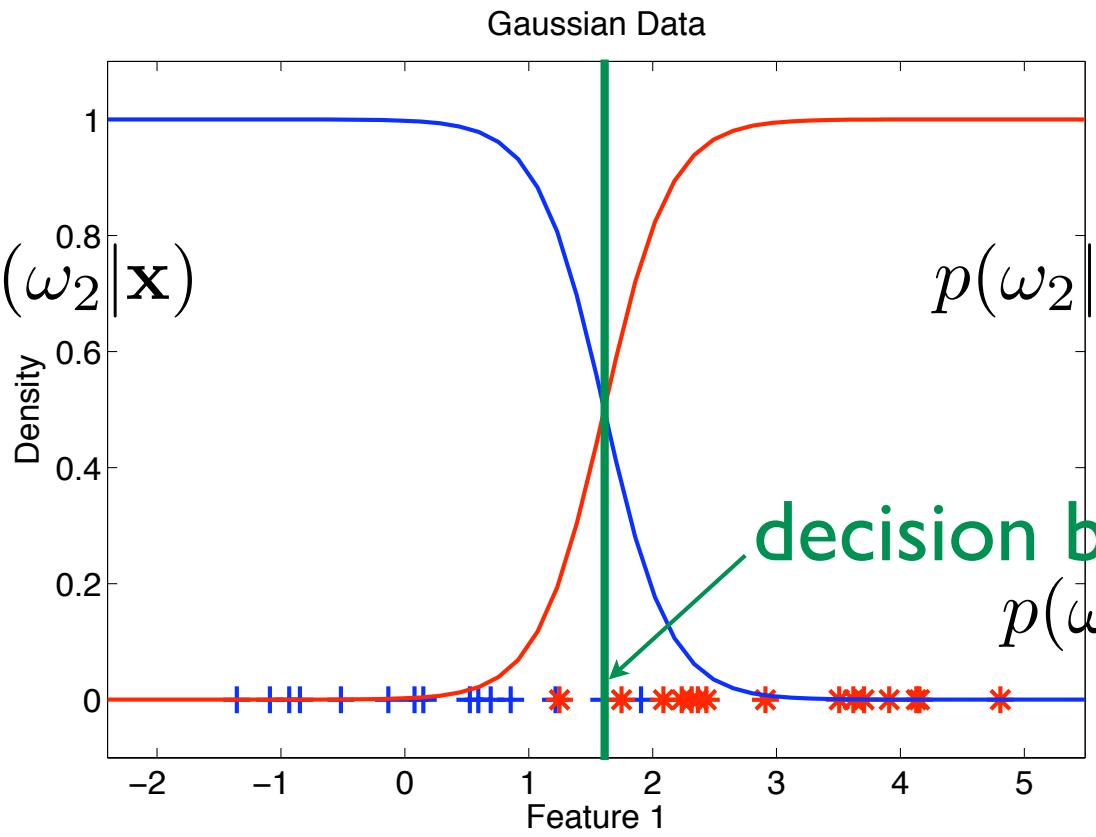


$$\sum_i p(\omega_i|\mathbf{x}) = 1$$

Classify new objects

- Assign the label of the class with the largest posterior probability

$p(\omega_1|\mathbf{x}) > p(\omega_2|\mathbf{x})$



$p(\omega_2|\mathbf{x}) > p(\omega_1|\mathbf{x})$

decision boundary

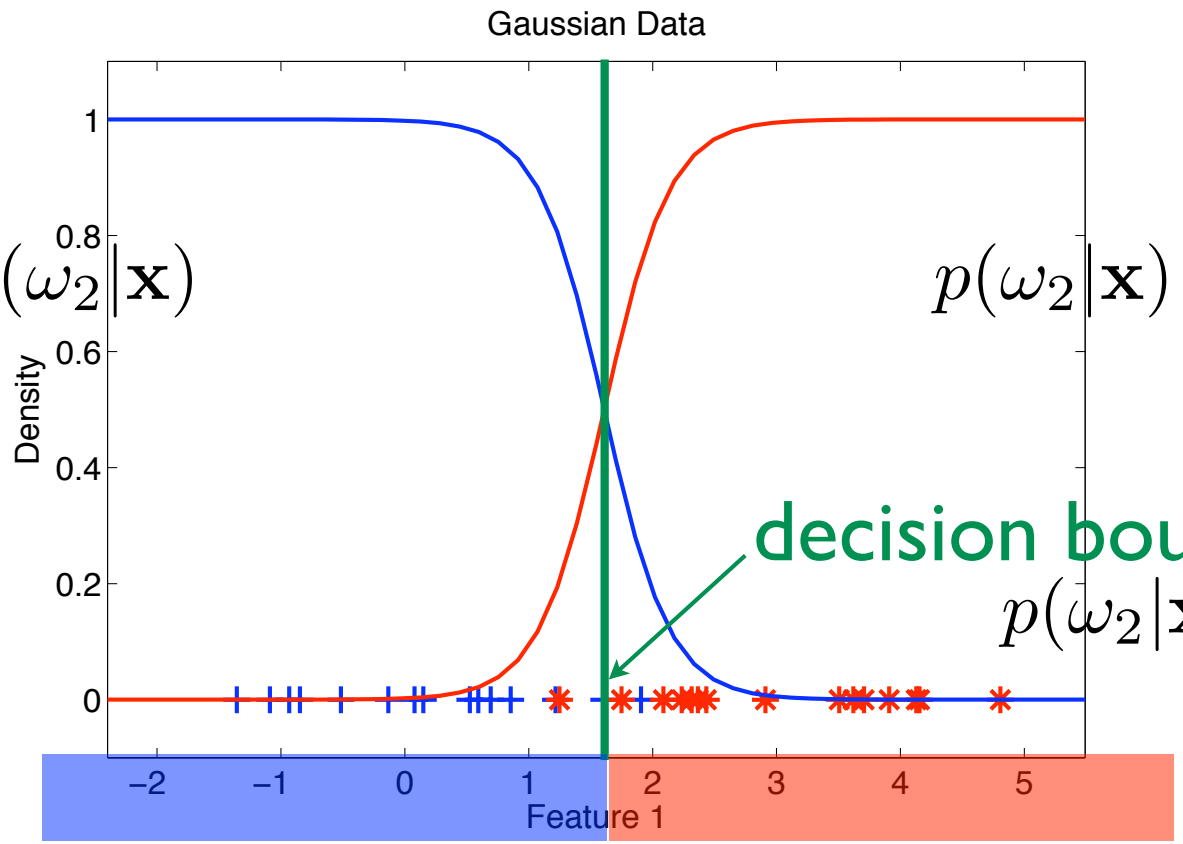
$p(\omega_2|\mathbf{x}) = p(\omega_1|\mathbf{x})$

Classify new objects

- Assign the label of the class with the largest posterior probability

$p(\omega_1|\mathbf{x}) > p(\omega_2|\mathbf{x})$

$p(\omega_2|\mathbf{x}) > p(\omega_1|\mathbf{x})$



decision boundary

$p(\omega_2|\mathbf{x}) = p(\omega_1|\mathbf{x})$

Ω_1

Ω_2

Bayes' theorem

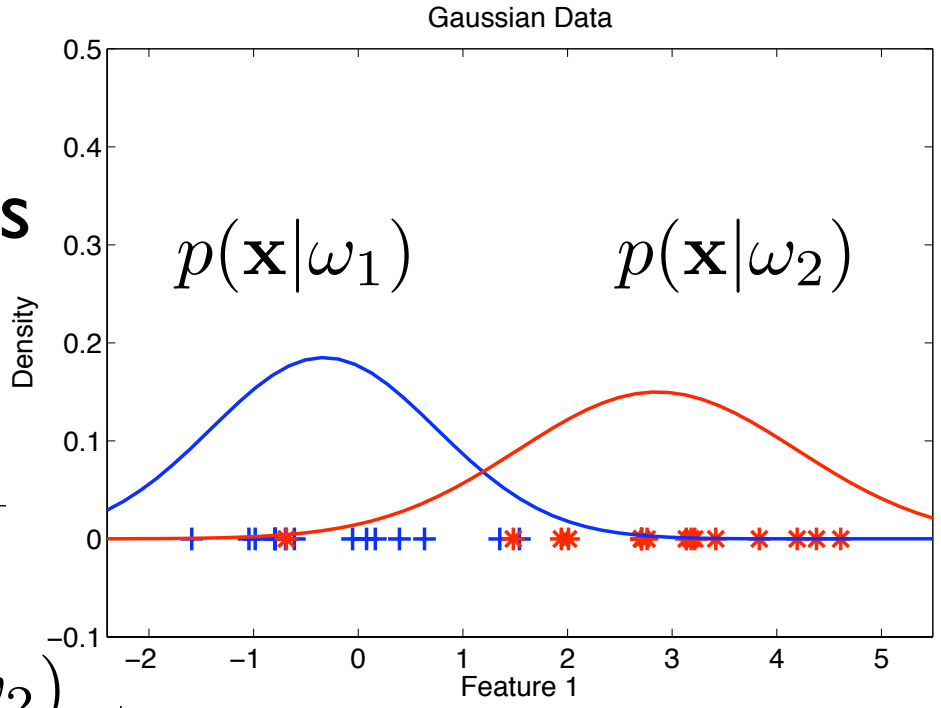
- In many cases the posterior is hard to estimate
- Often a functional form of the class distributions can be assumed
- Use Bayes' theorem to rewrite one into the other:

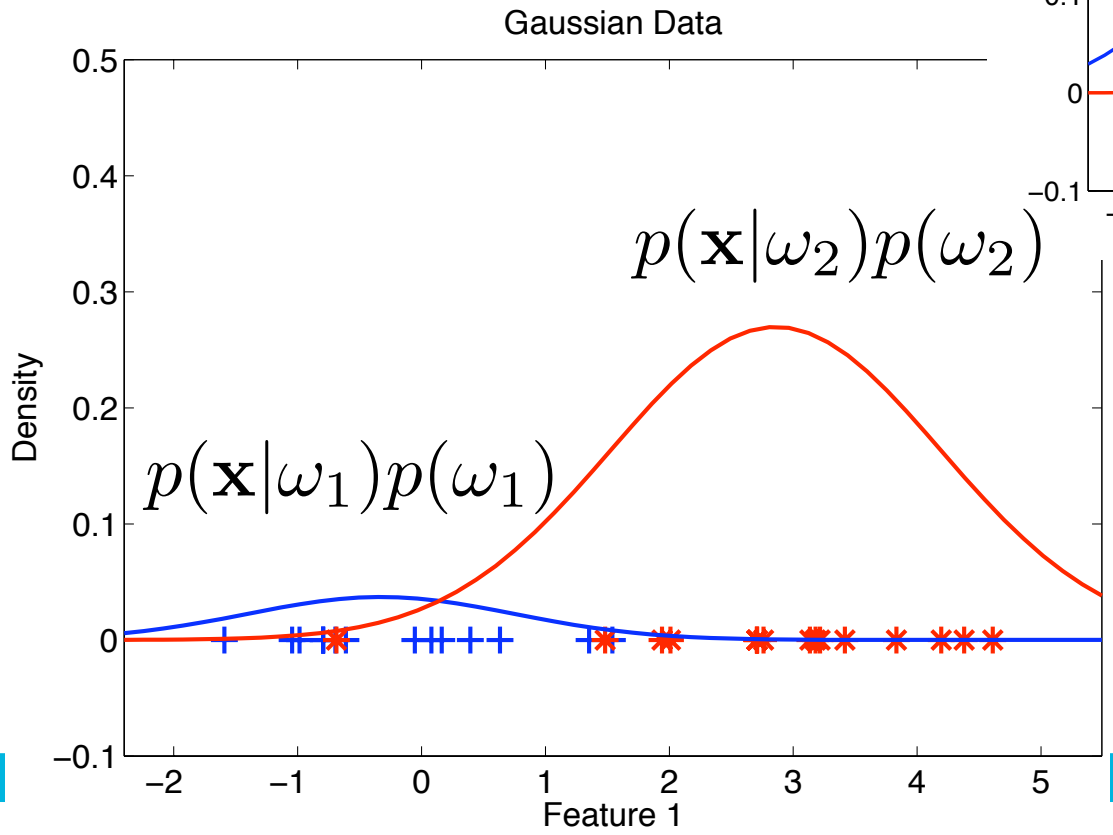
$$p(\omega|\mathbf{x}) = \frac{p(\mathbf{x}|\omega)p(\omega)}{p(\mathbf{x})}$$

class (conditional) distribution $p(\mathbf{x}|\omega)$
class prior $p(\omega)$
(unconditional) data distribution $p(\mathbf{x})$

Bayes' rule

1. Estimate the class conditional probabilities



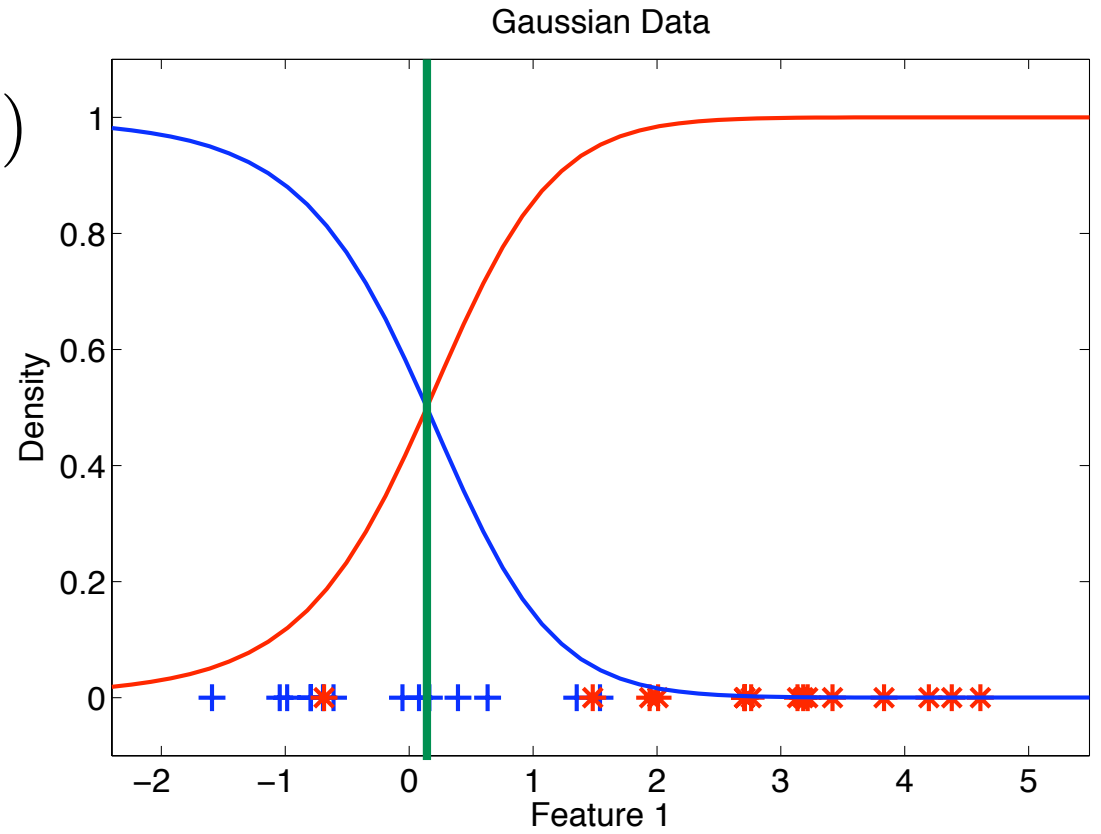


2. Multiply with the class priors

Bayes' rule

3. Compute the class posterior probabilities
4. Assign objects to the class with the highest posterior probability

$p(\omega_1|\mathbf{x})$

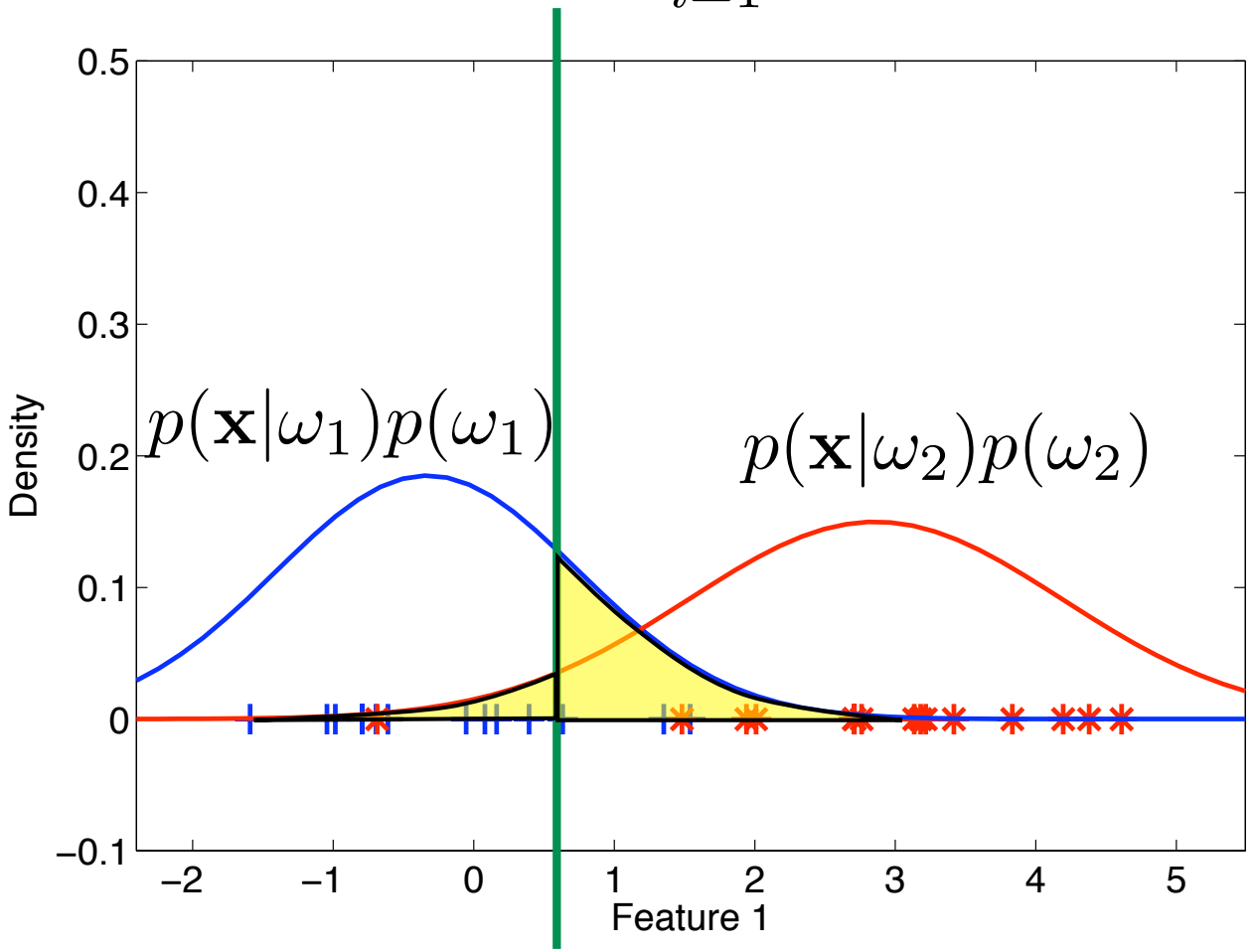


$p(\omega_2|\mathbf{x})$

Classification error

- The error:

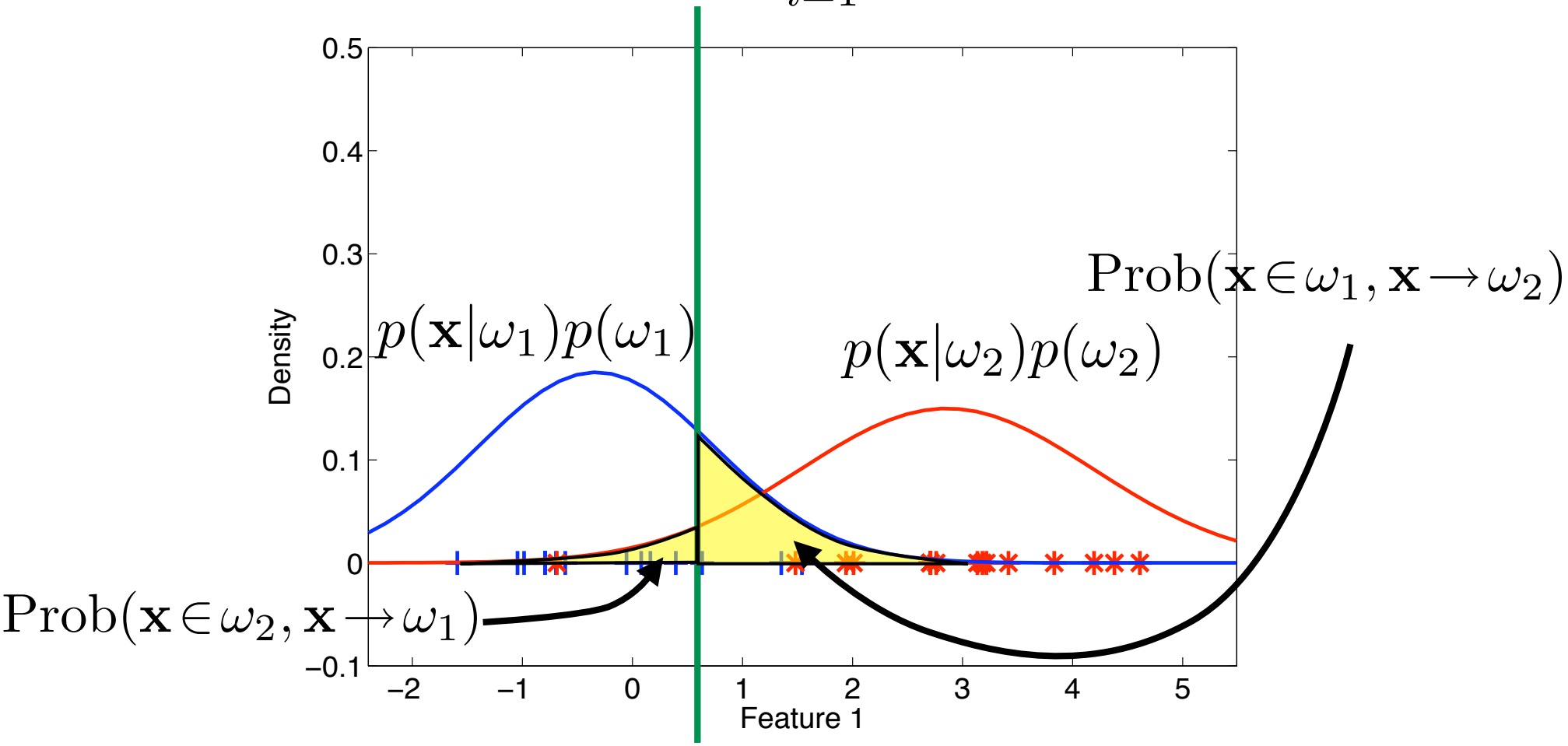
$$P(\text{error}) = \sum_{i=1}^{C} P(\text{error}|\omega_i)P(\omega_i)$$



Classification error

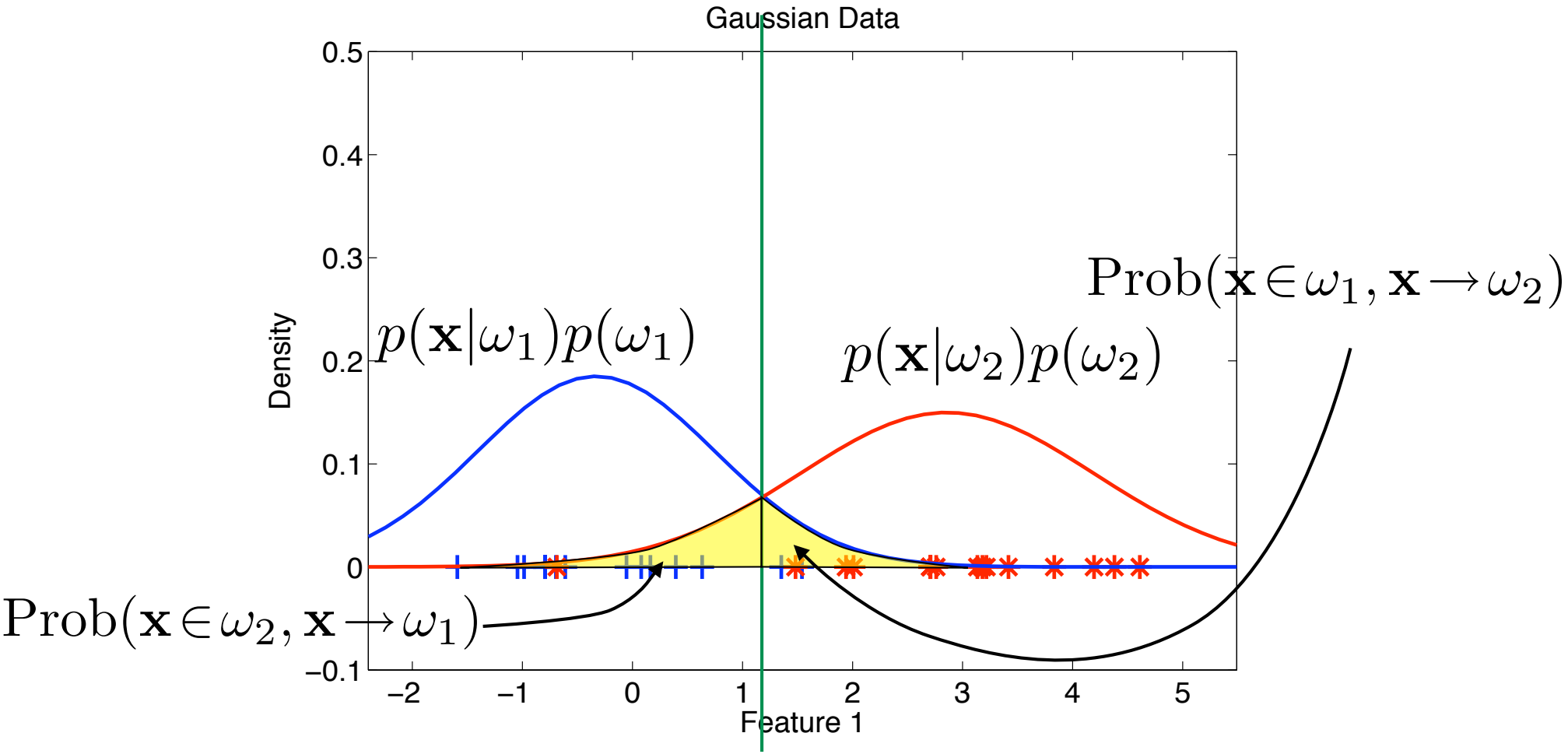
- The error:

$$P(\text{error}) = \sum_{i=1}^{C} P(\text{error}|\omega_i)P(\omega_i)$$



ε^* Bayes error

Bayes error is the **minimum** error: typically >0 !!



TU Delft

Bayes' Error

- Bayes' error is the **minimum** attainable error
- In practice, we do not have the true distributions, and we can not obtain ε^*
- The Bayes' error does not depend on the classification rule that you apply, but on the distribution of the data
- In general you can not compute the Bayes' error:
  - you don't know the true class conditional probabilities
  - the (high) dimensional integrals are very complicated

Misclassification Costs

- Sometimes: misclassification of class A to class B is much more dangerous than misclassification of class B to class A

misclassification:
classify 'healthy' to 'ill'

misclassification:
classify 'ill' to 'healthy'

Misclassification cost

- Introduce a loss that measures the cost of assigning an object that came from class ω_j to class ω_i : λ_{ji}

$\lambda_{\text{healthy,ill}} = 10$

$\lambda_{\text{ill,healthy}} = 100$

Conditional risk, total risk

- The conditional risk of assigning object **x** to class ω_i :

$$l^i(\mathbf{x}) = \sum_{j=1}^{C} \lambda_{ji} p(\omega_j|\mathbf{x})$$

- The average risk over a region:

$$\begin{aligned} r^i &= \int_{\Omega_i} l^i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \\ &= \int_{\Omega_i} \sum_{j=1}^{C} \lambda_{ji} p(\omega_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} \end{aligned}$$

- Overall risk:

$$r = \sum_{i=1}^{C} r^i = \sum_{i=1}^{C} \int_{\Omega_i} \sum_{j=1}^{C} \lambda_{ji} p(\omega_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

Minimum total risk

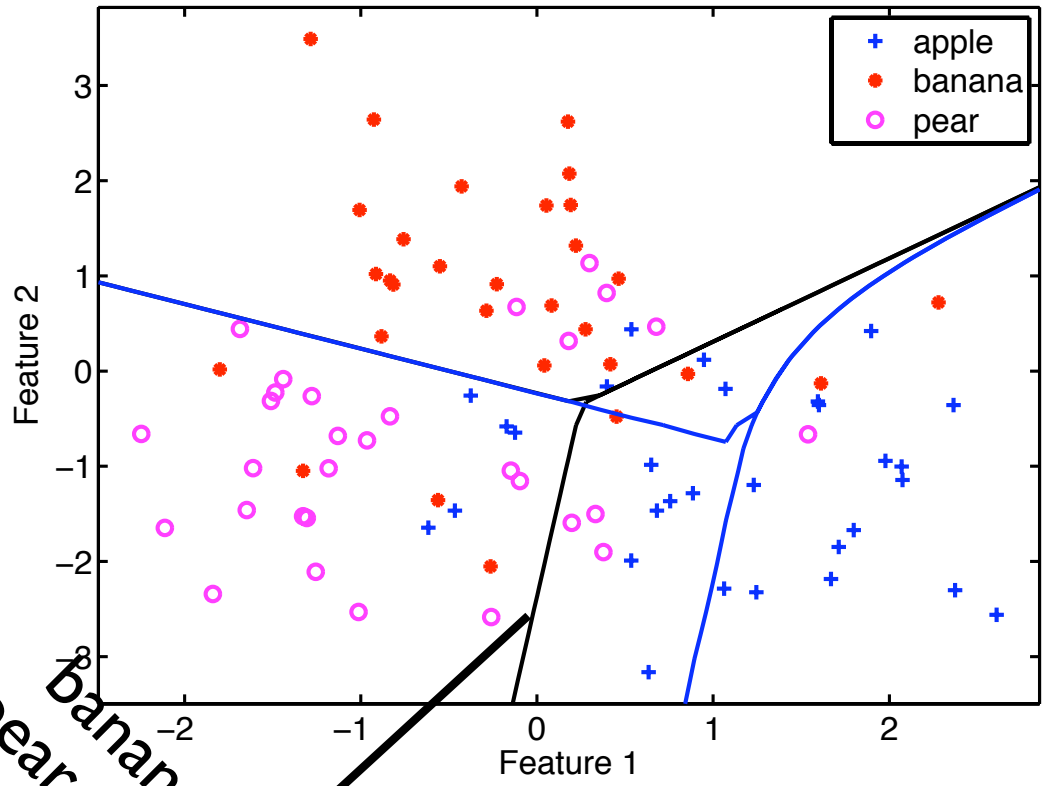
- We minimize the risk when we define the regions Ω_i are chosen such that each of the integrals are as small as possible:

$$r = \sum_{i=1}^{C} r^i = \sum_{i=1}^{C} \int_{\Omega_i} \sum_{j=1}^{C} \lambda_{ji} p(\omega_j|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

- So make **x** part of Ω_i if:

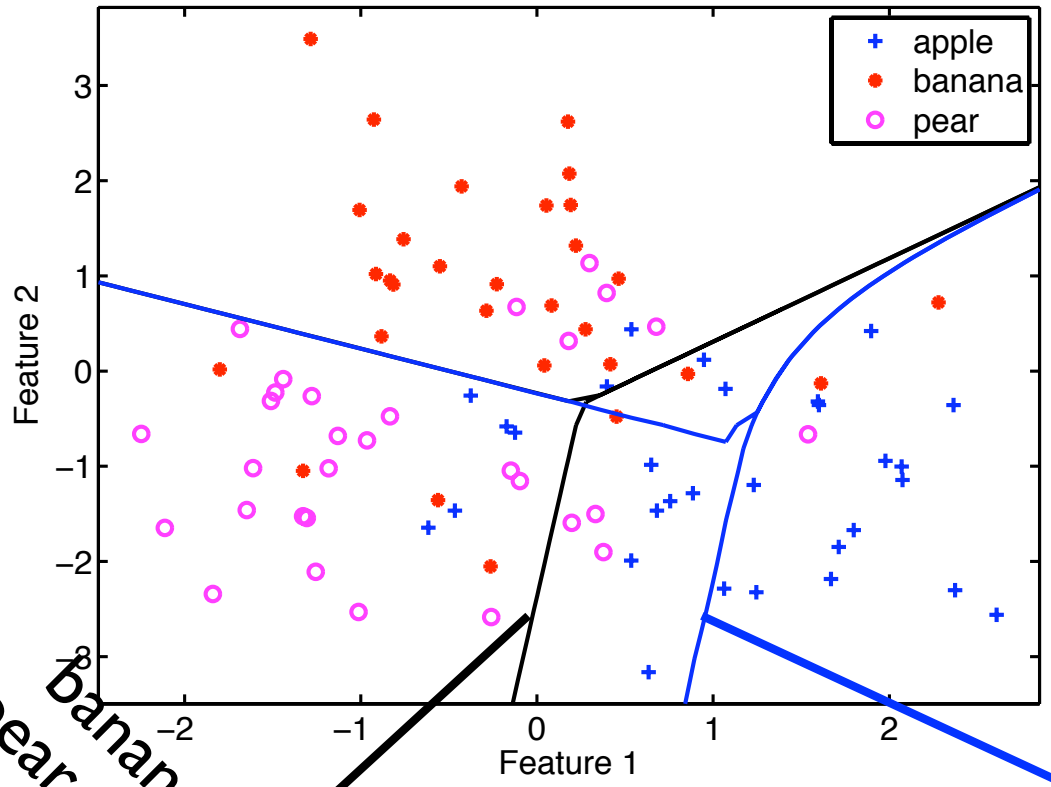
$$\sum_{j=1}^{C} \lambda_{ji} p(\omega_j|\mathbf{x}) \leq \sum_{j=1}^{C} \lambda_{jk} p(\omega_j|\mathbf{x}) \quad k = 1, .., C$$

Example cost



True \ classified	apple	pear	banana
apple	0	1	1
pear	1	0	1
banana	1	1	0

Example cost



$$\lambda_{2,1} = 9$$

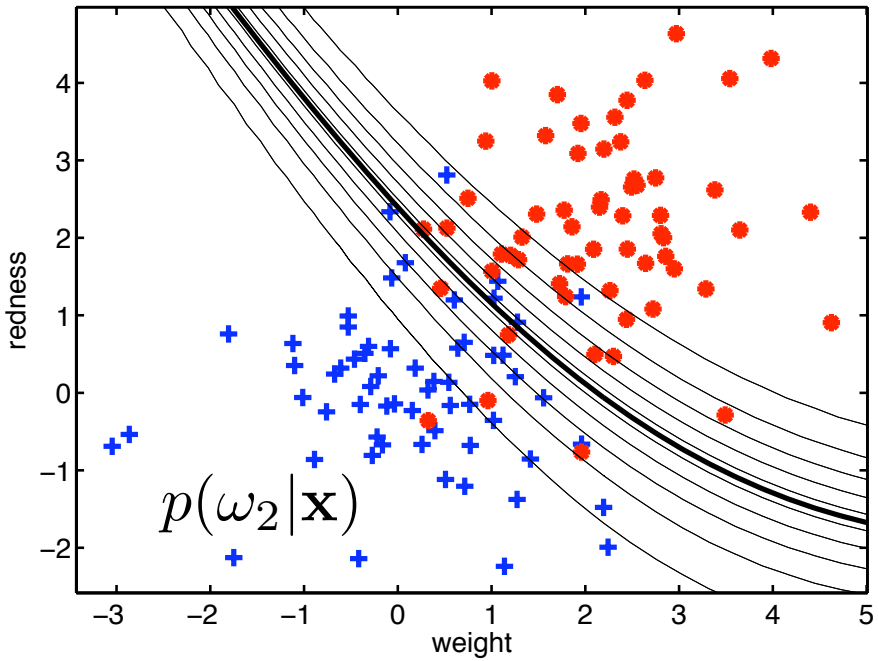
predicted

True	apple	pear	banana
apple	0	1	1
pear	1	0	1
banana	1	1	0

	apple	pear	banana
apple	0	1	1
pear	9	0	1
banana	1	1	0

Questions for you...

- True or not: when you can estimate the posterior probabilities better, you can decrease the Bayes' error.
- Give an estimate of the classification error for this data:
- Will this estimate of the error be larger, equal of smaller than the Bayes' error?
- What happens when $p(\omega_1) = 2p(\omega_2)$?



Bayes' theorem

- In many cases the posterior is hard to estimate
- Often a functional form of the class distributions can be assumed
- Use Bayes' theorem to rewrite one into the other:

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}$$

class (conditional) distribution	$p(\mathbf{x}\|y)$
class prior	$p(y)$
(unconditional) data distribution	$p(\mathbf{x})$