Machine Learning CSE2510 – Lecture 1.2: Probability theory / Bayes

Odette Scharenborg

Welcome to week 1 - lecture 2

- Administrative questions?
- Recap previous lecture
- Probability theory: Introduction
- Bayes' Rule
- Decision theory
- Bayes error
- Misclassification costs

TU Delft

Administrative questions?

TU Delft

Recap of the previous lecture

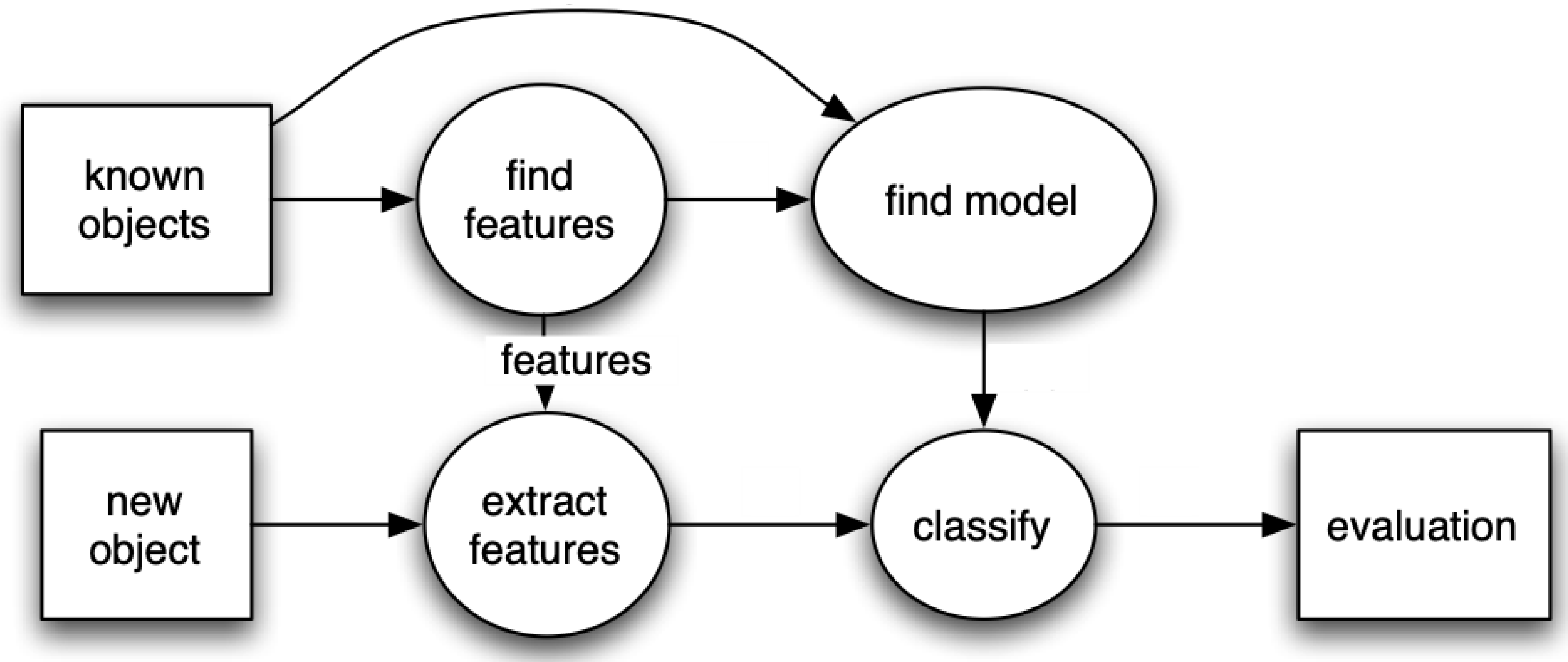
A(nother) definition of ML

- The learning of patterns or regularities in data by computer algorithms in order for these computer algorithms to carry out a specific task without using explicit instructions, but instead relying on these patterns and inference

[Wikipedia]

ML pipeline

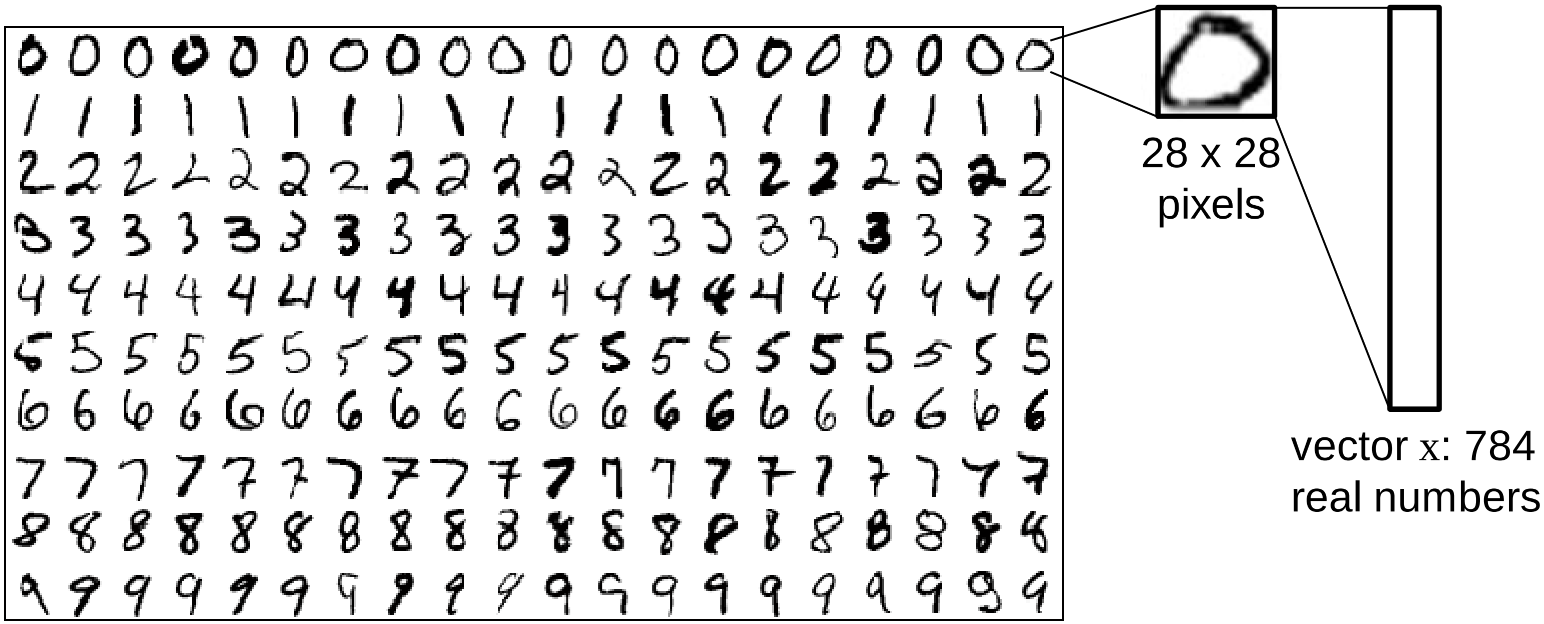
applying, generalisation



testing

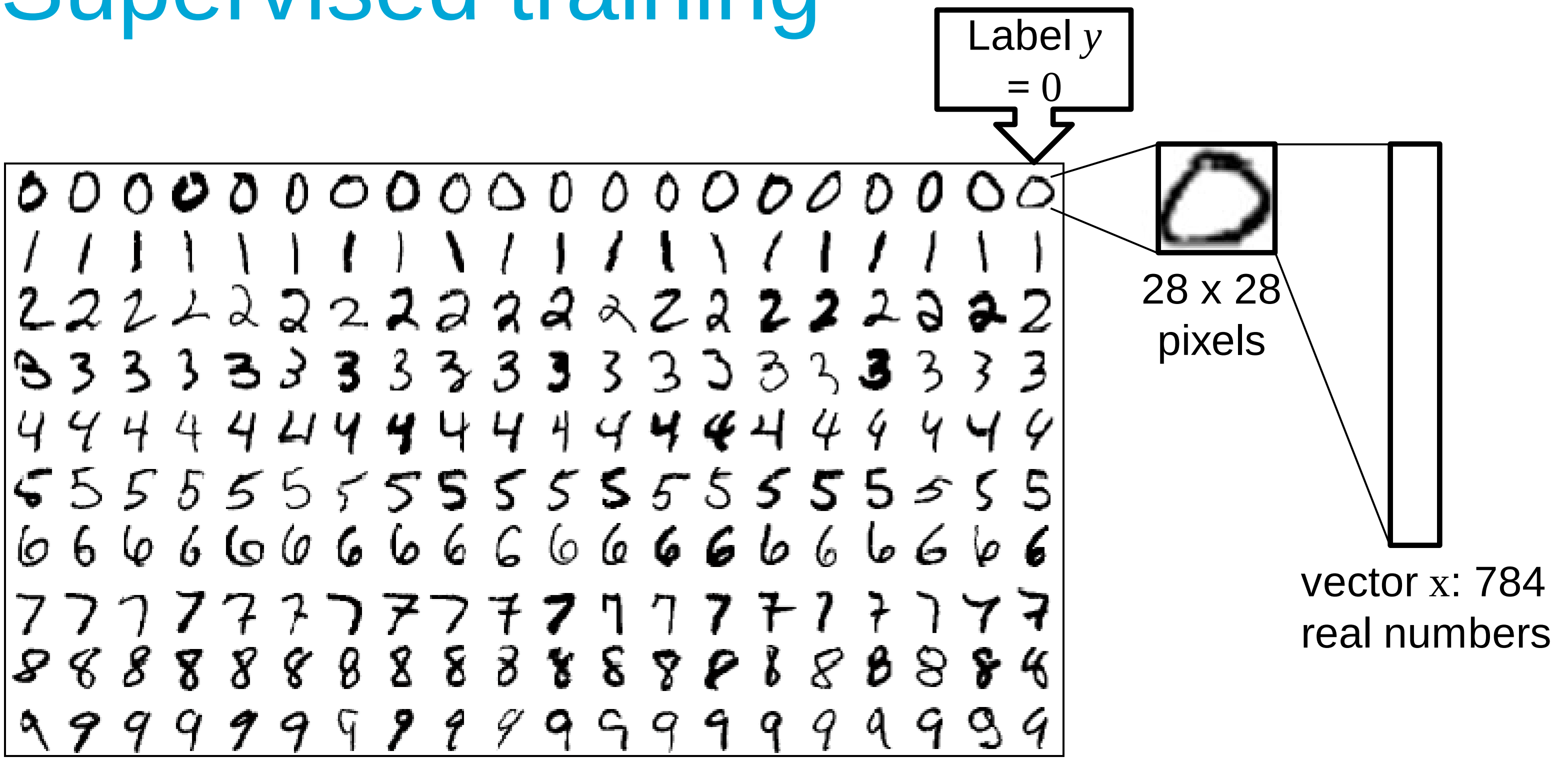
Slide adapted from Dr. David Tax

Example: handwritten digit recognition



Goal: build machine that takes a new input x and outputs the identity of the digit $0 \,..\, 9$

Supervised training



Training set: N digits $\{\mathrm{x}_1, .., \mathrm{x}_N\}$

with for each digit: label y (= target)

TUDelft

Train/test performance

1. Train on training data

- Performance is measured on the training set to guide the learning process

➔ Optimisation error

2. Evaluate the model

- Test model on *independent* test set for an unbiased estimate of the generalisability
  - No overlap with examples in training set
  - Similar to training data

➔ Test or generalisation error

Different ML tasks

- Supervised learning:
  - Classification: categorization into a prespecified number of discrete categories
  - Regression: predicting a continous value
  - Clustering: split the data into a number of groups with similar examples

Irrespective of the task; underlying it all $p(y|x)$:

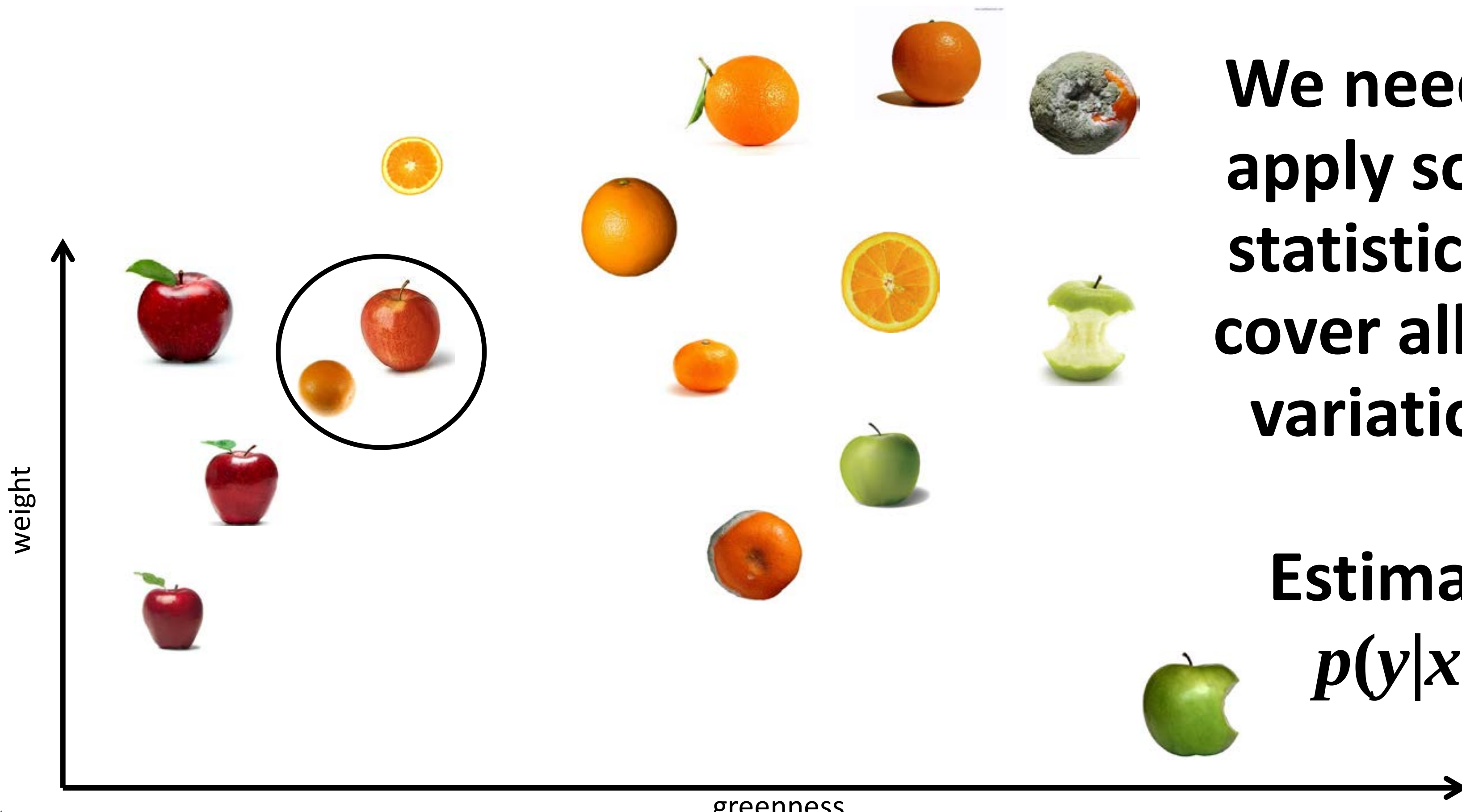
Probability theory ➔ Today!

TU Delft

ML techniques are based on measurements

Slide adapted from Dr. Marco Loog

Noise in the measurements



We need to apply some statistics to cover all the variations

Estimate
$$p(y|x)$$

Probabilistic classifiers

- Tuesday: “Hard” classification
  - Assign a sample to a class
  - Output the class label

- In the ideal world: “Probabilistic” classification
  - Estimate the probability distribution over a set of classes
  - Estimate the probability that sample belongs to a class
  - “Hard” classification: give sample label of the most likely class

  ➔ Probabilistic classifiers are a generalisation of the first type of classifiers

Today’s learning objectives

After practicing with the concepts of this week you are able to:

- Explain the basic ideas of probability theory, decision theory, and Bayes Rule and their application in Machine Learning

Introduction to Probability Theory

ML: design classifiers that classify an unknown object in the most likely class

→ Our task: how do we determine what is "most likely"?

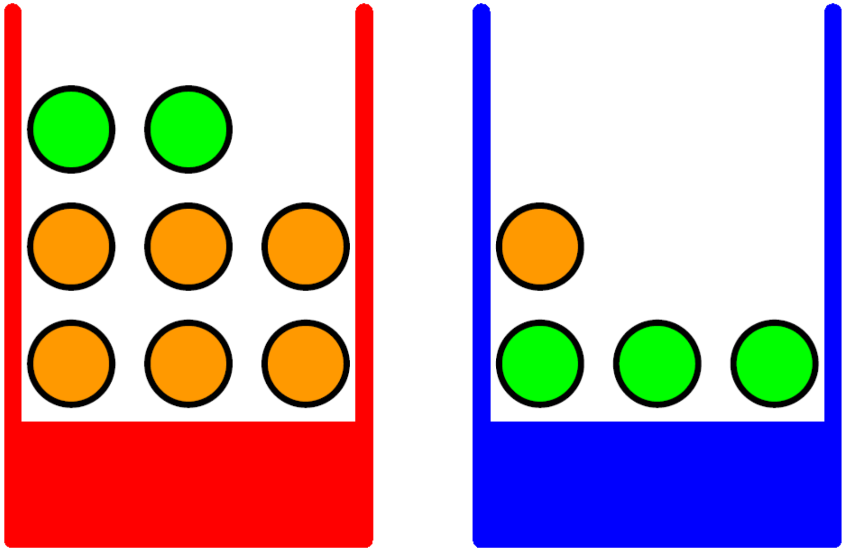
→ Estimate $p(y|x) = p(class|object)$

$$p(label|feature\ vector)$$

Probability theory: The discrete case

- Prob of selecting the red box $= p(B = \mathrm{r})$: $\frac{4}{10}$
- Prob of selecting the blue box $= p(B = \mathrm{b})$: $\frac{6}{10}$

Mutually exclusive?
Probability must sum to 1

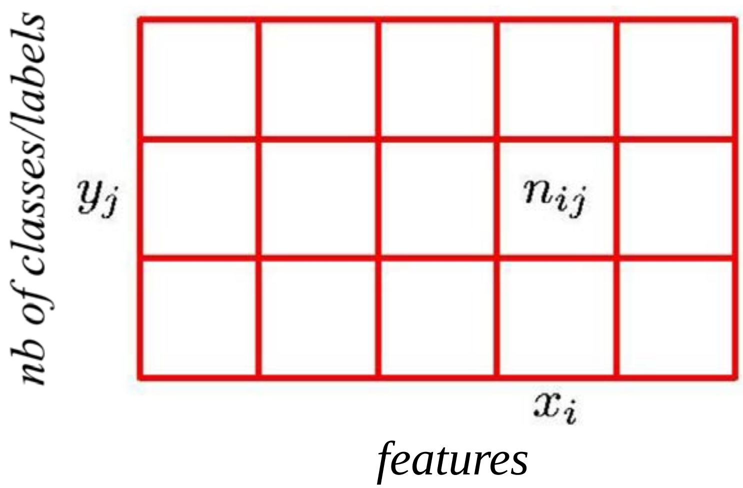


$\frac{4}{10}$ $\frac{6}{10}$

TU Delft

Joint probability: The discrete case

- Probability that $X = x_i$ (feature, F) and $Y = y_j$ (label, B): $p(X = x_i, Y = y_j)$
- E.g., probability that F=o (x) and B=r (y)



Joint Probability

$$p(X = x_i, Y = y_j) = \frac{n_{ij}}{N}$$

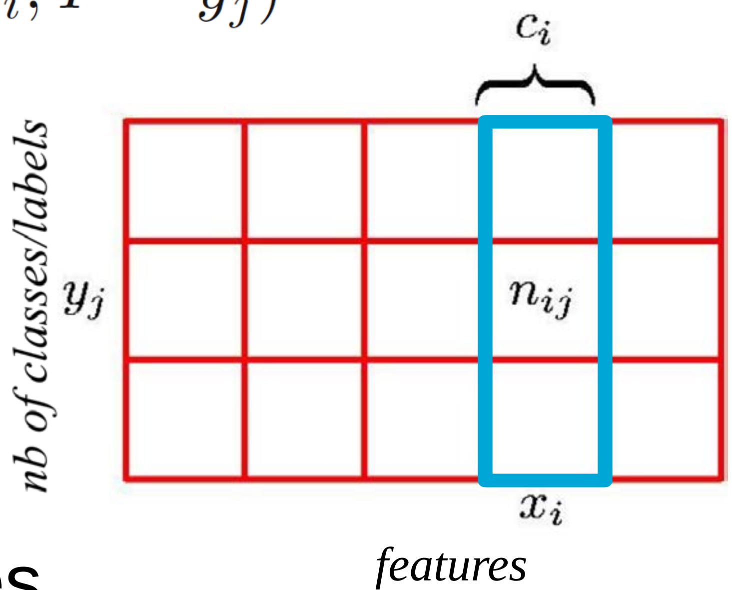
Sum rule of probability

- Probability that $X = x_i$ irrespective of Y: $p(X = x_i)$

$$p(X = x_i) = \sum_{j=1}^{L} p(X = x_i, Y = y_j)$$

c_i: nb of trials that $X = x_i$ irrespective of Y

L: total number of classes



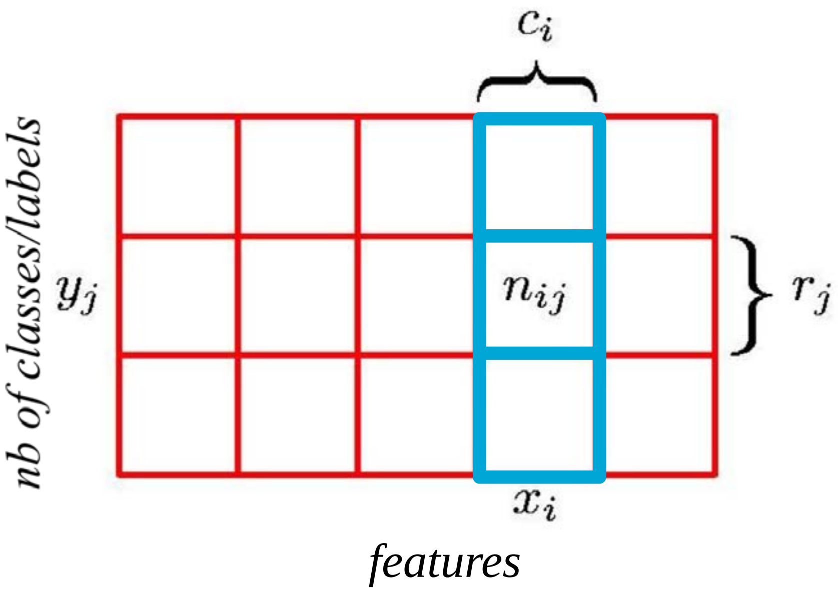
Marginal Probability

$$p(X = x_i) = \frac{c_i}{N}.$$

Conditional probability: $p(y|x)$

- Probability $Y = y_j$ given that $X = x_i$: $p(Y = y_j \mid X = x_i)$
- E.g., probability that $B = r$ given that $F = o$

$$p(Y = y_j | X = x_i) = \frac{n_{ij}}{\boxed{c_i}}$$



- Q: What is the difference between the conditional and the joint probability?

Fundamental rules of probability

- $p(X,Y)$: joint probability = probability of X **and** Y
- $p(Y|X)$: conditional probability = probability of Y **given** X
- $p(X)$: marginal probability of X

- $p(X,Y) = p(Y,X)$: symmetry property
- $p(X,Y) = p(Y|X)\, p(X)$: product rule

Bayes' Rule

With labeled examples of the classes → estimate a probability density per class

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

Difficult to estimate for continuous variables

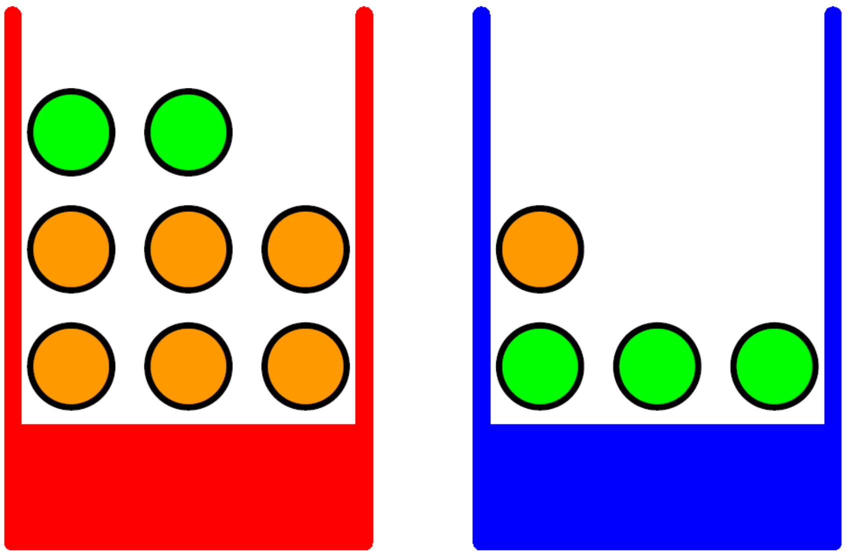
→ Central role in machine learning

TU Delft

An exercise

- Probability of picking an apple? = $p(F{=}a)$

$$
\begin{aligned}
p(F=a) &= p(F=a|B=r)p(B=r) + p(F=a|B=b)p(B=b) \\
&= \frac{1}{4} \times \frac{4}{10} + \frac{3}{4} \times \frac{6}{10} = \frac{11}{20}
\end{aligned}
$$

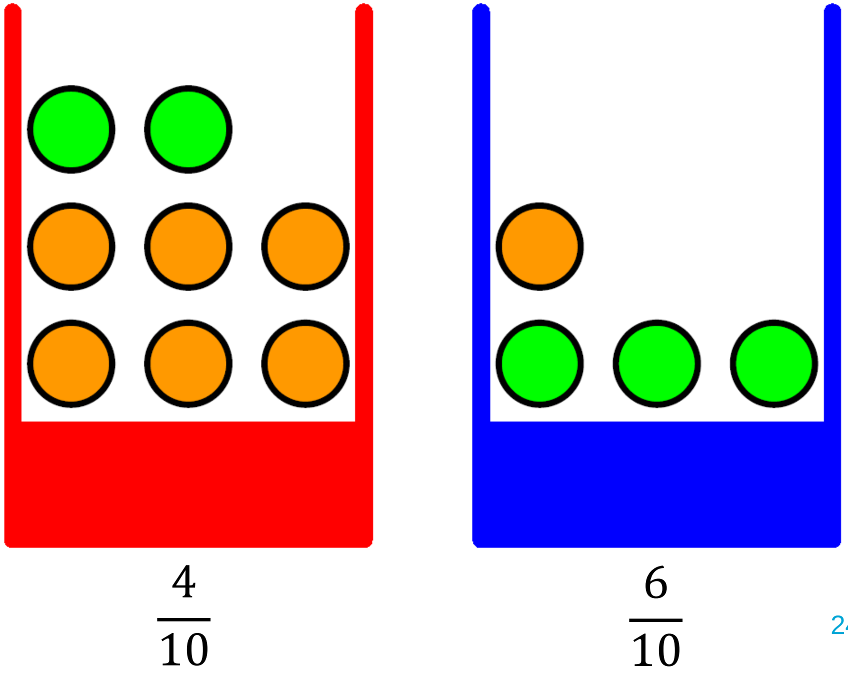


$\frac{4}{10}$ $\frac{6}{10}$

Another exercise

- Probability of picking an orange? = $p(F=o)$

➔ Sum rule: $p(F=o) = 1 - 11/20 = 9/20$



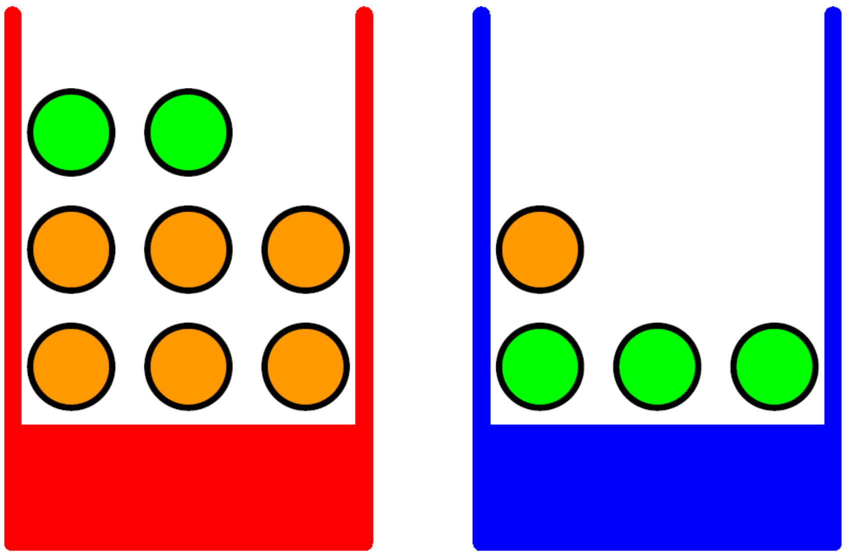
TU Delft

Conditional probability: $p(y|x)$

- Q: Probability of B=r given that F=o?

$$p(B=r|F=o) = \frac{p(F=o|B=r)p(B=r)}{p(F=o)} = \frac{3}{4} \times \frac{4}{10} \times \frac{20}{9} = \frac{2}{3}$$

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$



$\frac{4}{10}$ $\frac{6}{10}$

TU Delft

Bayes' Rule

Prior probability

Posterior probability

$$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

Prior prob of selecting the red box (Y), i.e., *before* observing an orange: $p(B{=}\mathrm{r})$: $\frac{4}{10}$

Posterior prob of selecting the red box (Y), i.e., *after* observing an orange: $p(B{=}\mathrm{r} \mid F{=}\mathrm{o})$: $\frac{2}{3}$

Classification

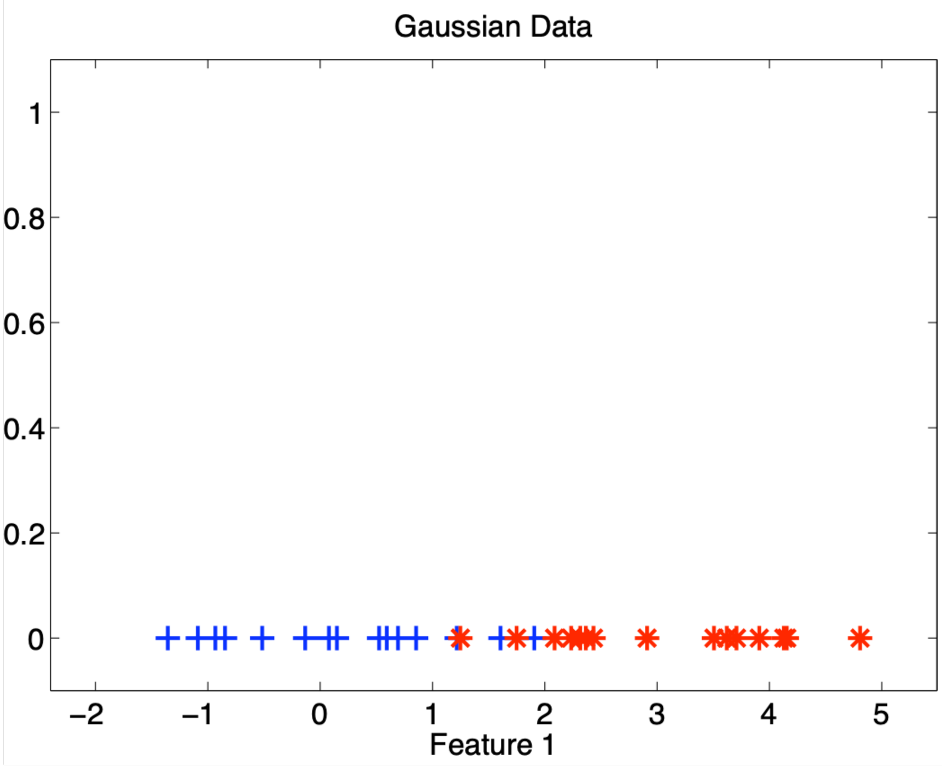
- Training: Estimate the probability distribution over a set of classes

- Testing: Estimate the probability that sample belongs to a class

TU Delft

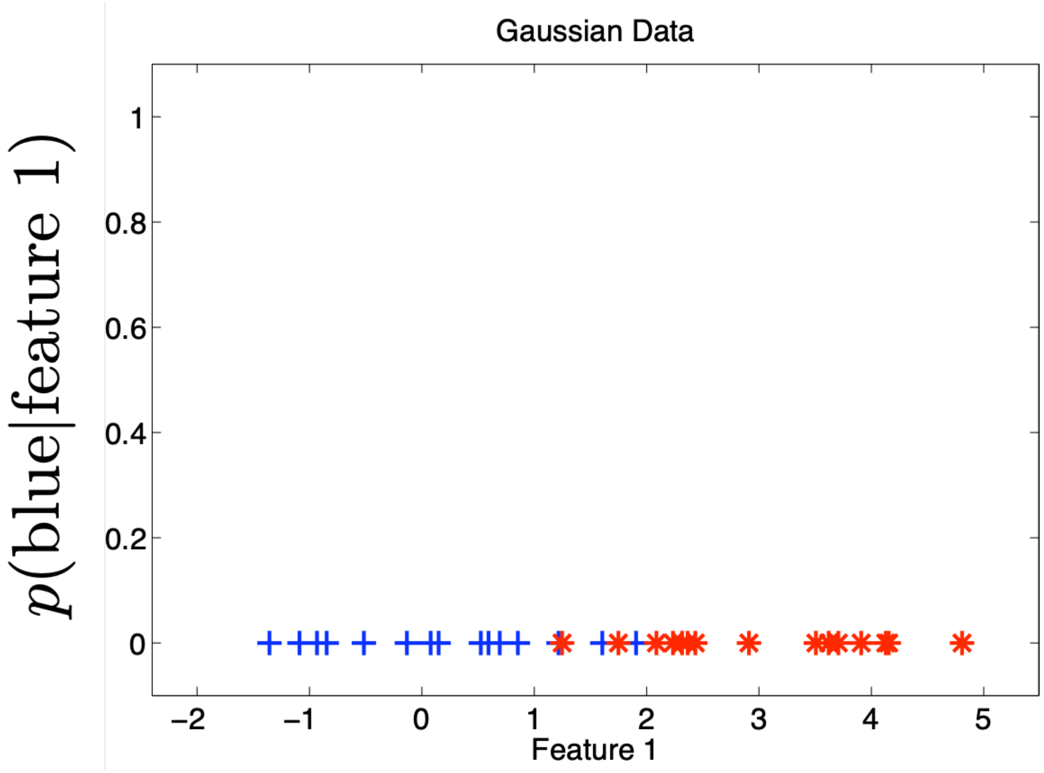
Estimating the probability distribution over a set of classes: The continuous case

- Given a feature, and a training set, where is the blue (e.g., apples) class?



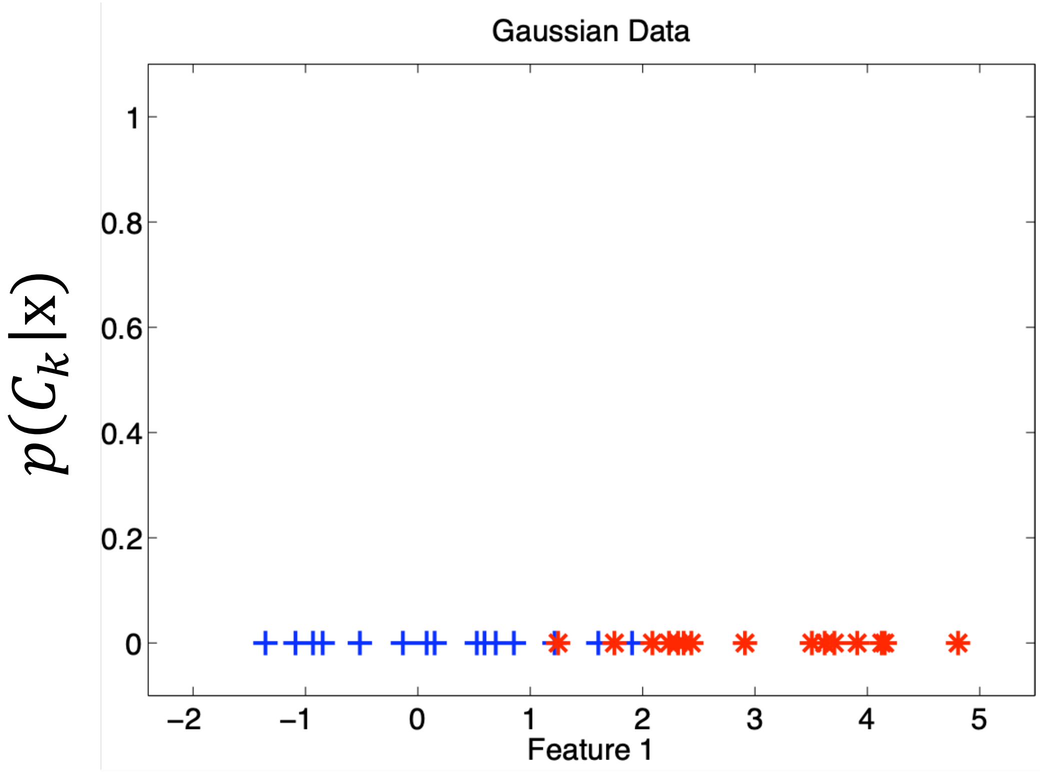
Estimating the probability distribution over a set of classes: The continuous case

- For each object we want to estimate $p(\text{blue}|\text{feature } 1)$



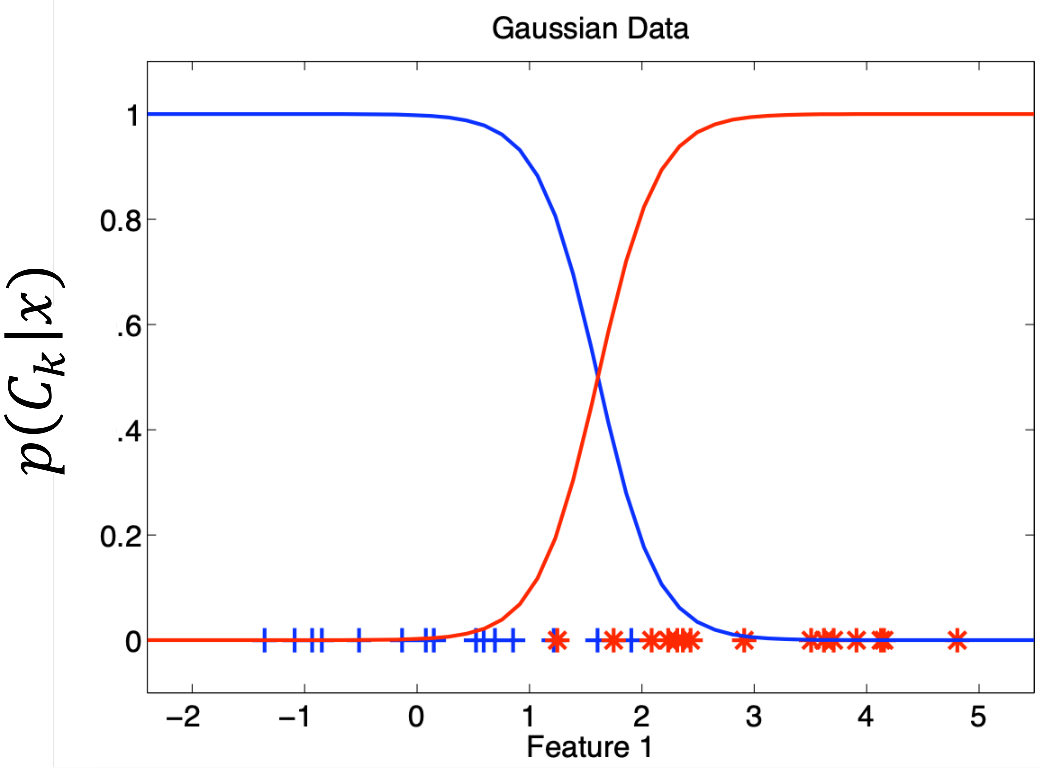
Class conditional probability

- For each object we want to estimate $p(C_k|x)$



Probability distribution over the classes

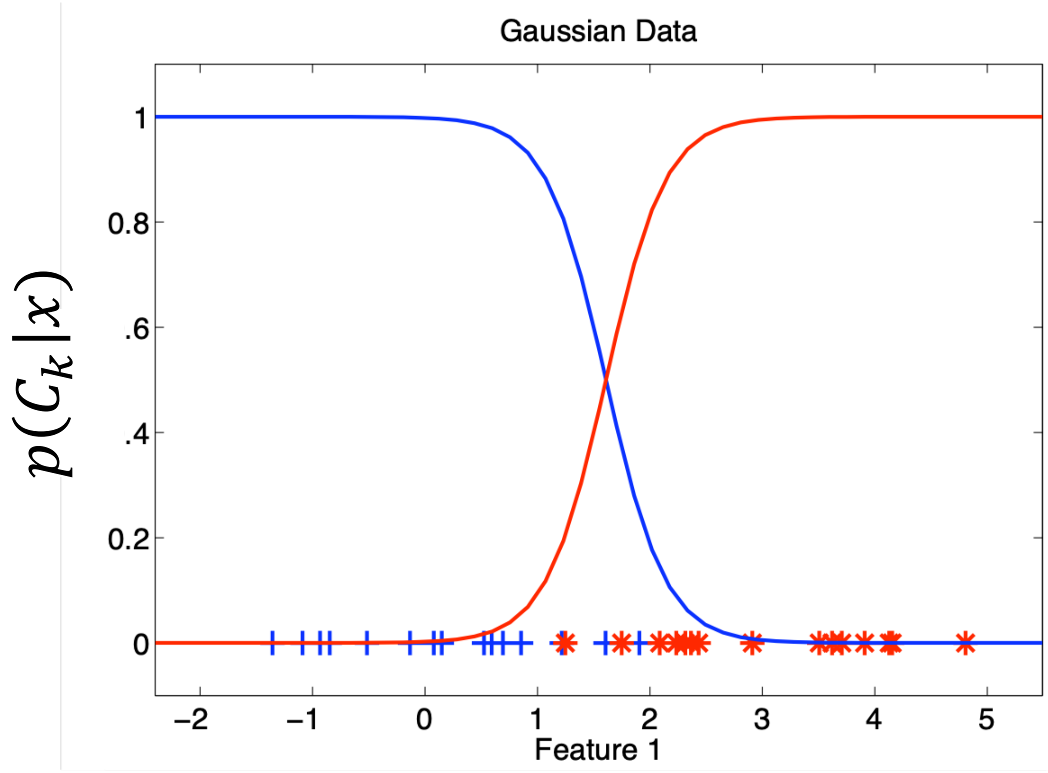
- For each object we have to estimate $p(C_k|x)$



$$\sum_k p(C_k|x) = 1$$

In order to classify a new x

- Estimate the posterior probability $p(C_k|x)$

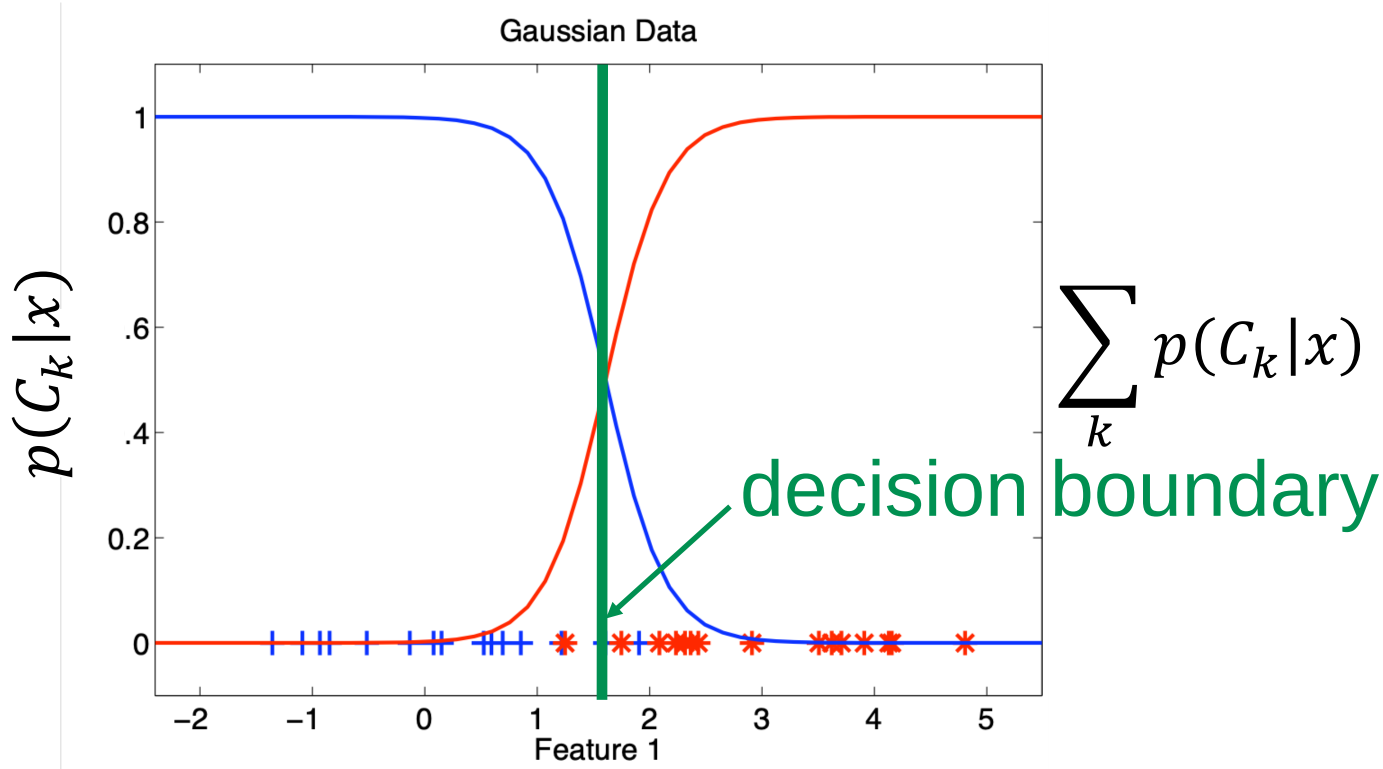


$$\sum_k p(C_k|x) = 1$$

Slide adapted from Dr. David Tax

In order to classify a new x

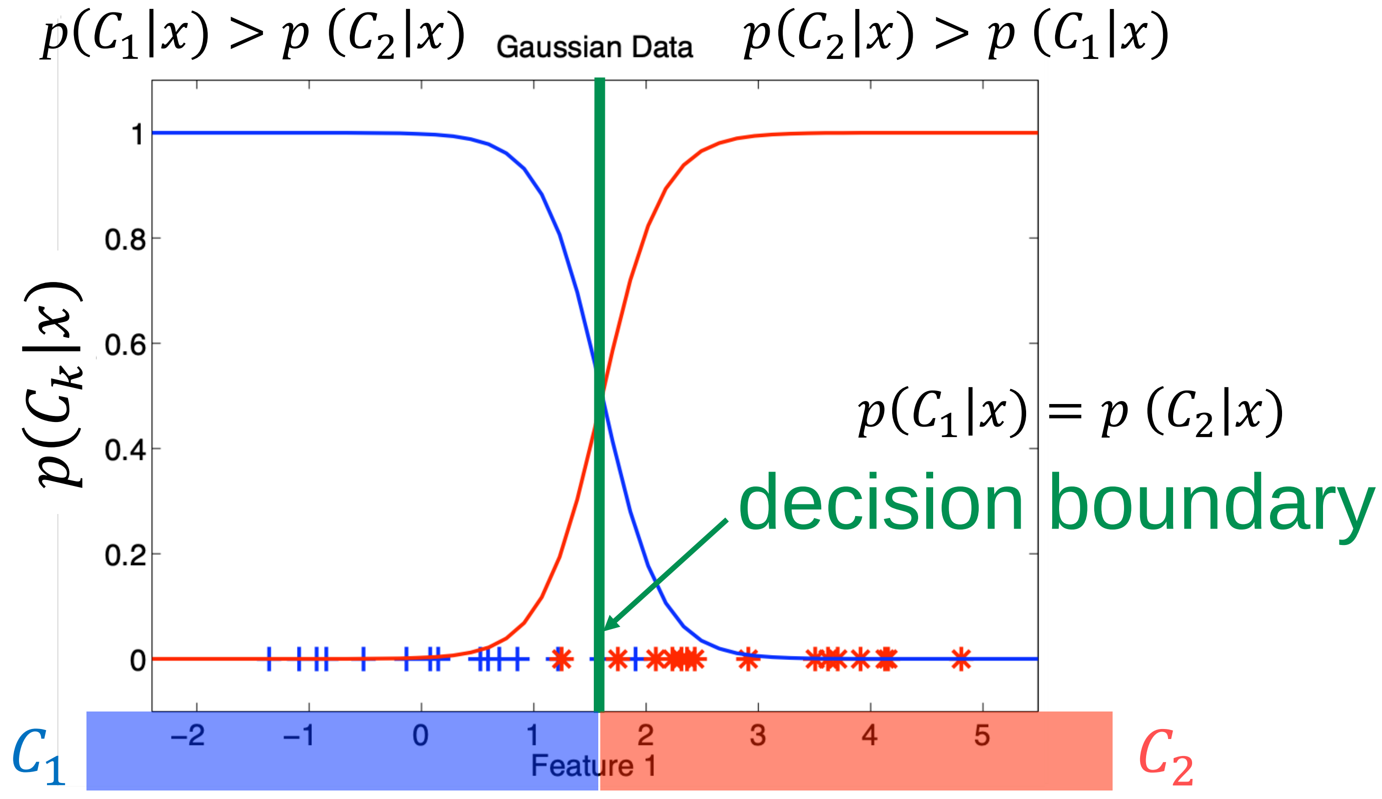
- Estimate the posterior probability $p(C_k|x)$



$$\sum_k p(C_k|x) = 1$$

"Hard" classification of the new object x

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



Slide adapted from Dr. David Tax

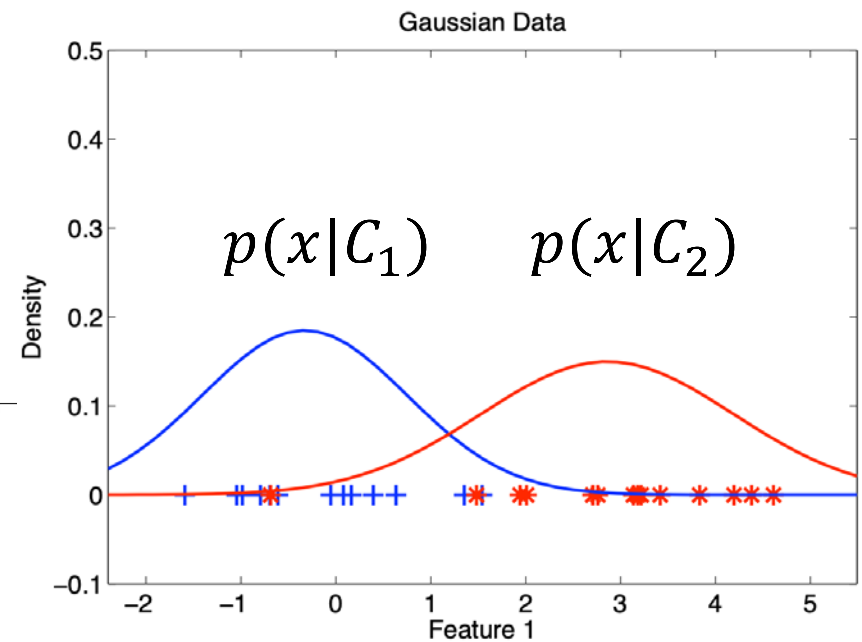
Description of a classifier: Decision theory

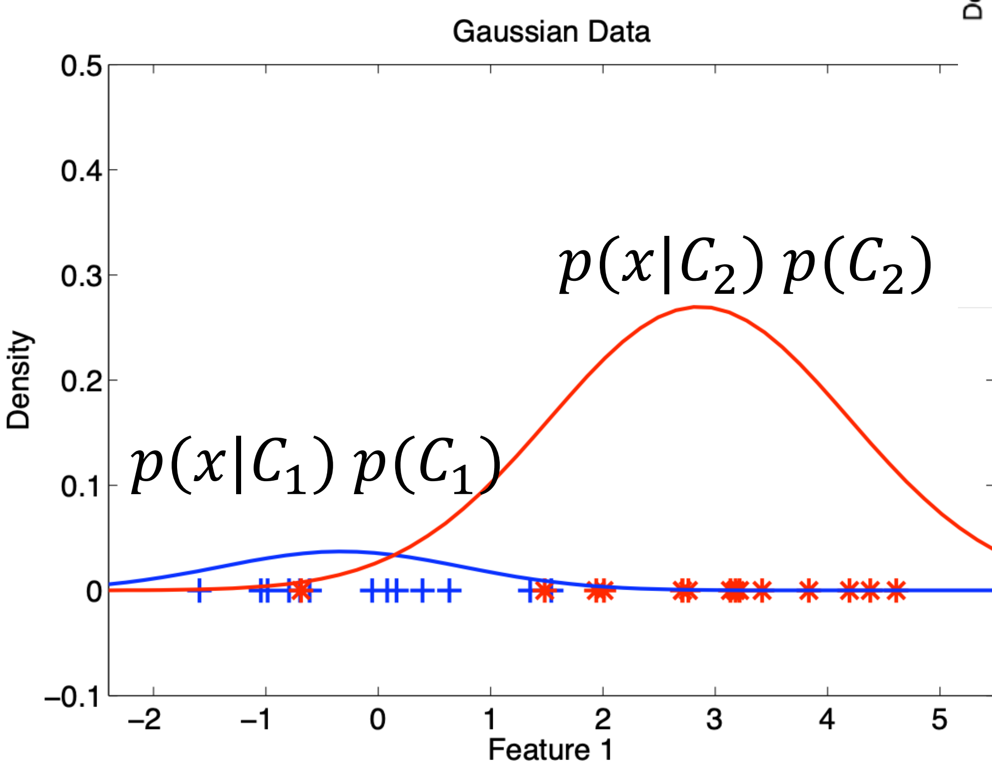
There are several ways to describe a classifier:

- if $p(C_1|\mathrm{x}) > p\,(C_2|\mathrm{x})$ then assign to C_1
  otherwise C_2
- if $p(C_1|\mathrm{x}) - p(C_2|\mathrm{x}) > 0$ then assign to C_1
- or $\frac{p(C_1|\mathrm{x})}{p(C_2|\mathrm{x})} > 1$ then assign to C_1
- or …

How do we calculate the posterior probabilities?

➔ Bayes' Rule





1. Estimate the class probabilities

2. Multiply with the class priors

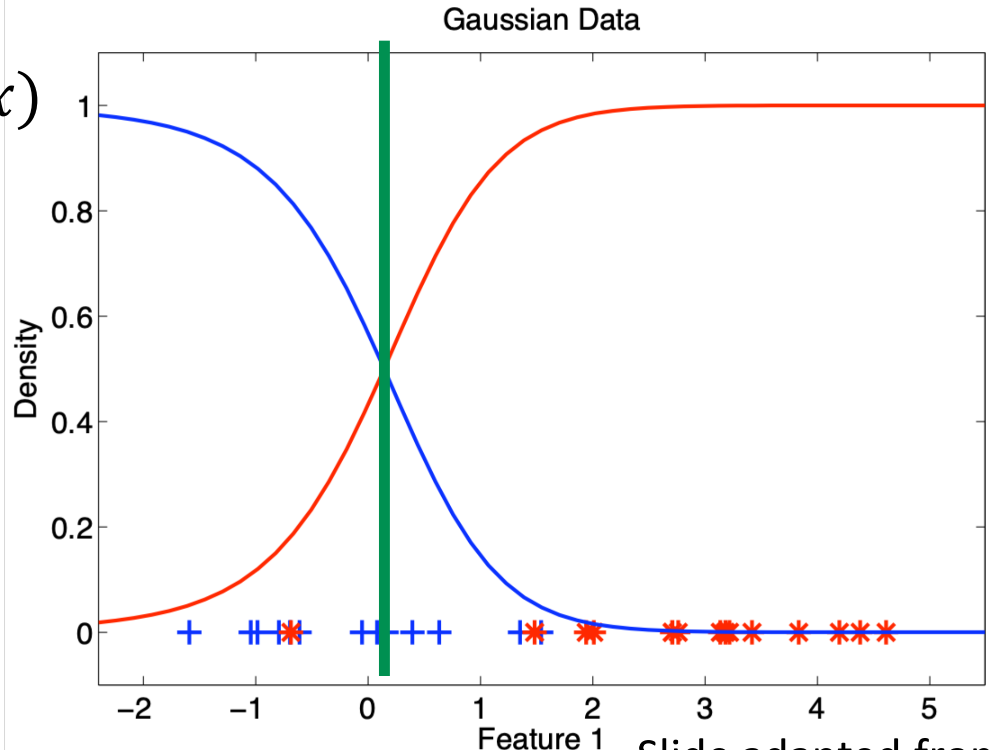
TUDelft

Slide adapted from Dr. David Tax

Bayes' Rule

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

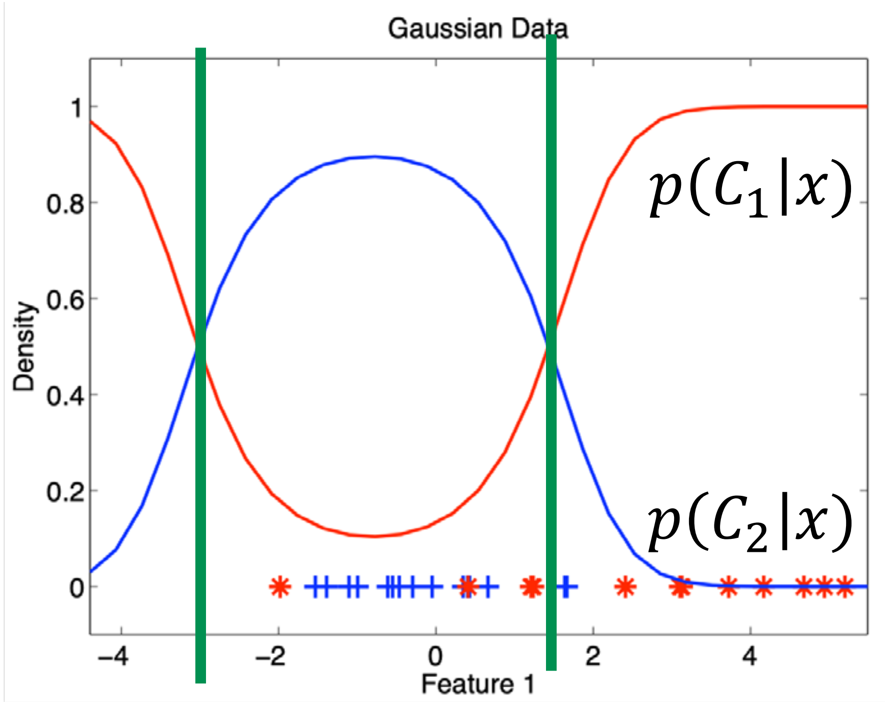
$p(C_1|x)$



$p(C_2|x)$

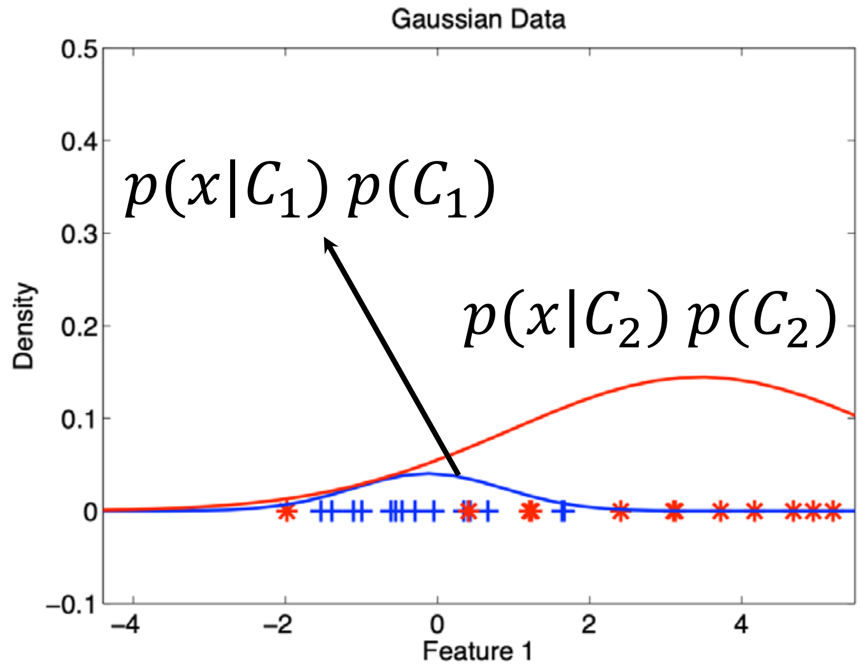
Slide adapted from Dr. David Tax

Complicated decision boundary



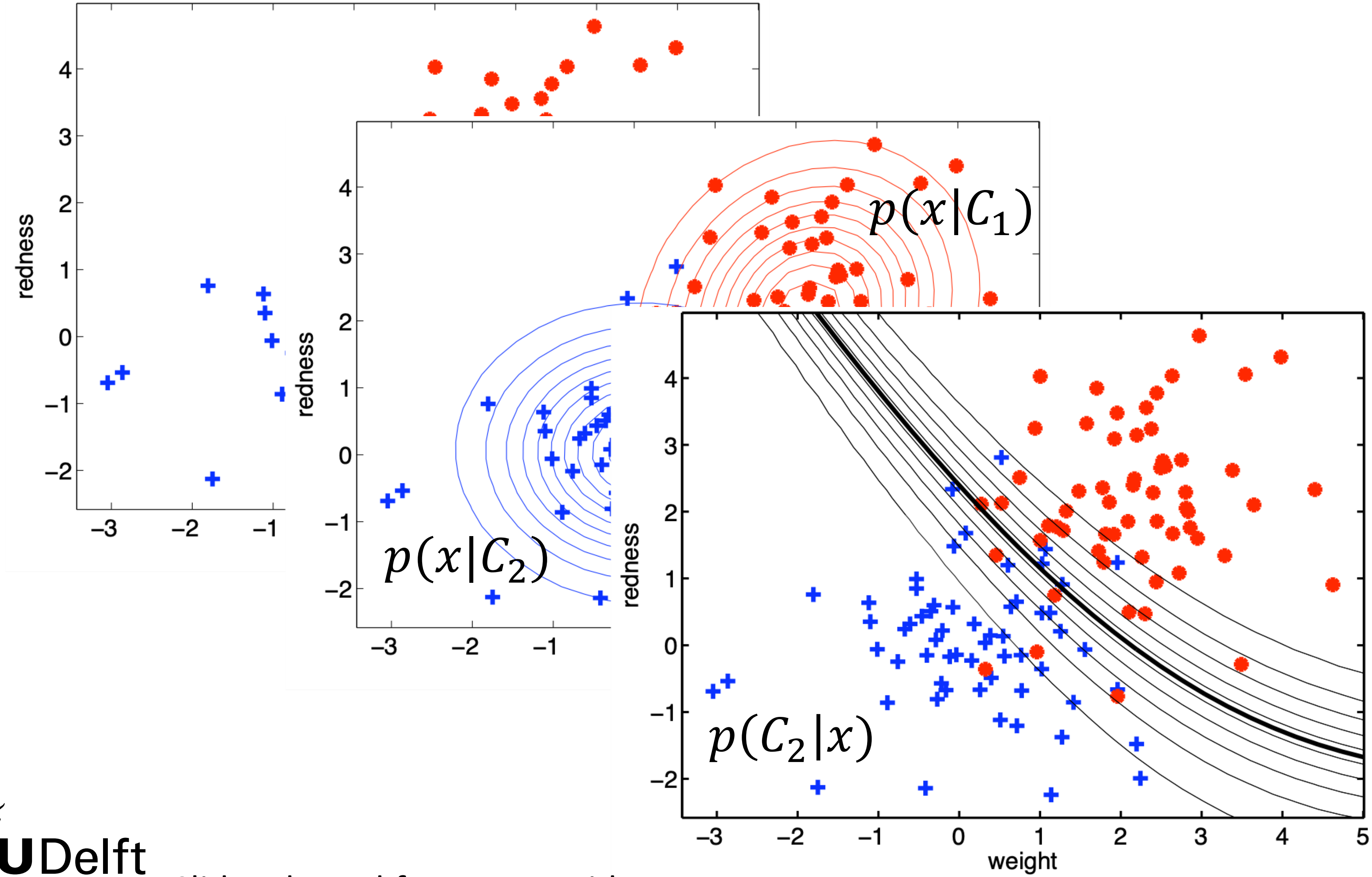
- Depending on the class conditional probability densities, complicated decision boundaries can appear

Missing decision boundary

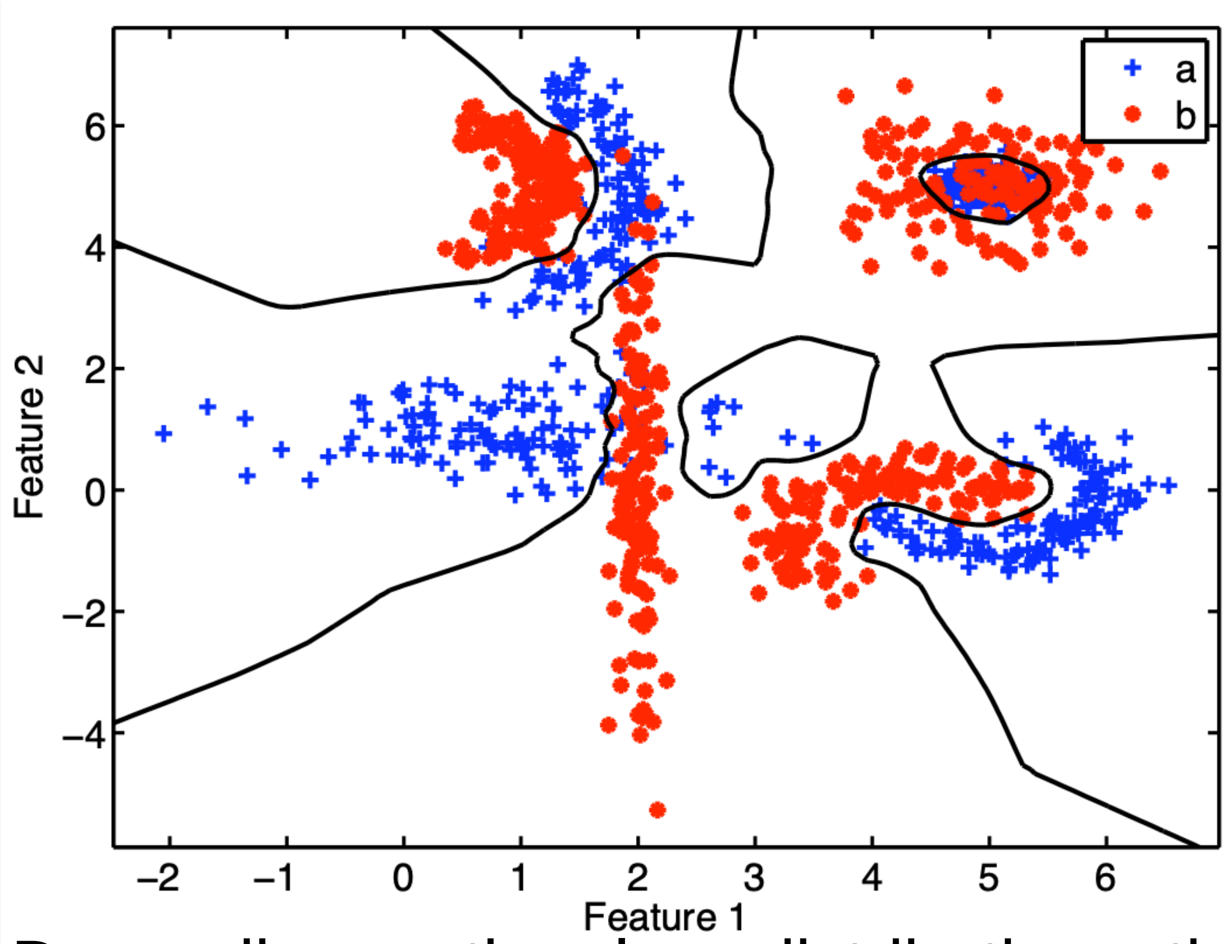


- A class can be too small (class prior is low) or too dispersed, that no objects are assigned to that class

Higher (2)-dimensional feature space



Multi-modal distributions



- Depending on the class distributions, the decision boundary can have arbitrary shapes

TU Delft

Slide adapted from Dr. David Tax

The class conditional probabilities

- How do we obtain the class conditional probabilities $p(x|C_k)$?
- We need a model
- During training, estimate the model parameters such that the example objects fit well: maximum likelihood estimators
- This will be the topic for the coming weeks

Bayes Error

- Even if $p(x, y)$ is perfectly known (true distribution), errors predicting y from x will occur because the posteriors $p(y|x)$ are often not exactly 0 or 1

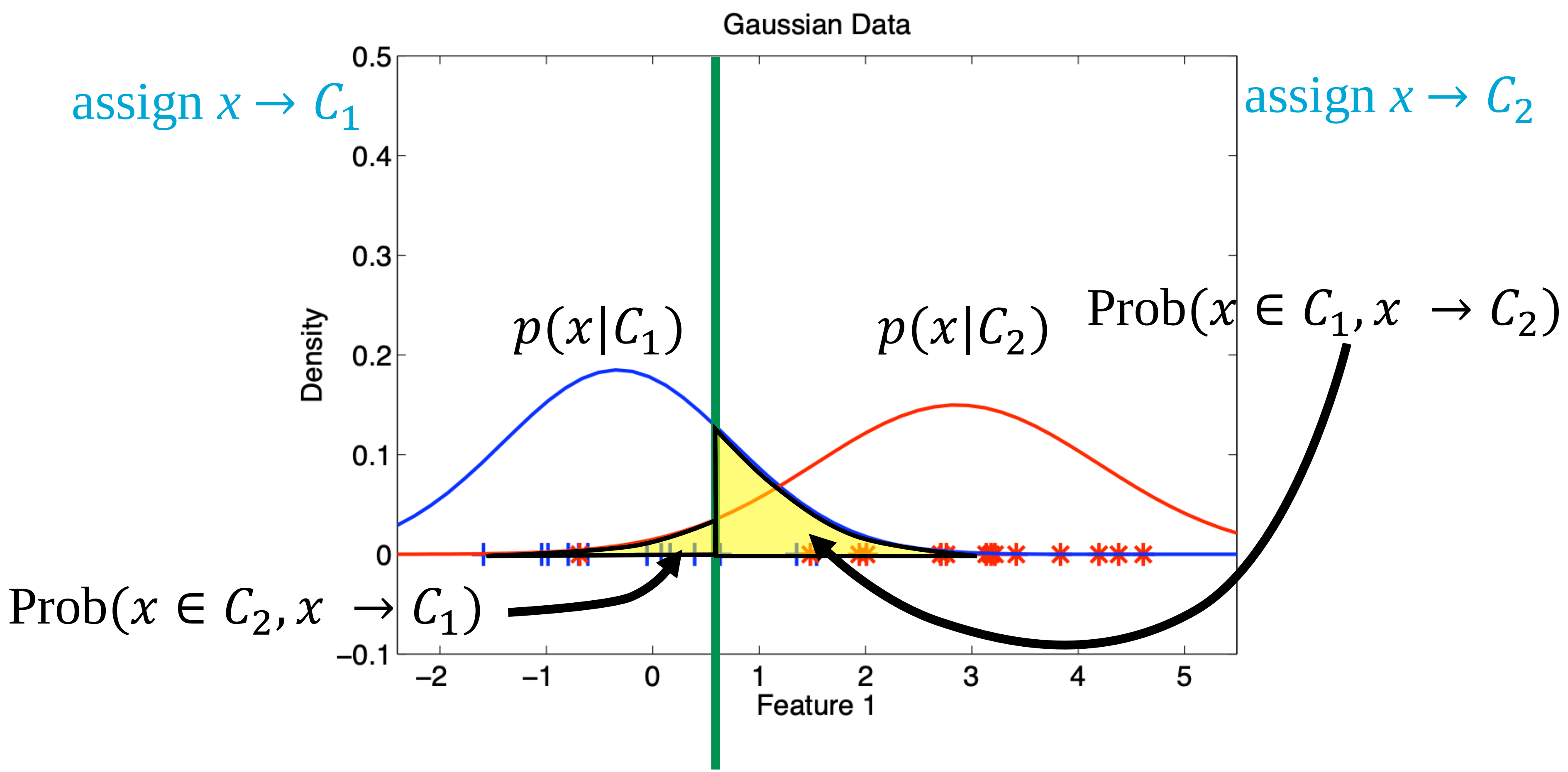
➔ lowest possible prediction error

== Bayes' error

== irreducible error

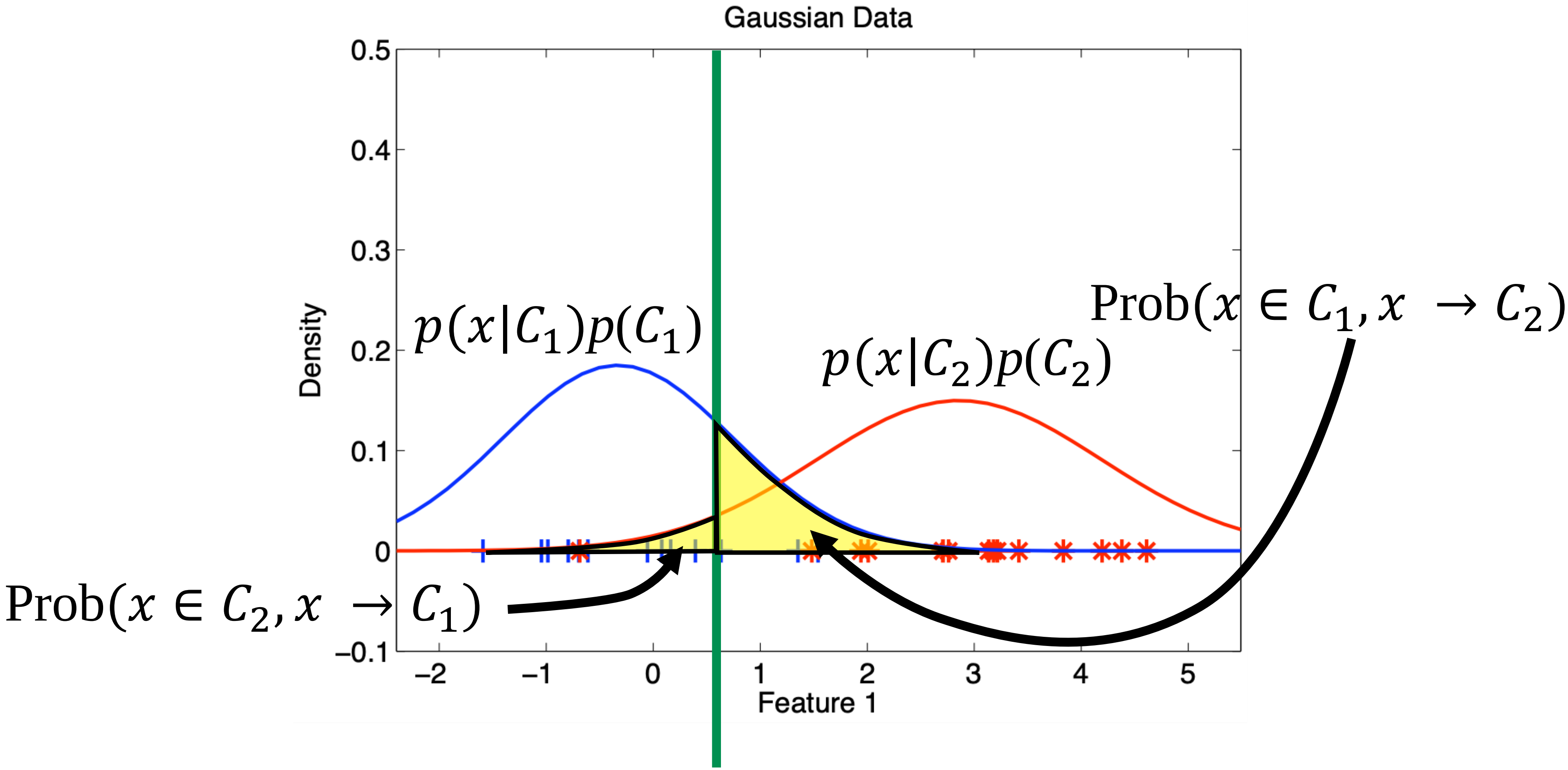
How good is the classifier?

- The error of the green decision boundary:



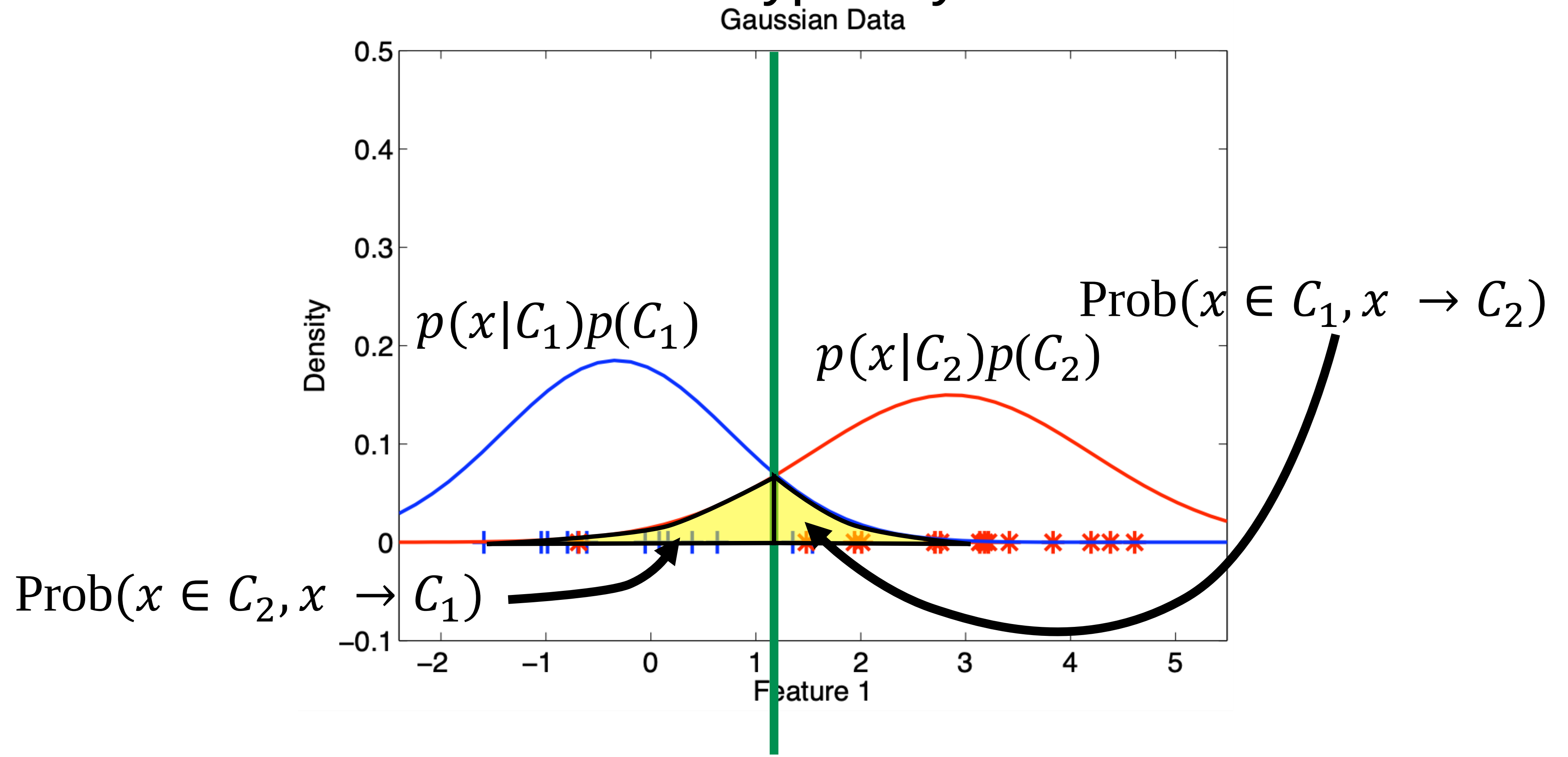
Classification error

- The error: $P(error) = \sum_{i=1}^{C} P(error|C_i)P(C_i)$



Bayes Error ε^*

- The **minimum** error: typically > 0!



Slide adapted from Dr. David Tax

Bayes Error

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

Misclassification costs

- We want to make as few errors as possible with the assignment of x to class C
- Error: x is assigned to C_1 but should have been assigned to C_2 and vv.

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’ as ‘ill’

misclassification:

classify ‘ill’ as ‘healthy’

Misclassification cost

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

misclassification:

$\lambda_{healthy,ill} = 1$

misclassification:

$\lambda_{ill,healthy} = 1000$

Misclassification cost

- Preferably fewer $\lambda_{ill,healthy}$ misclassifications than $\lambda_{healthy,\,ill}$ misclassifications
- Even if that means an increase in $\lambda_{healthy,\,ill}$ misclassifications

→ Loss function balances these wishes

- Measure of loss incurred in taking any of the available decisions or actions

→ Minimise the loss function

Solving decision problems

1. Generative models (week 2 & 3)
2. Discriminative models (week 4 & 6)

Generative vs discriminative models

- Generative models (week 2 & 3):
  - Model the actual distribution of each class
  - Learn the class conditional probability $p(X|Y)$
  - Allows you to *generate* new samples from all classes
  - Predict the posterior probability using Bayes' Rule

- Discriminative models (week 4 & 6):
  - Directly estimate $p(Y|X)$ which *discriminates* between classes
  - No $p(X|Y)$ → no generation of new data from classes

Conclusions

- ML is "probabilistic" classification:
  - Estimate the posterior conditional probability using

Bayes' Rule: $$p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$$

- "Hard classification" is done using decision theory