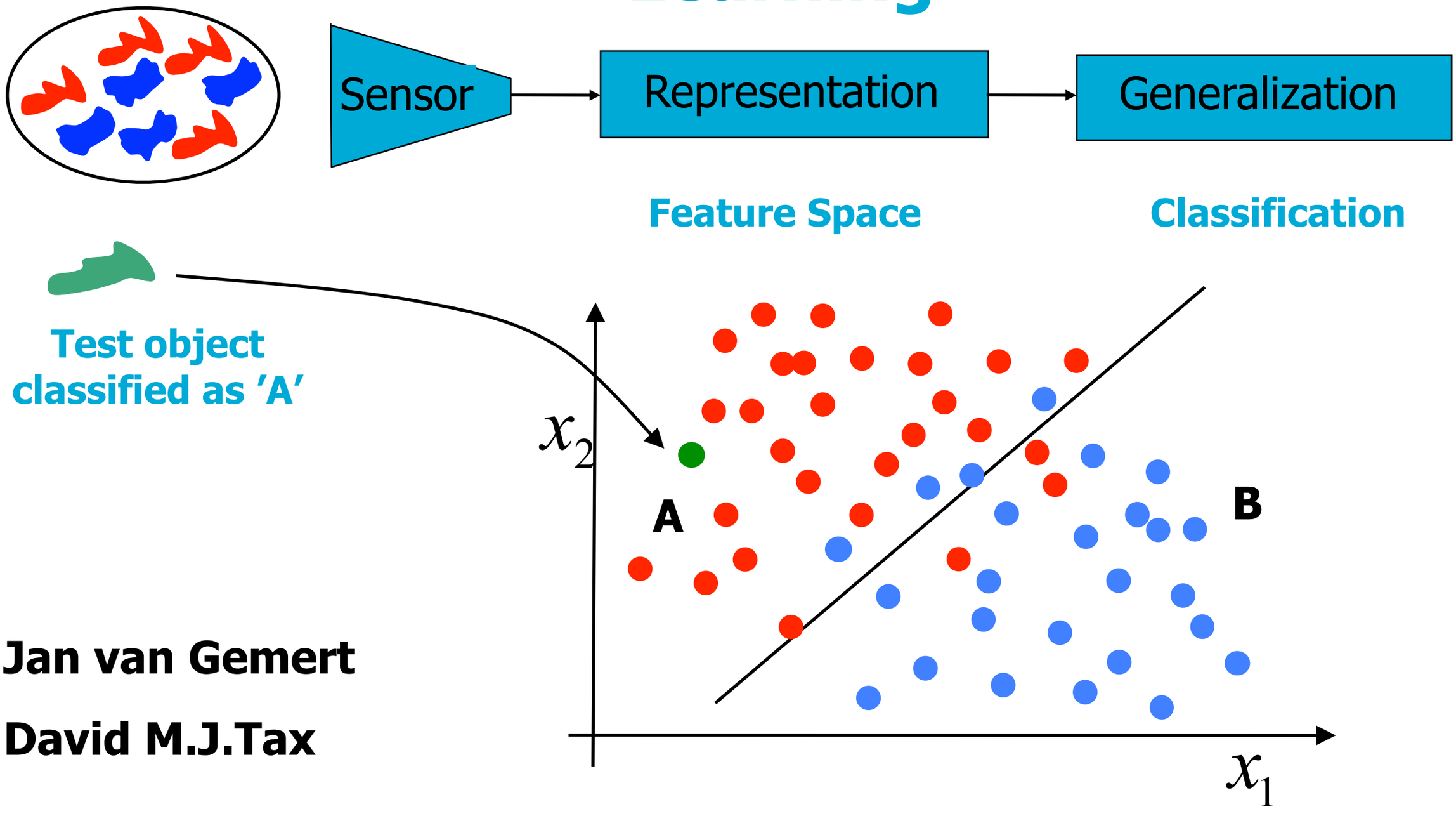
DSAIT 4005 Machine and Deep Learning



Jan van Gemert

David M.J.Tax

Pattern recognition: find the cat

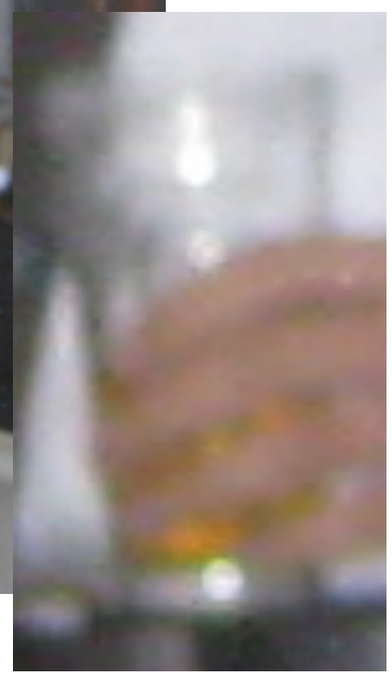
lft

Pattern recognition: find the cat

lft

Recognition problem

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



TU Delft

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**

Machine Learning

Unsupervised Learning

- Big Data Visualisation
- Structure Discovery
- Dimension Reduction
- Feature Extraction
- 2D to 3D Modelling
- Music Generation
- Generative Networks
- Pattern Modelling
- Image Generation
- Clustering
- Recommender Systems
- Customer Segmentati
- Targeted Marketing

Semi-Supervised Learning

- Text Classification
- Lane Finding

Reinforcement Learning

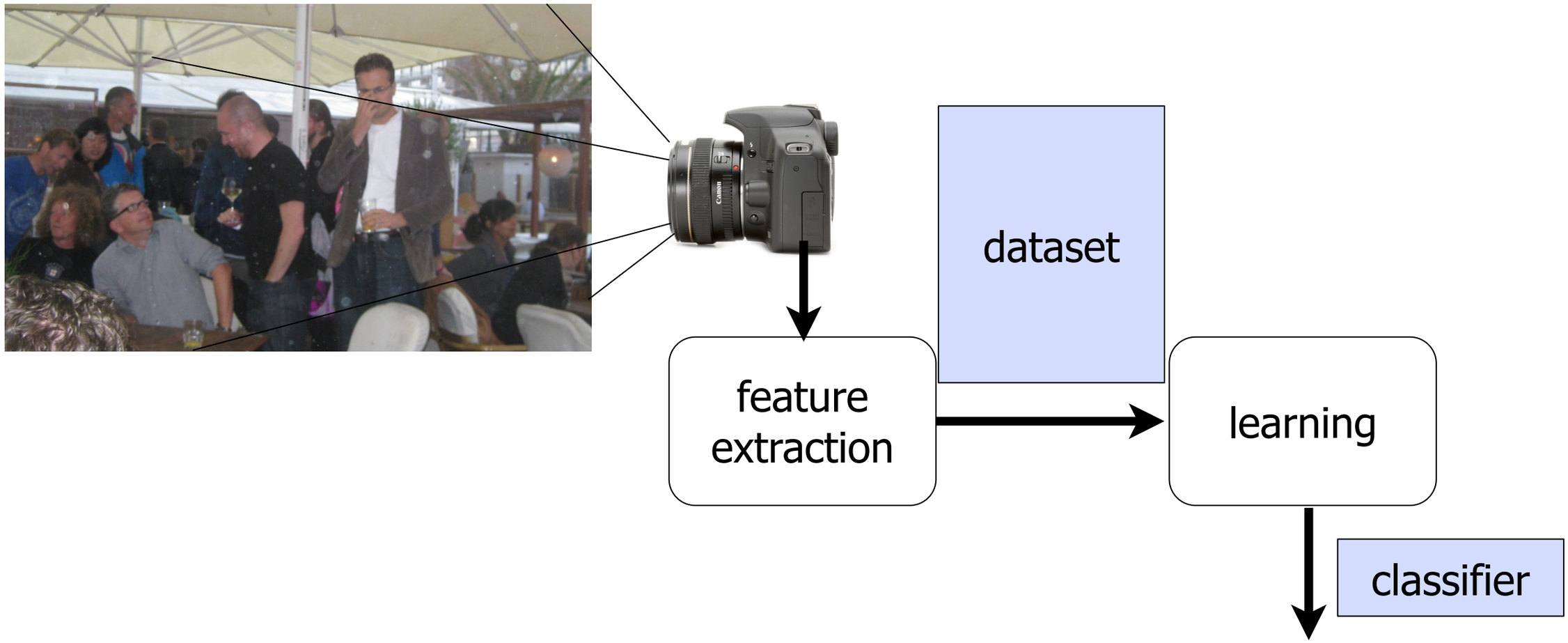
- Real-time Decisions
- Game AI
- Robot Navigation
- Skill Acquisition
- Learning Tasks

Supervised Learning

- Fraud Detection
- Gender Detection
- Classification
- Image Classification
- Medical Diagnosis
- Weather Forecasting
- Regression
- Advertising Reach Prediction
- Sales Growth Prediction
- Market Forecasting
- Estimating Life Expectancy

By Benjamin Timmermans. http://btimmermans.com

Pattern recognition pipeline



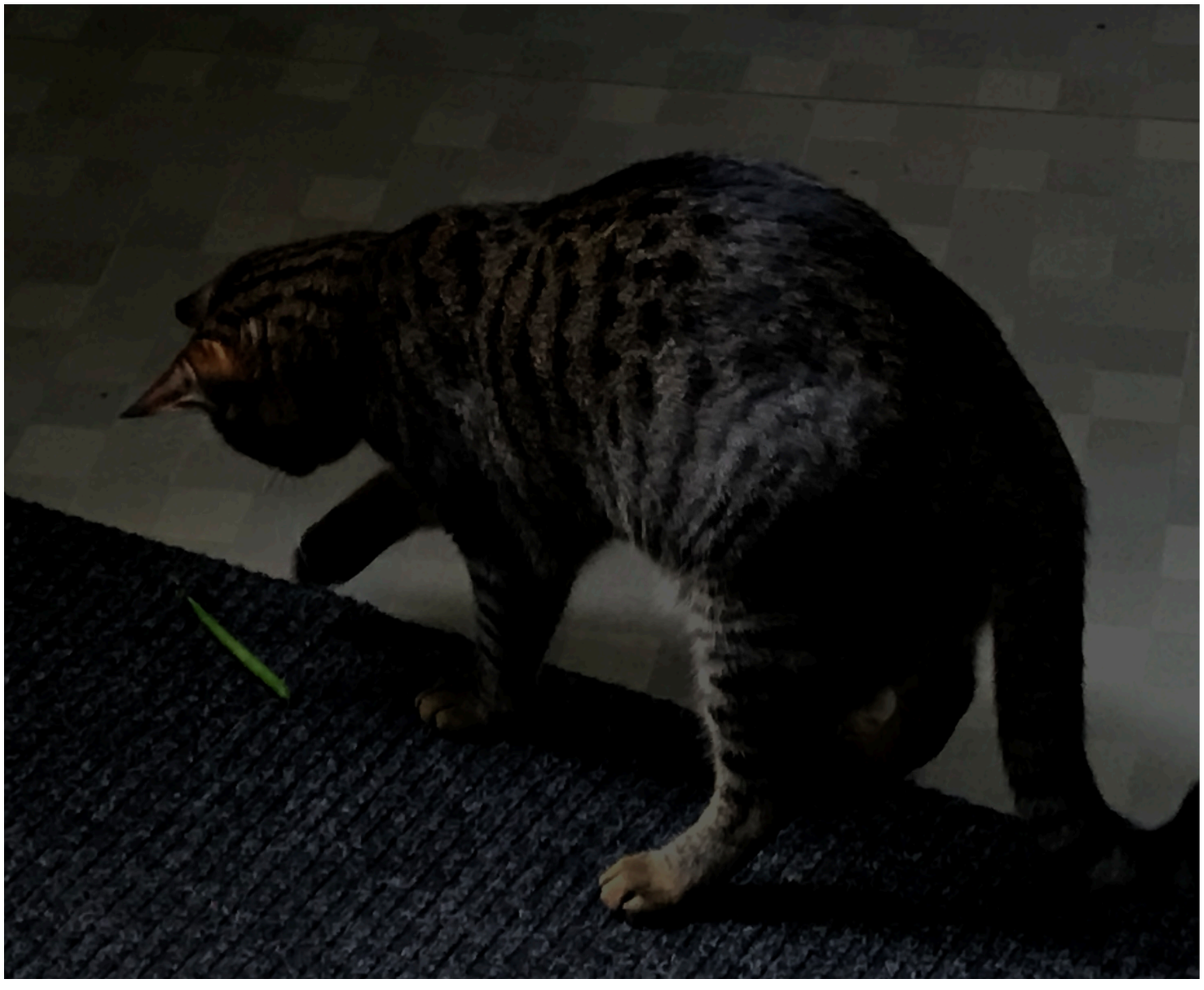
- From measurements, extract features, define labels, and train a classifier

Classification: What is the label?

What is the label of this object?

Cat? Why not green bean?

What is the label of this object?

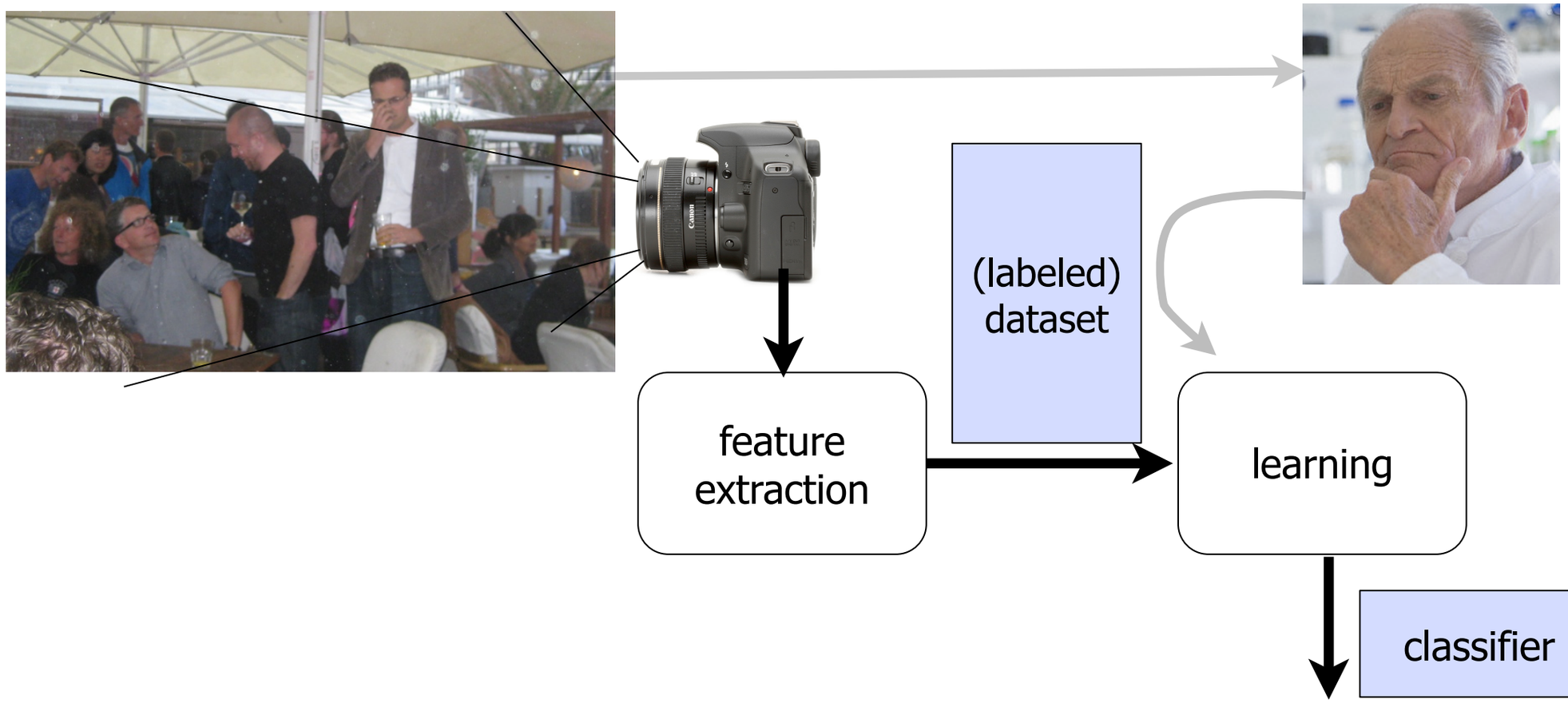


- And now?

What is the label of this object?

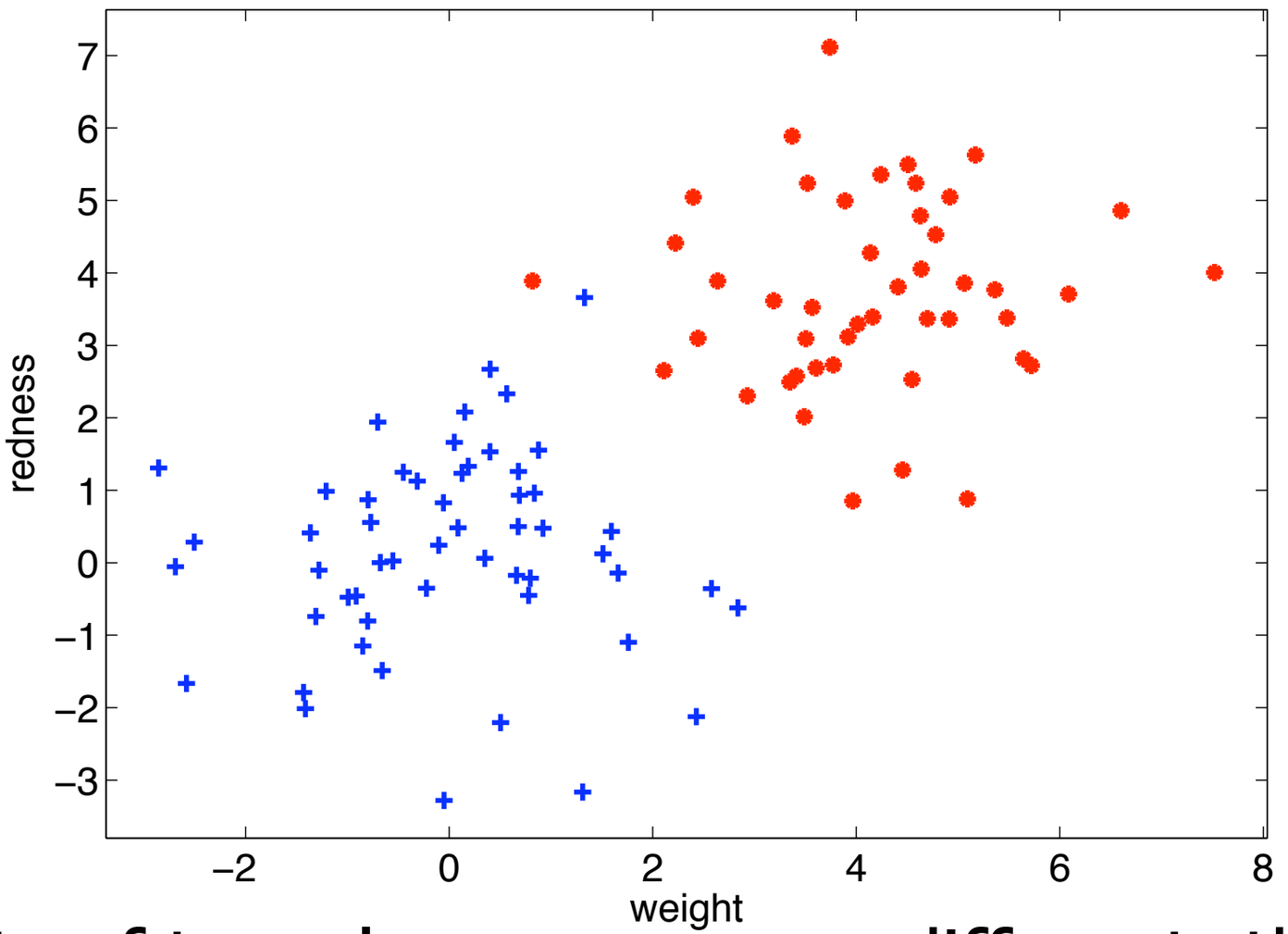
- And now?
- Is label generated from input image, or is there an external source?
- Is all relevant information available in the image for classification?

Pattern recognition pipeline



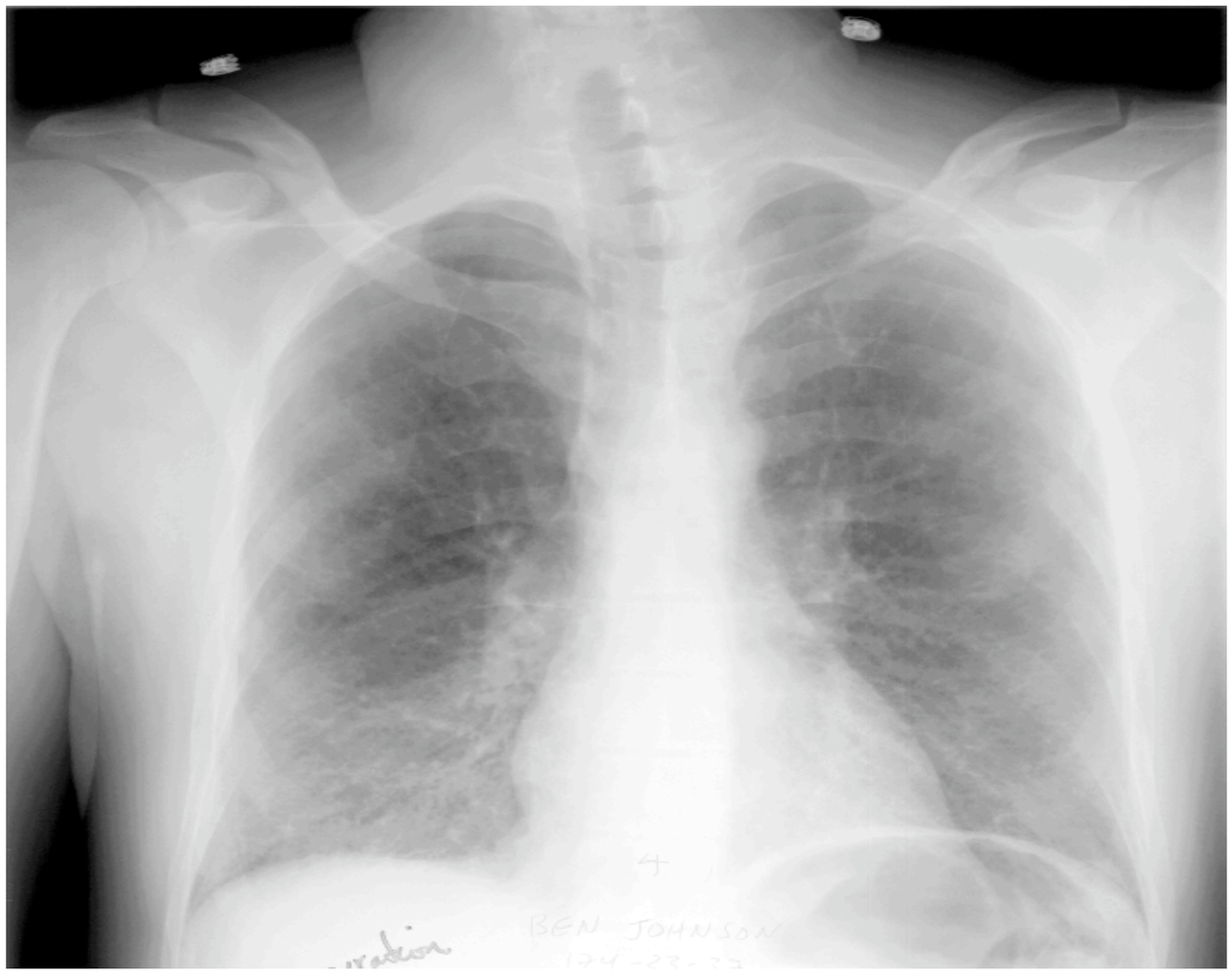
- Experts are needed for providing labels/targets

(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

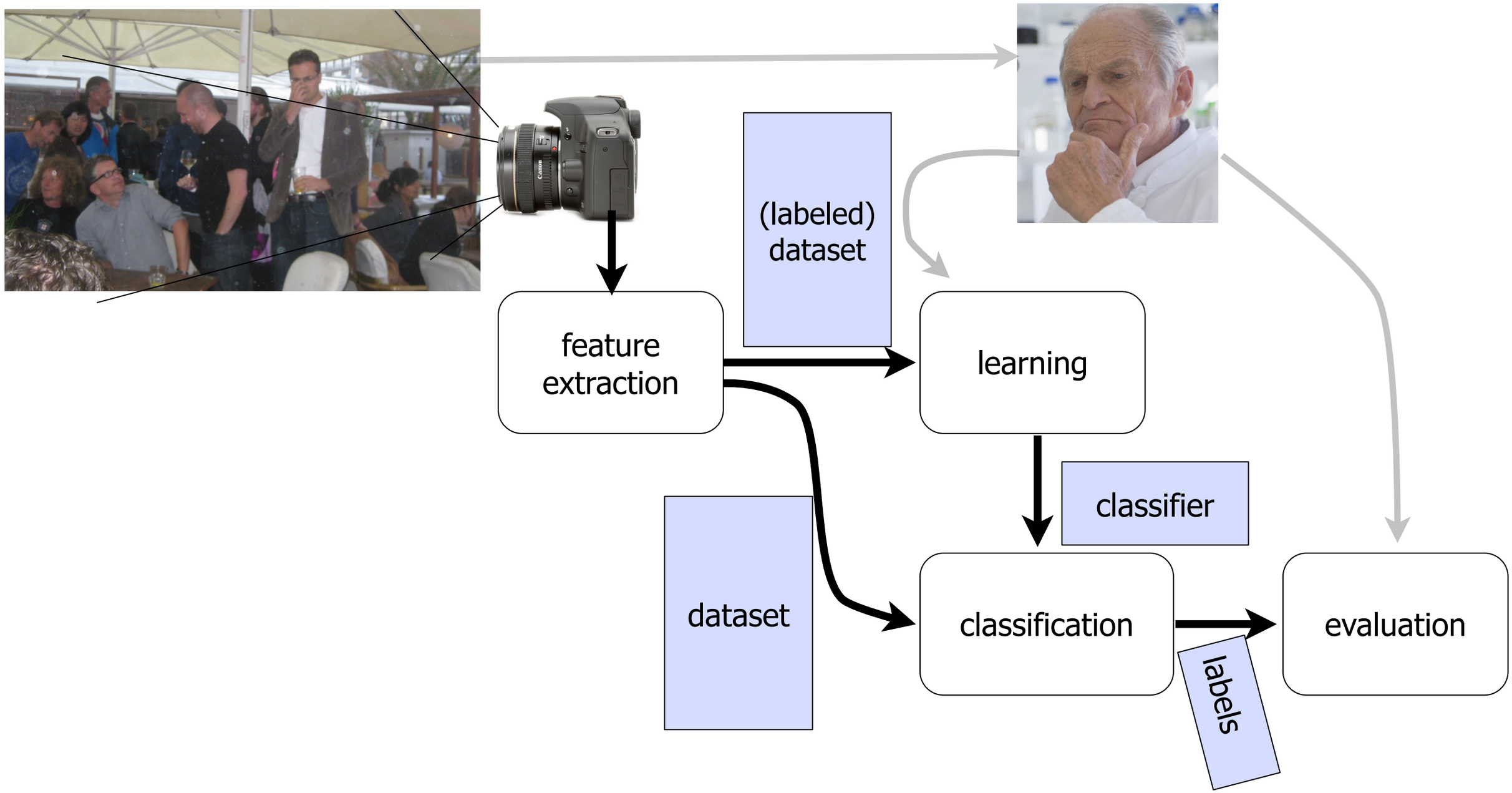
Again: what is the label of this object?

- ?

Again: what is the label of this object?

- Cat? Green bean?
- How do you know? You're overfitting to your training data!
- Not fair!

Pattern recognition pipeline



- Check with independent test data how well it works!

Admin: 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
- Look at Brightspace!

Admin: Schedule of the course

- Week 1: Basic Machine Learning: classification
- Week 2: Multi-layer perceptron
- Week 3: Discriminative classifiers
- Week 4: Optimization and Evaluation
- Week 5: Complexity and Regularization
- Week 6: CNNs and Recurrent Neural Networks
- Week 7: Self-attention and Unsupervised Learning
- Week 8: Foundation models and Q&A

- Exam!

Admin: Lectures, exercises and questions

- Lectures are not mandatory
- (Non-mandatory) exercises in Python
- All info on Brightspace (check regularly!)
- Two hours per week, TAs present to answer questions
- Brightspace has all information (in particular 'Overview')
- Questions? Use Brightspace Forum
- For personal things: use

  `MDL-course-cs-ewi@tudelft.nl`

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?)

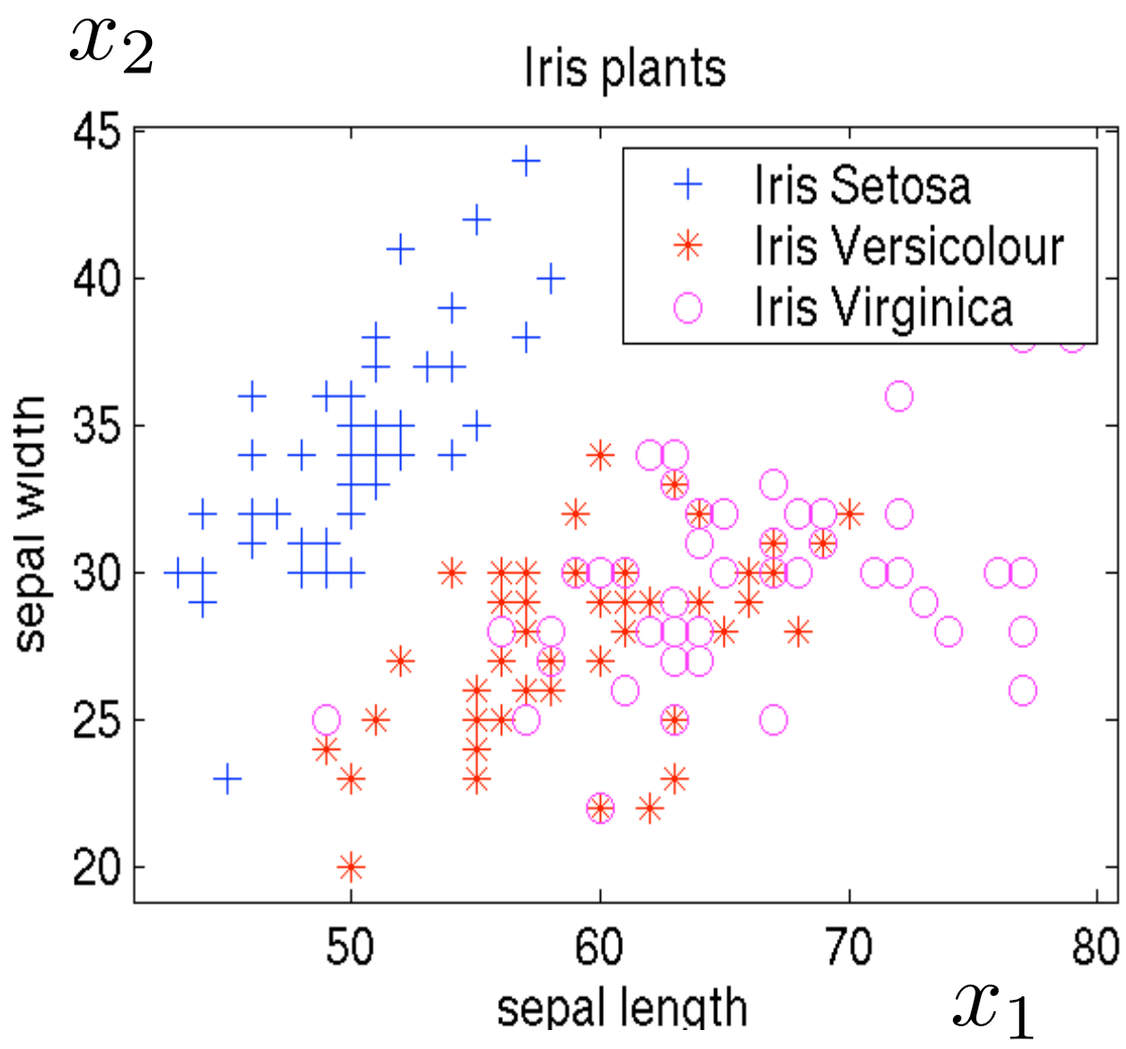
Objects in feature space

- 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

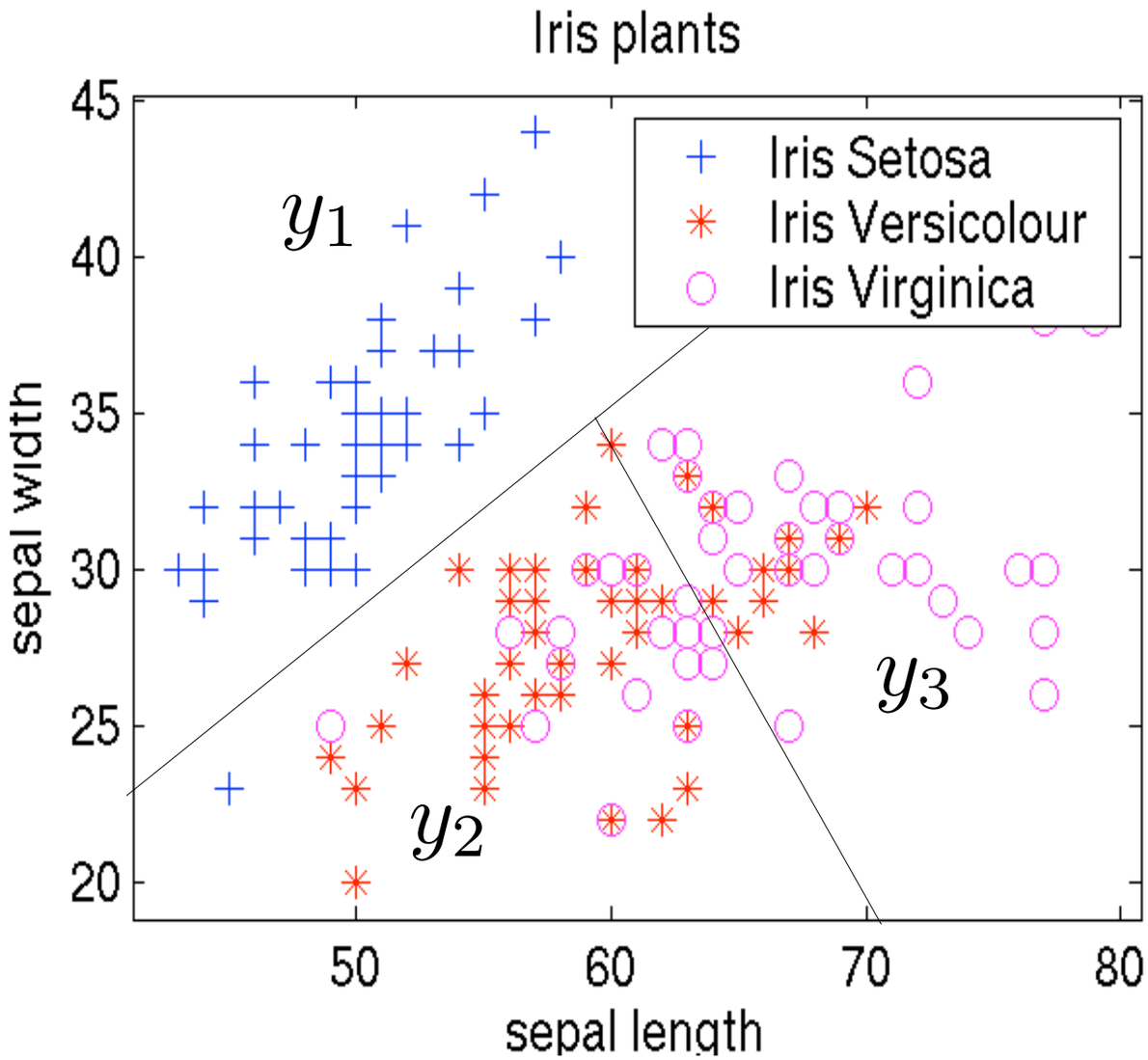
$$p(\mathbf{x}, y)$$



TU Delft

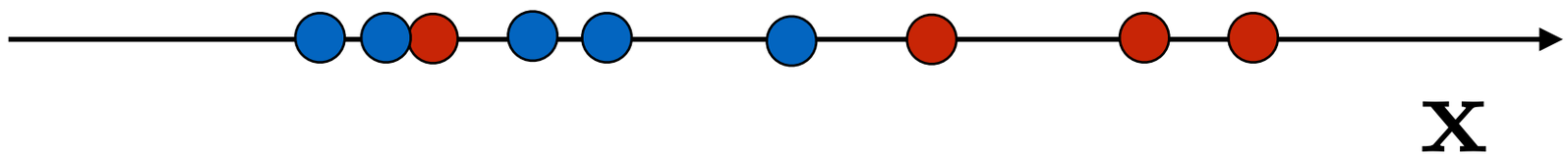
Classification

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



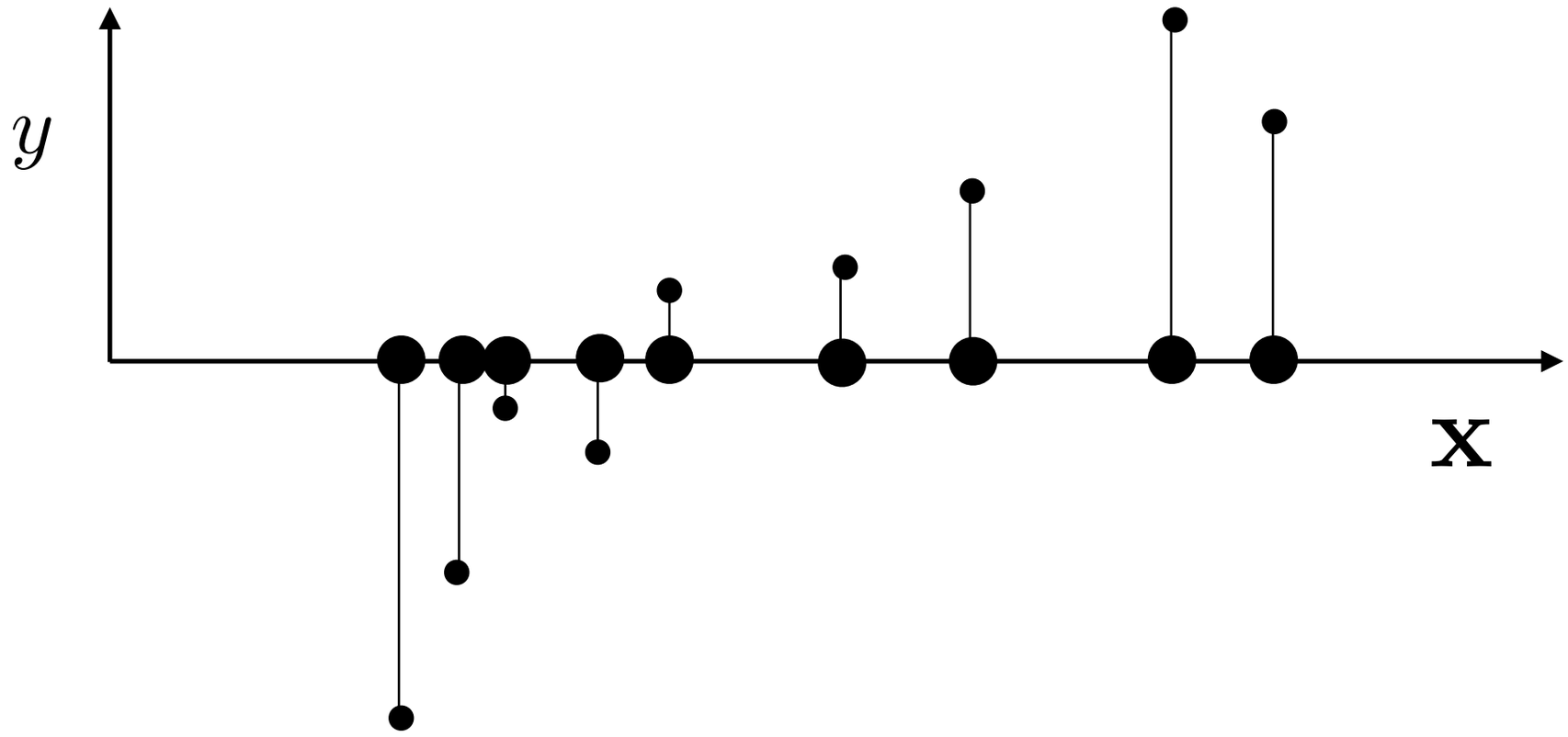
TU Delft

Also for regression, clustering, …



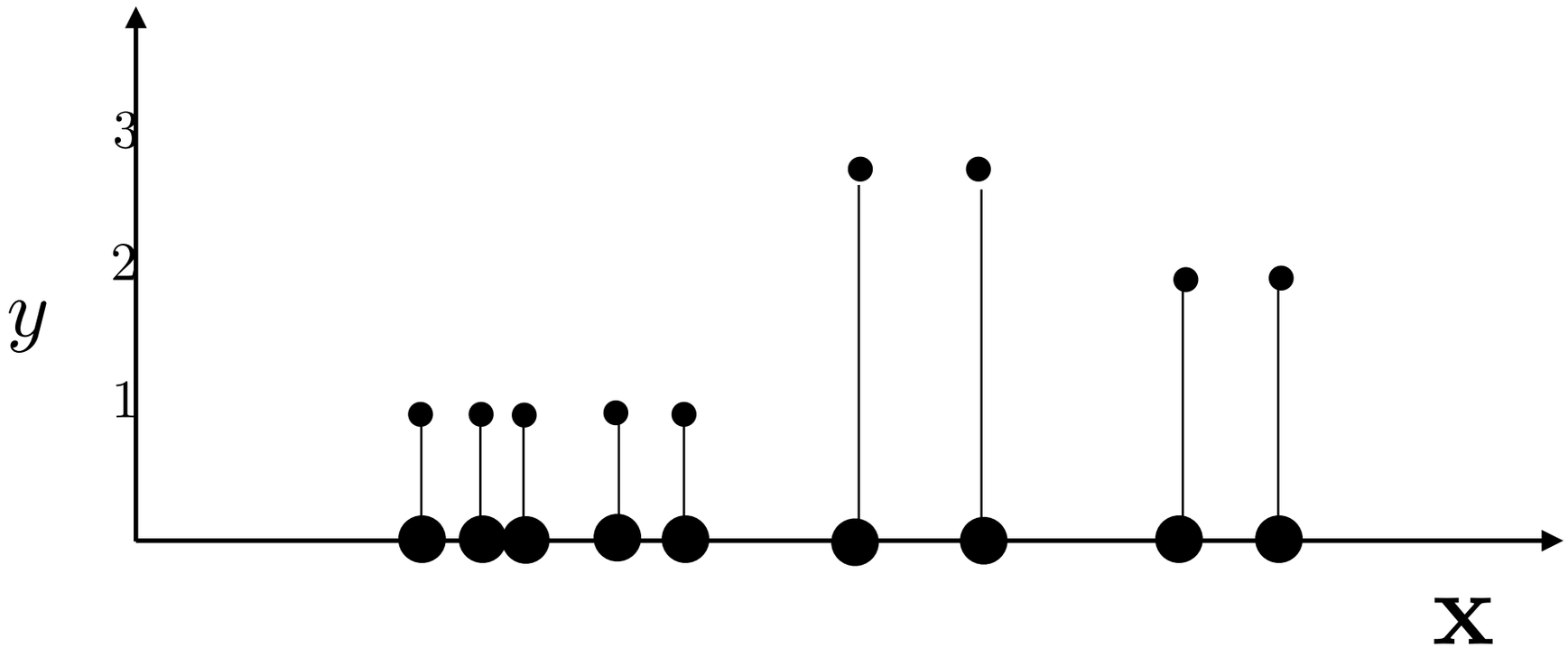
- Instead of class labels, real values: regression
- Instead of supervised labels, no labels: clustering

Also for regression, clustering, …



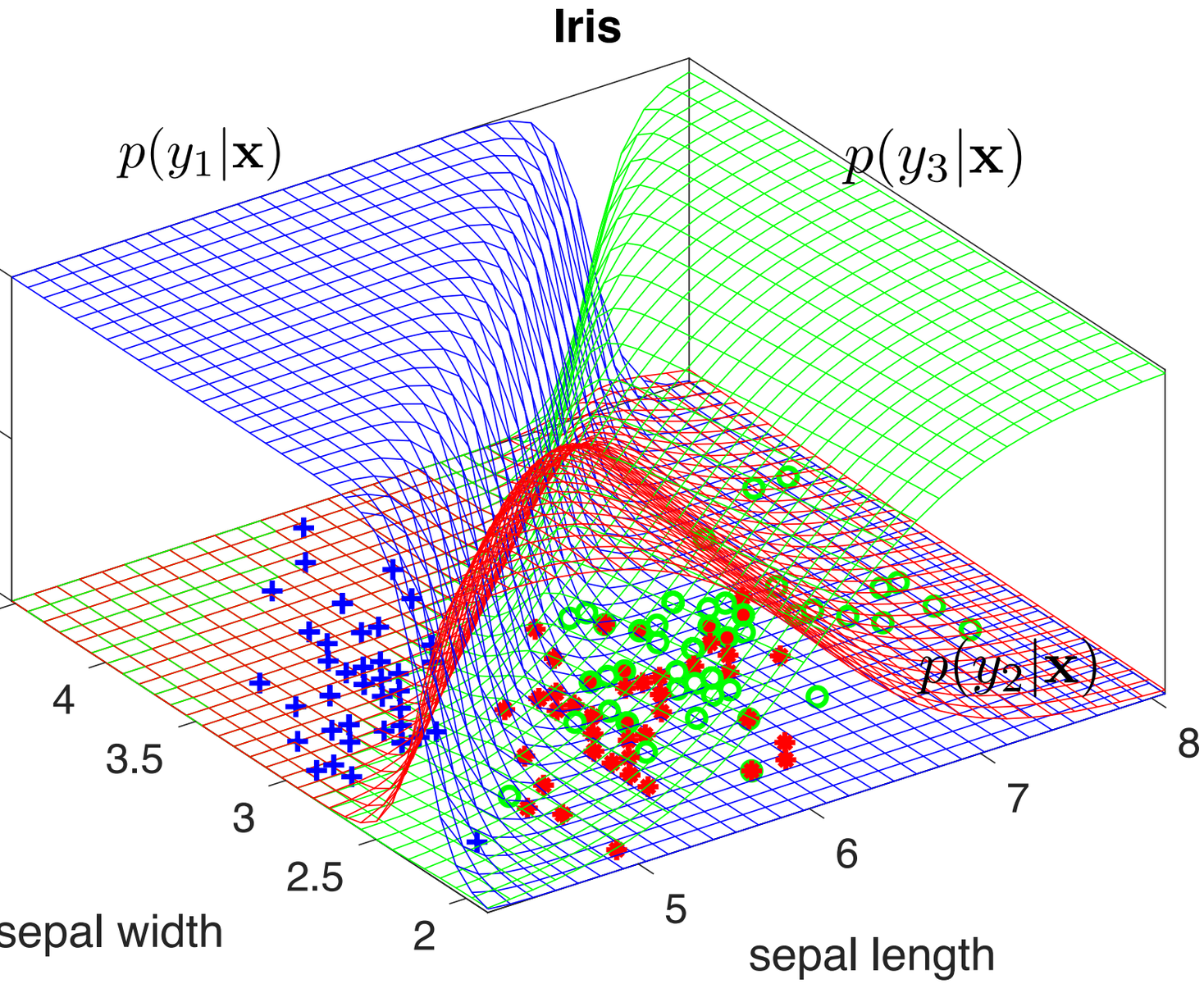
- Instead of class labels, real values: regression
- Instead of supervised labels, no labels: clustering

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

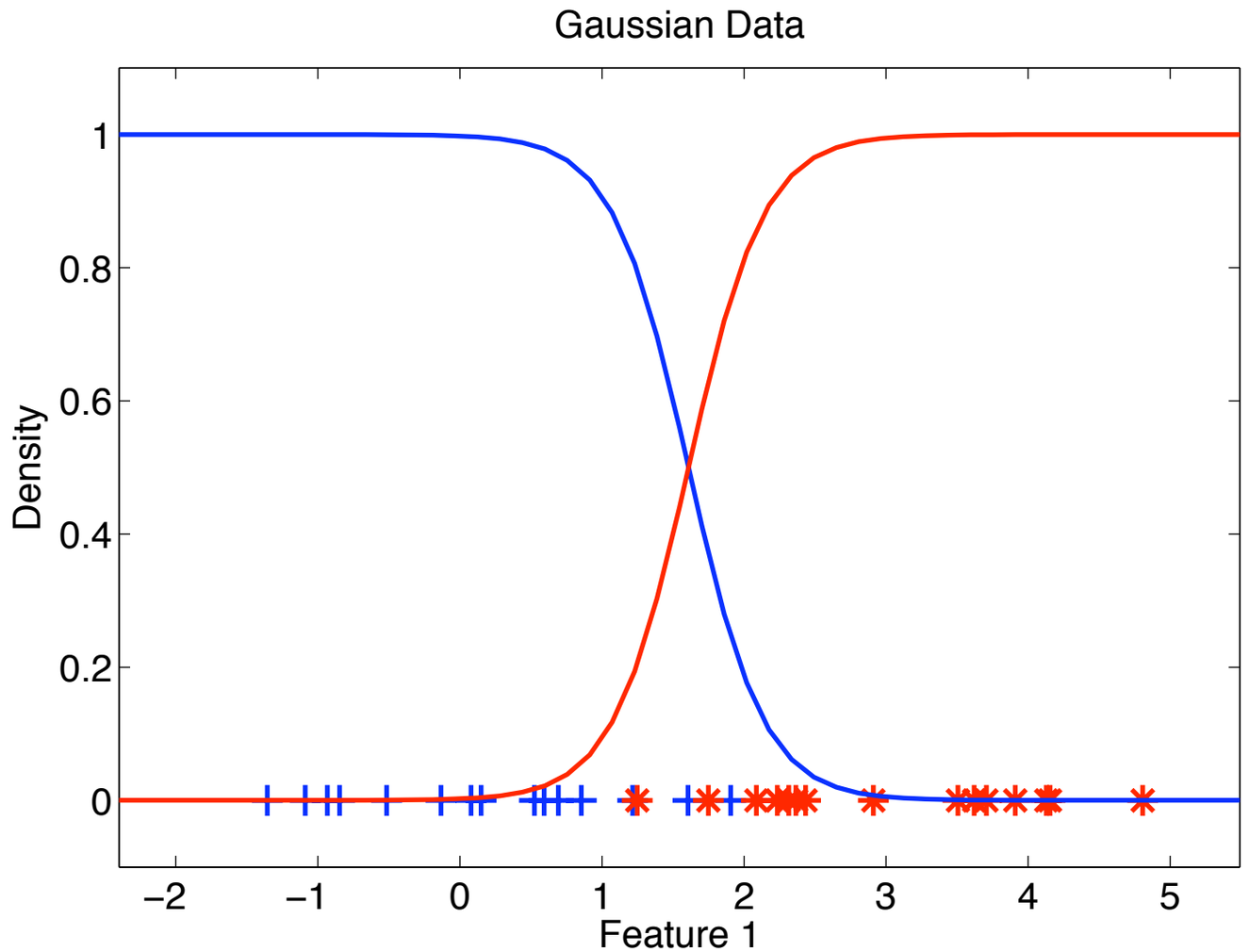
- or $\dfrac{p(y_1|\mathbf{x})}{p(y_2|\mathbf{x})} > 1$

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

Class posterior probability

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

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

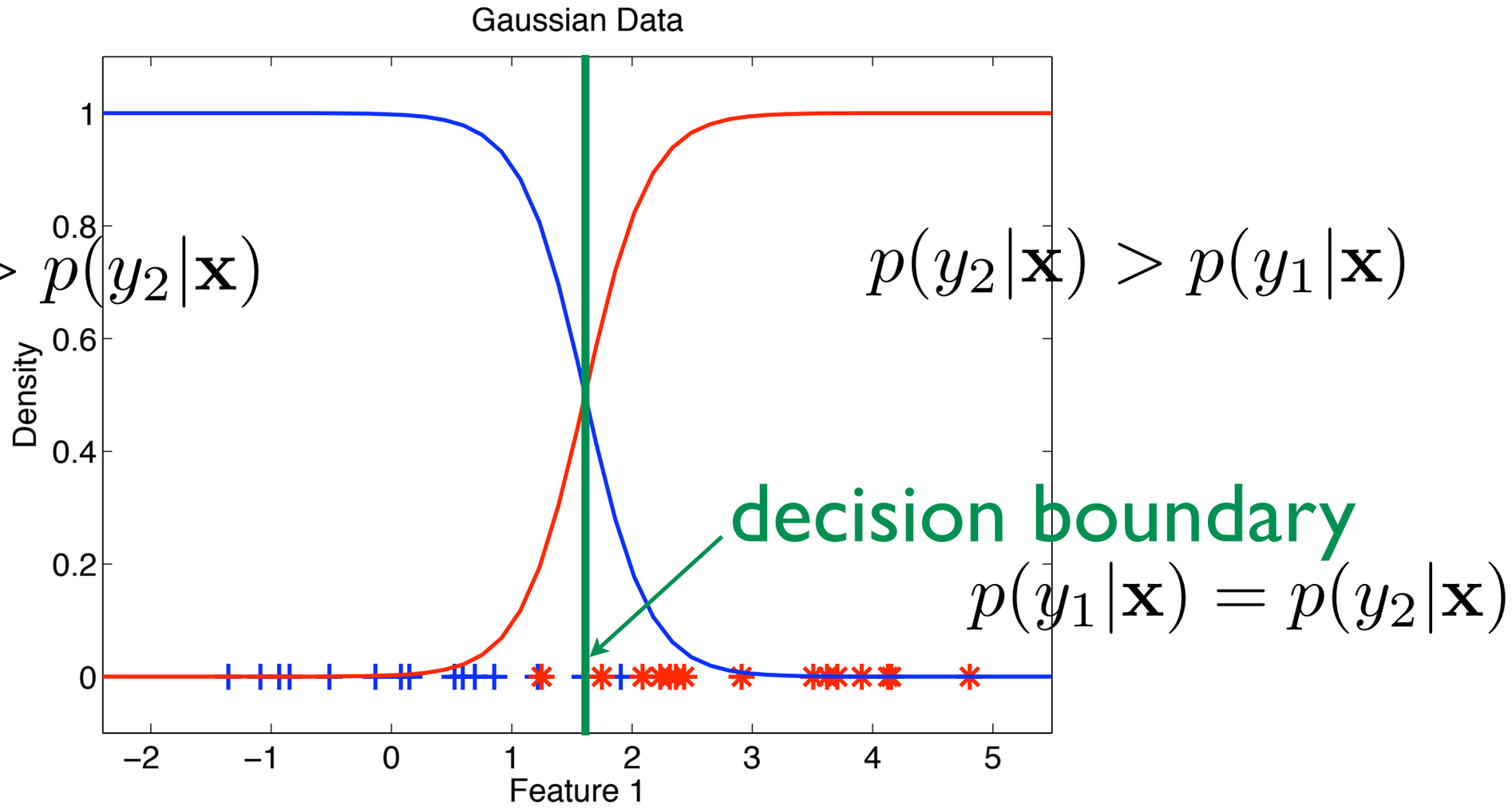


Classify new objects

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

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

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



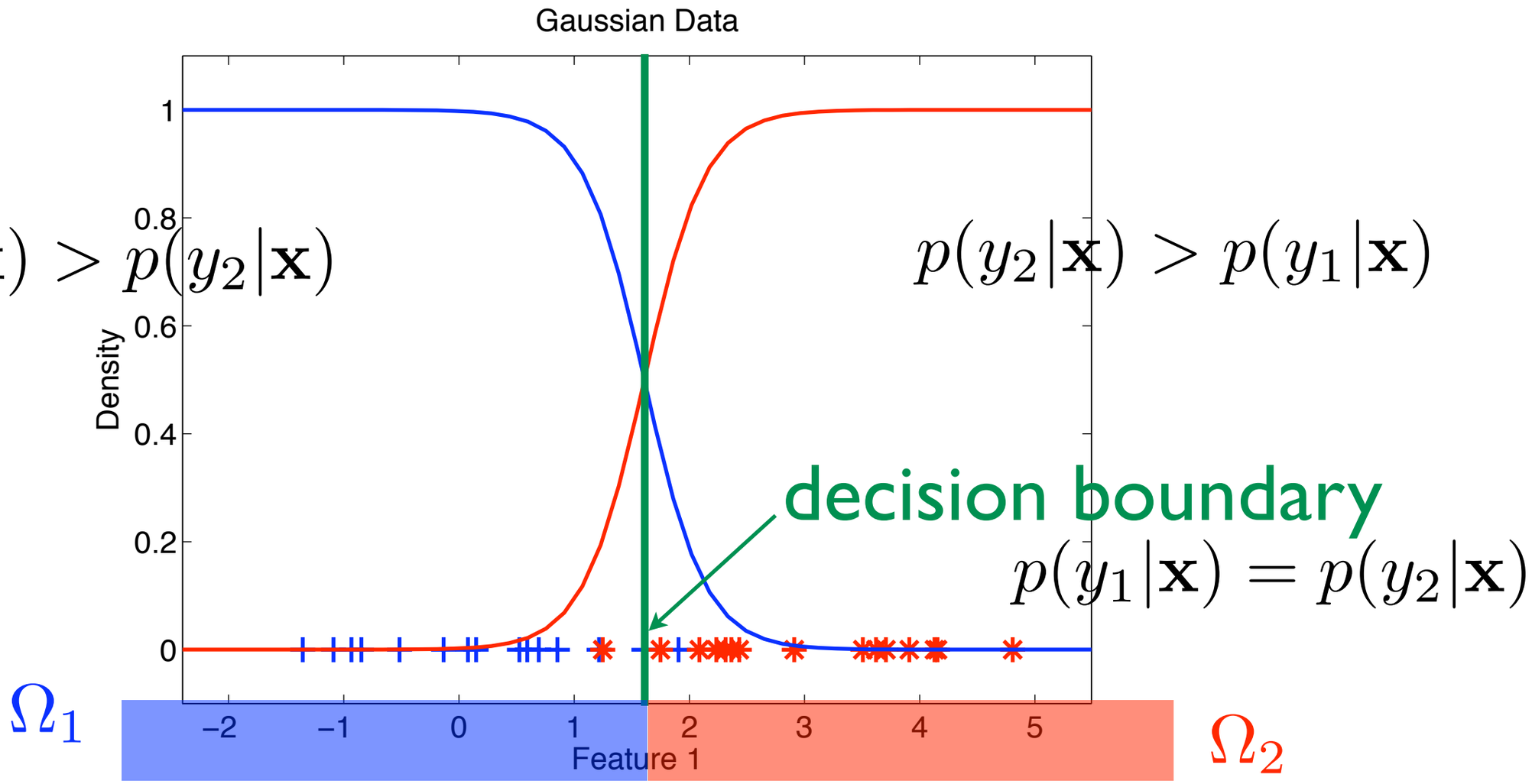
$$p(y_1|\mathbf{x}) = p(y_2|\mathbf{x})$$

Classify new objects

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

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

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



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	$p(\mathbf{x})$

Law of total probability

- Note that you don't have to estimate p(x) separately

- You can derive it from the individual class conditional probabilities:

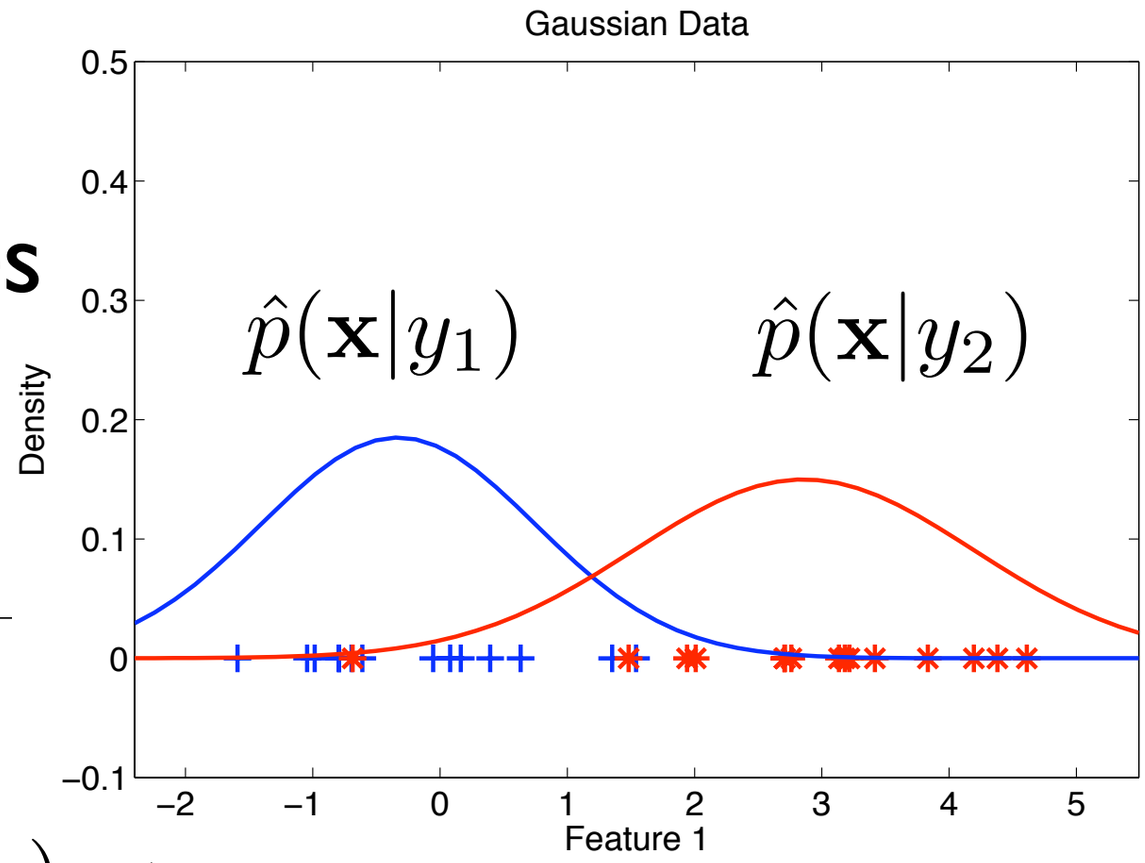
$$p(\mathbf{x}) = \sum_{i=1}^{C} p(\mathbf{x}|y_i)p(y_i)$$

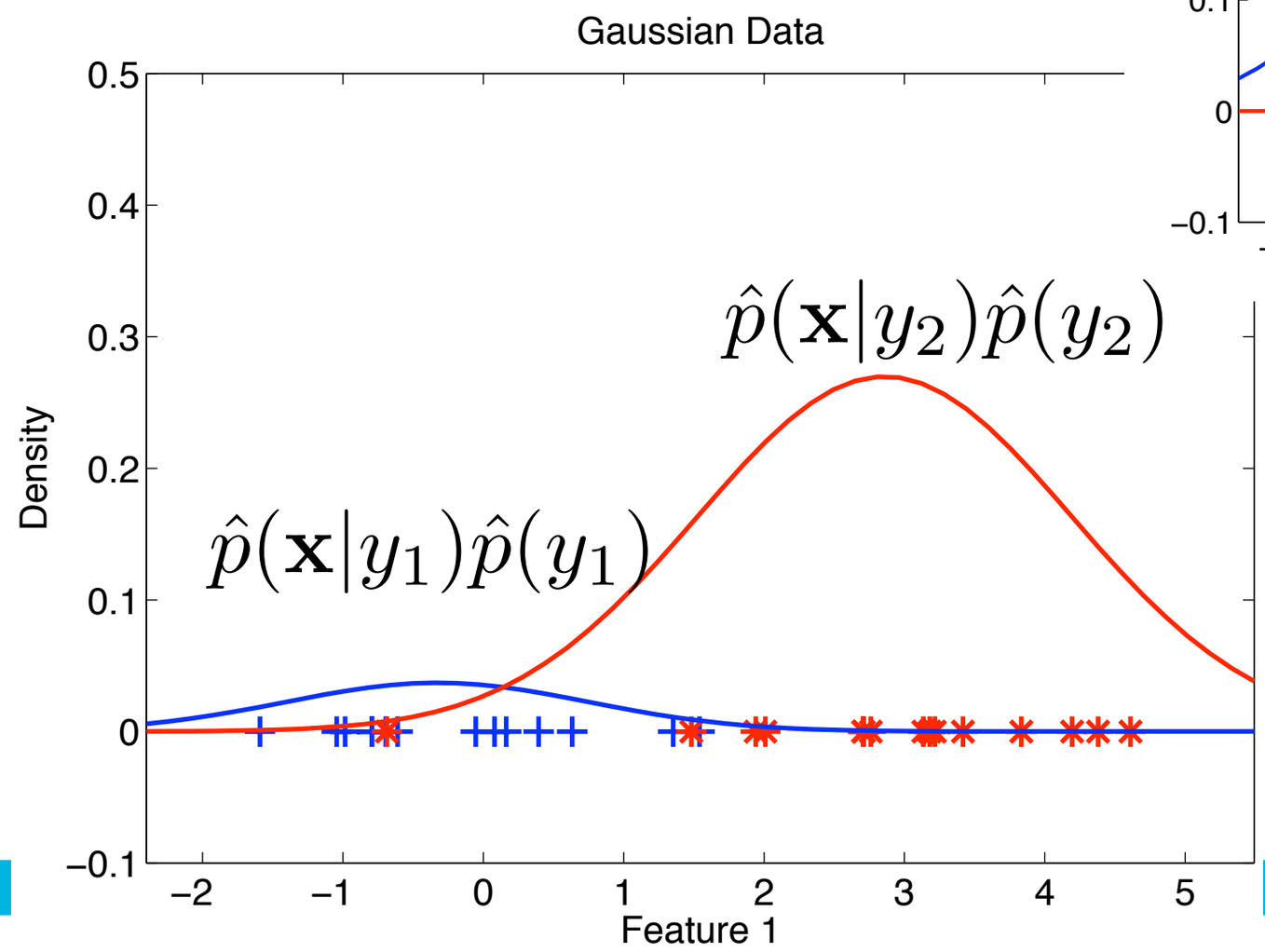
- For two-class classification problems:

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

Bayes' rule

1. Estimate the class conditional probabilities



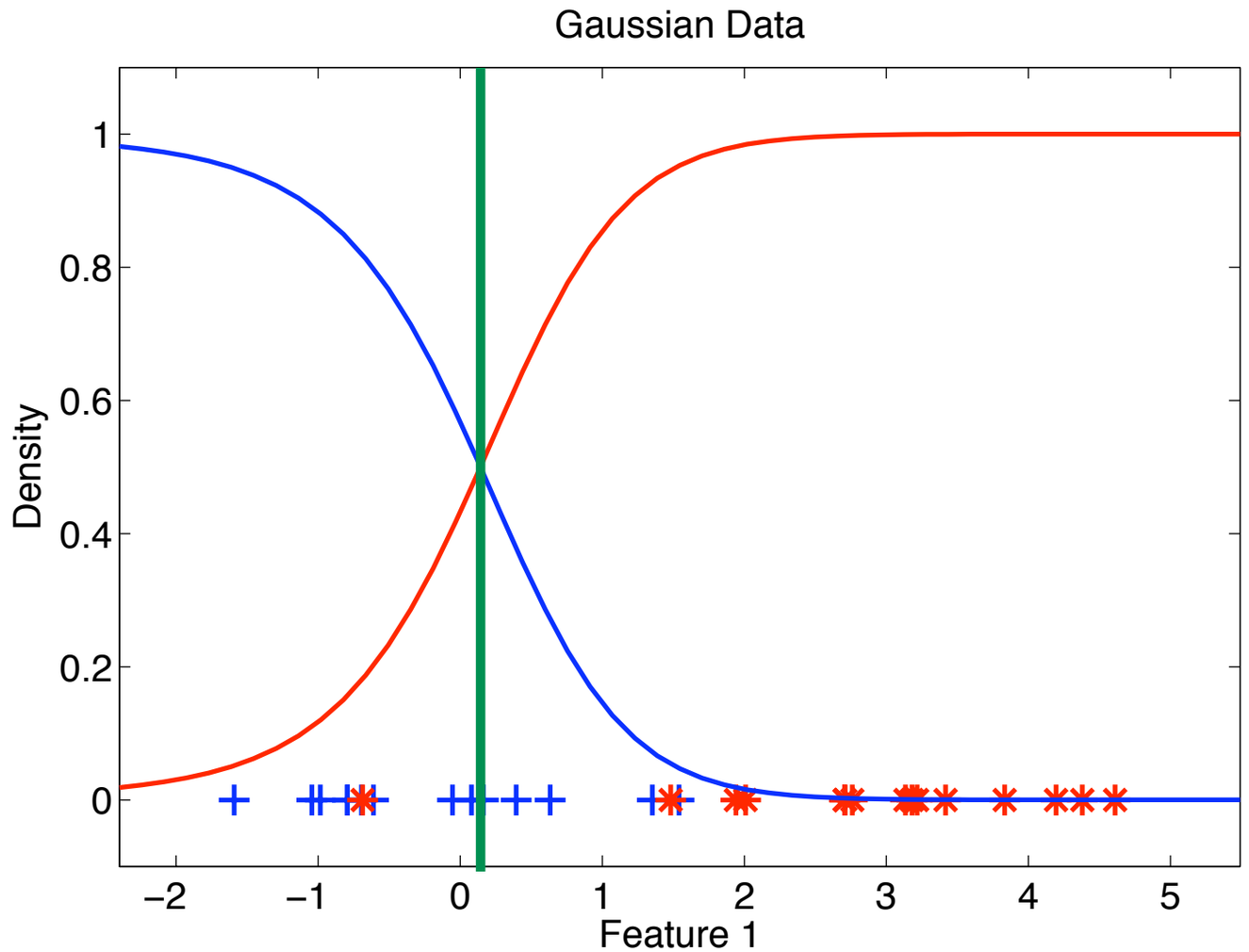


2. Multiply with the class priors

Bayes' rule

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

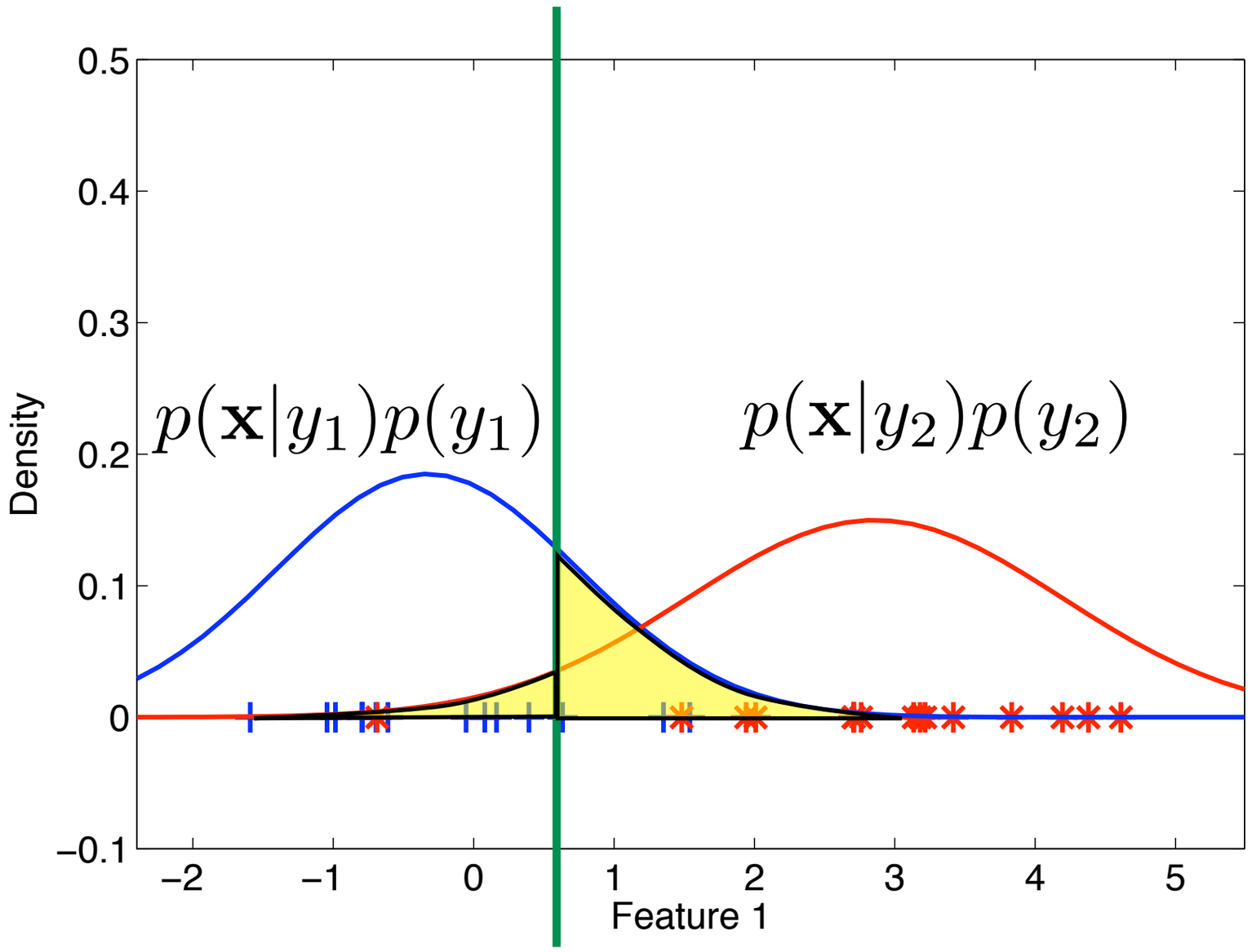
$\hat{p}(y_1|\mathbf{x})$



$\hat{p}(y_2|\mathbf{x})$

Classification error

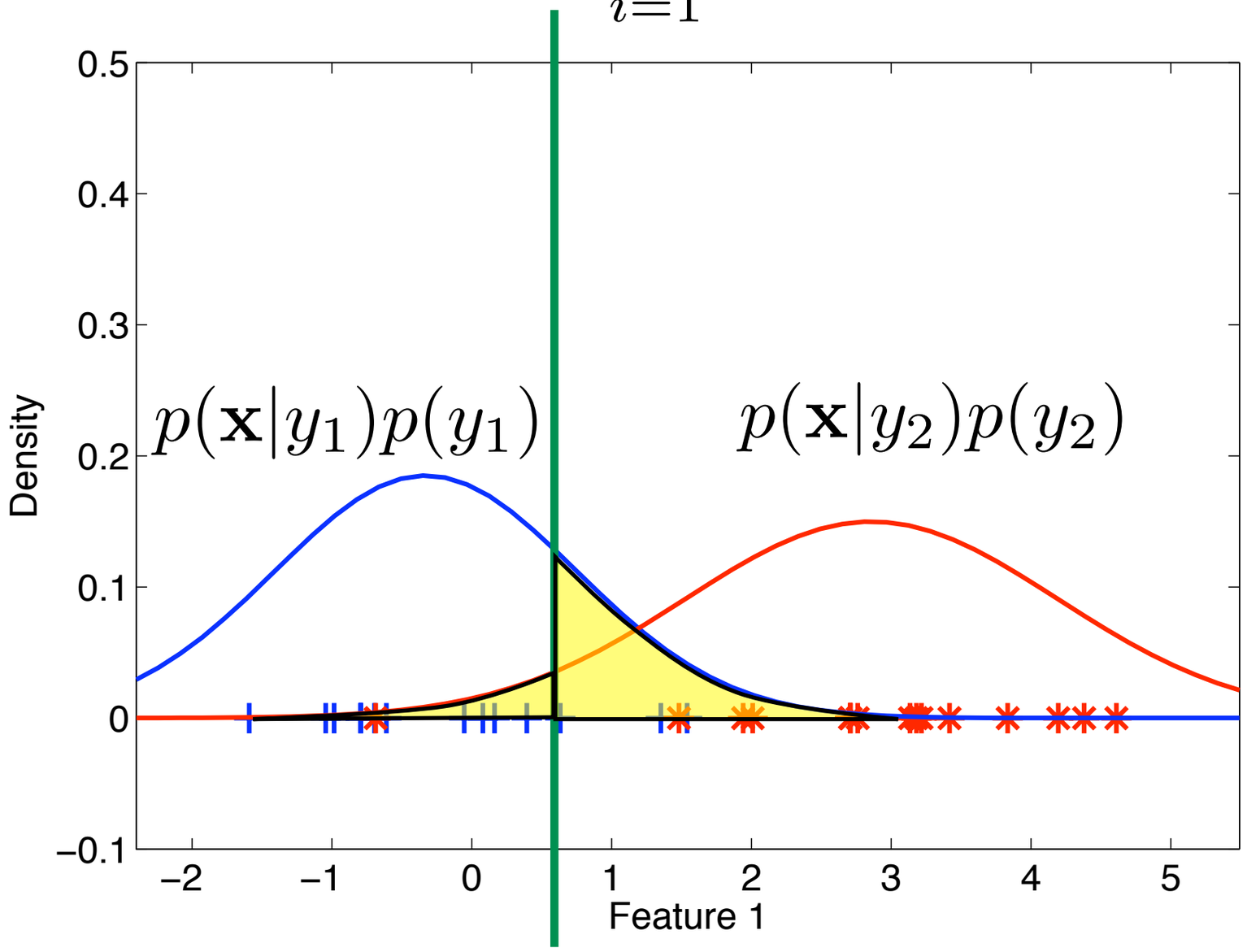
- If I have found a decision boundary, how good is it?



TU Delft

Classification error

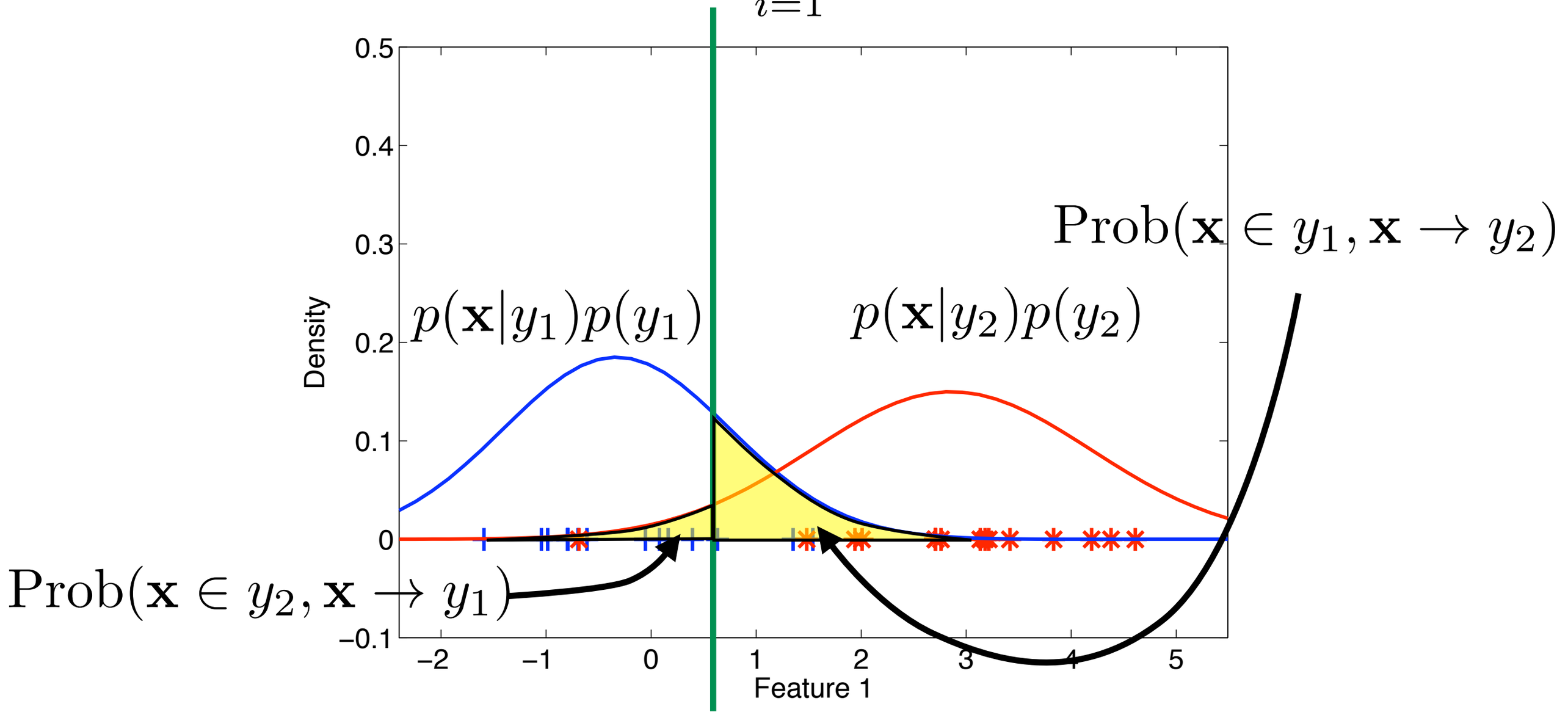
- The error: $$p(\text{error}) = \sum_{i=1}^{C} p(\text{error}|y_i)p(y_i)$$



TU Delft

Classification error

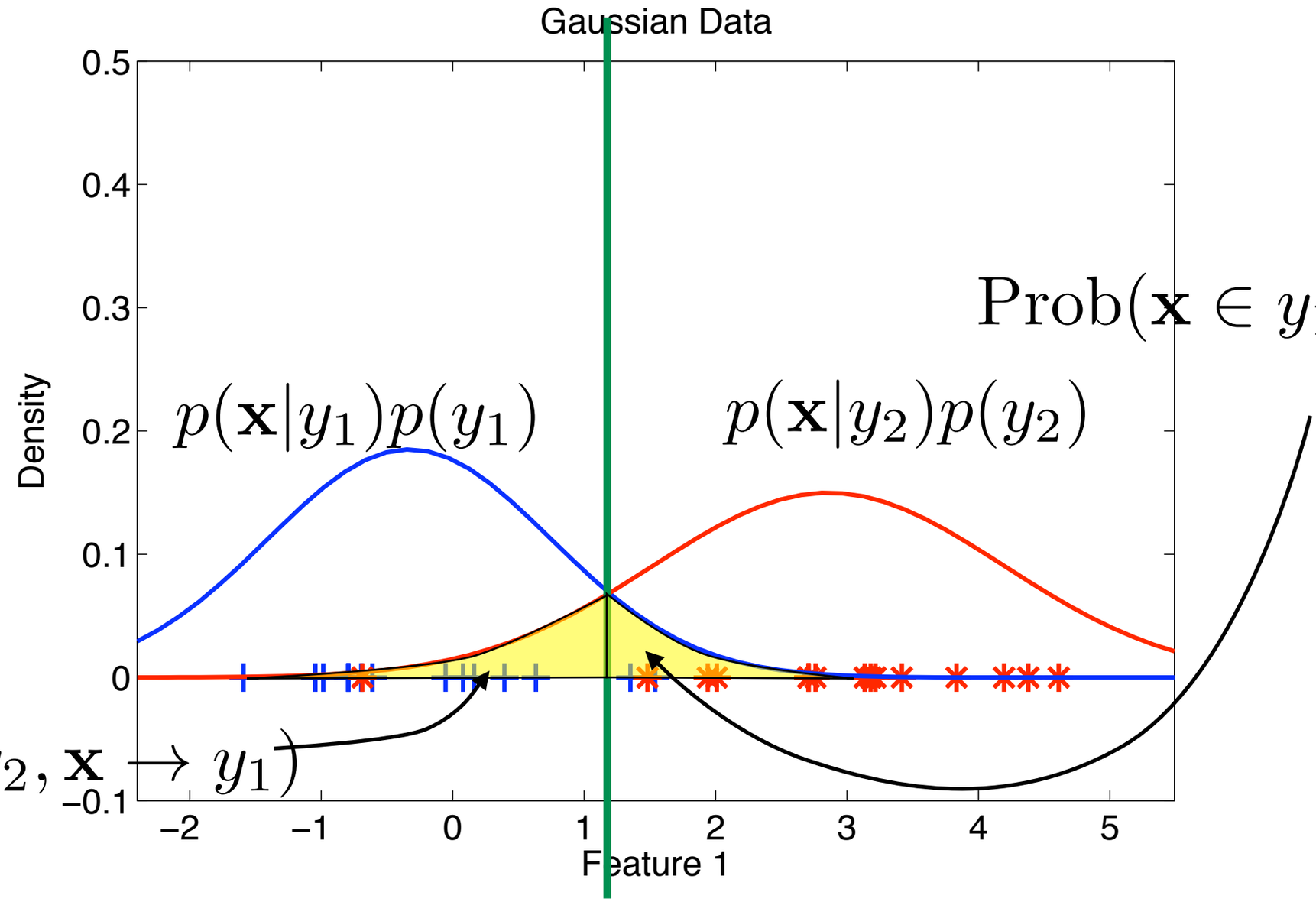
- The error: $$p(\text{error}) = \sum_{i=1}^{C} p(\text{error}|y_i)p(y_i)$$



TU Delft

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

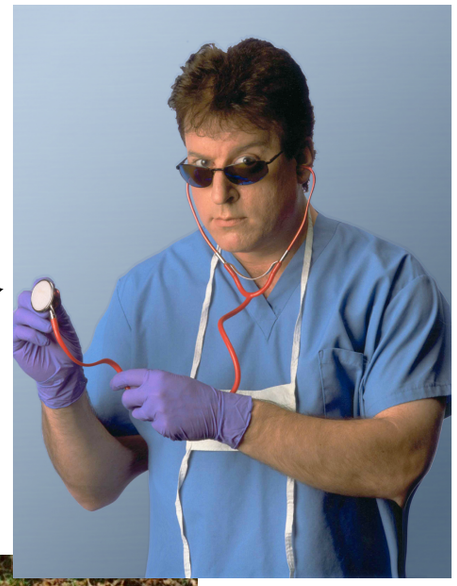
misclassification:
classify 'healthy' to 'ill'

misclassification:
classify 'ill' to 'healthy'

Misclassification cost

- Introduce a loss that measures the cost of assigning an object that came from class y_j to class y_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 :

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

- The average risk over a region:

$$r^i = \int_{\Omega_i} l^i(\mathbf{x}) p(\mathbf{x}) d\mathbf{x}$$

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

- Overall risk:

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

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:

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

- So make **x** part of Ω_i if:

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

Minimum total risk: two classes

- When you predict class y_i for an object of class y_i you would typically define:

$$\lambda_{y_i,y_i} = 0$$

- For a two-class problem, you therefore have to compare:

$$\sum_{j=1}^{2} \lambda_{j1} p(y_j|\mathbf{x}) = \lambda_{11} p(y_1|\mathbf{x}) + \lambda_{21} p(y_2|\mathbf{x}) = 0 + \lambda_{21} p(y_2|\mathbf{x})$$

$$>?$$

$$\sum_{j=1}^{2} \lambda_{j2} p(y_j|\mathbf{x}) = \lambda_{12} p(y_1|\mathbf{x}) + \lambda_{22} p(y_2|\mathbf{x}) = \lambda_{12} p(y_1|\mathbf{x}) + 0$$

Minimum total risk: two classes

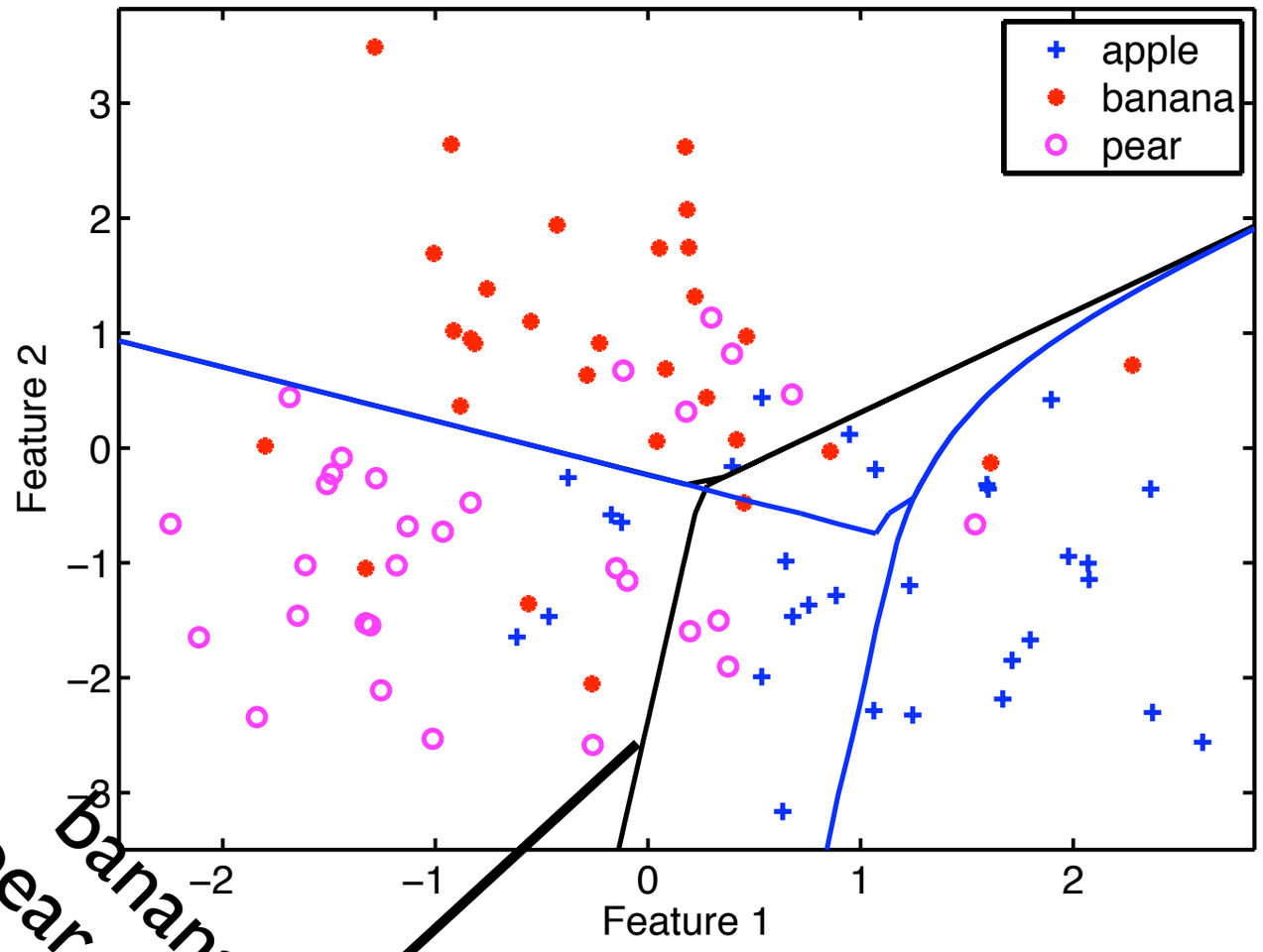
- When you predict class y_i for an object of class y_i you would typically define:

$$\lambda_{y_i,y_i} = 0$$

- For a two-class problem, you therefore have to compare:

$$\lambda_{21} p(y_2|\mathbf{x}) \quad \overset{?}{>} \quad \lambda_{12} p(y_1|\mathbf{x})$$

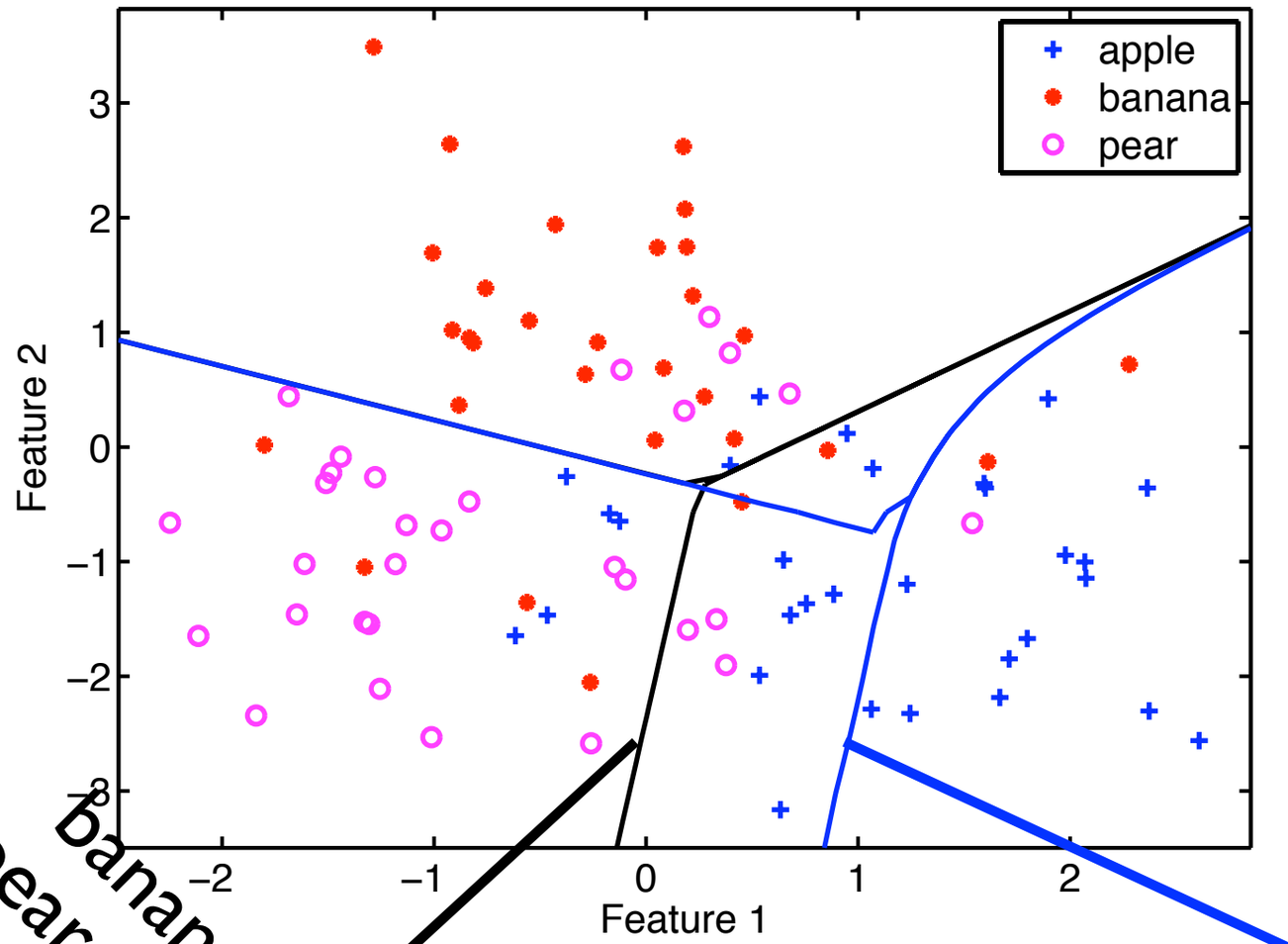
Example cost



classified

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

Example cost



$\lambda_{2,1} = 9$

Predicted: apple, pear, banana

True:

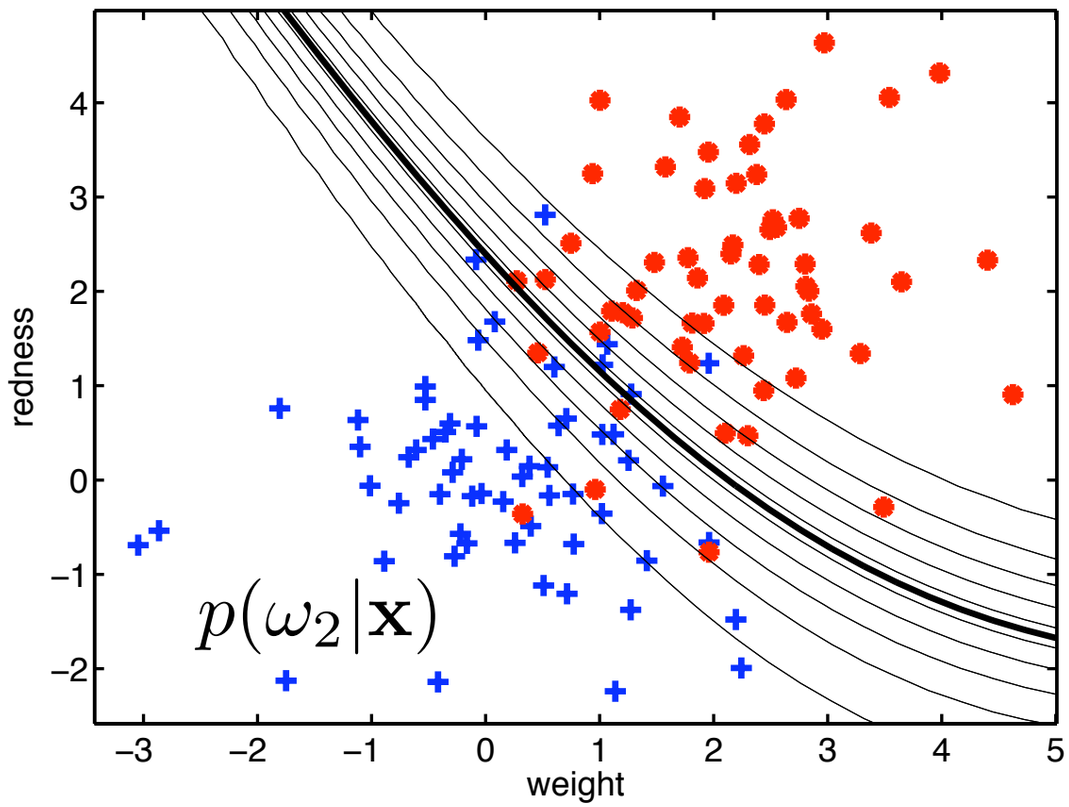
	apple	pear	banana
apple	0	1	1
pear	1	0	1
banana	1	1	0

Predicted: apple, pear, banana

	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 $p(y_1) = 2p(y_2)$?



Conclusions

- Machine Learning: learn from examples
- Classification: try to predict label of objects
- Best classifier is the Bayes' classifier

- Next lecture: quadratic classifier, linear classifier (LDA), nearest mean, more flexible classifiers