Non-parametric density estimation

Gosia Migut

Slides credit: David Tax

TU Delft

Admin stuff

- Lab 3 downloads: 220
- Questions lab 3: 200+

- Keep practicing!

After practicing with the concept of this lecture you should be able to:

- Explain what are and how to use the learning curves
- Explain the Naive Bayes classifier, including the following:
  - components and their function
  - independence assumption
  - dealing with missing data
  - Continuous example
  - Discrete example
  - Pros and cons

TU Delft

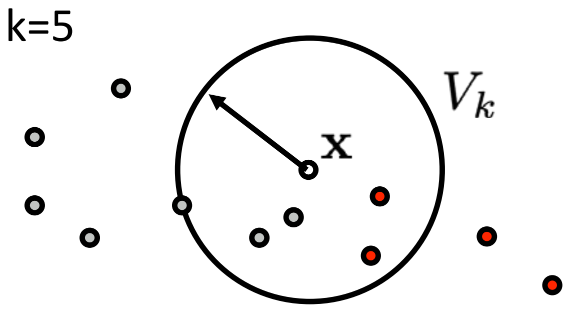
Literature

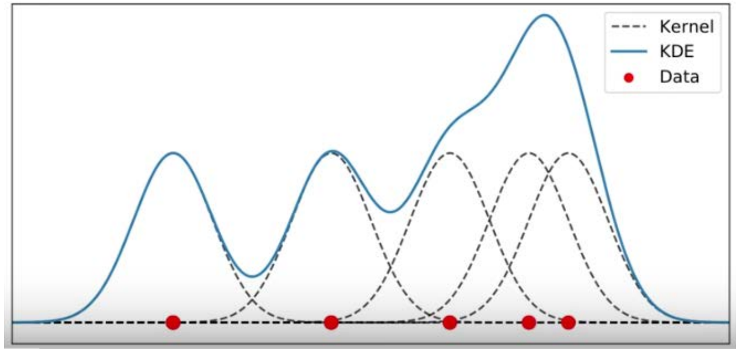
- Naive bayes
  - Lecture notes CS229: section 2 and 2.1 (excluding 2.2). Andrew Ng, Standford University. http://cs229.stanford.edu/notes/cs229-notes2.pdf
- Learning curves
  - Section 8.2 from "Pattern Recognition: Introduction and Terminology" by R.P.W. Duin and E. Pekalska.

  http://www.37steps.com/data/pdf/PRIntro_large.pdf

Recap last lecture

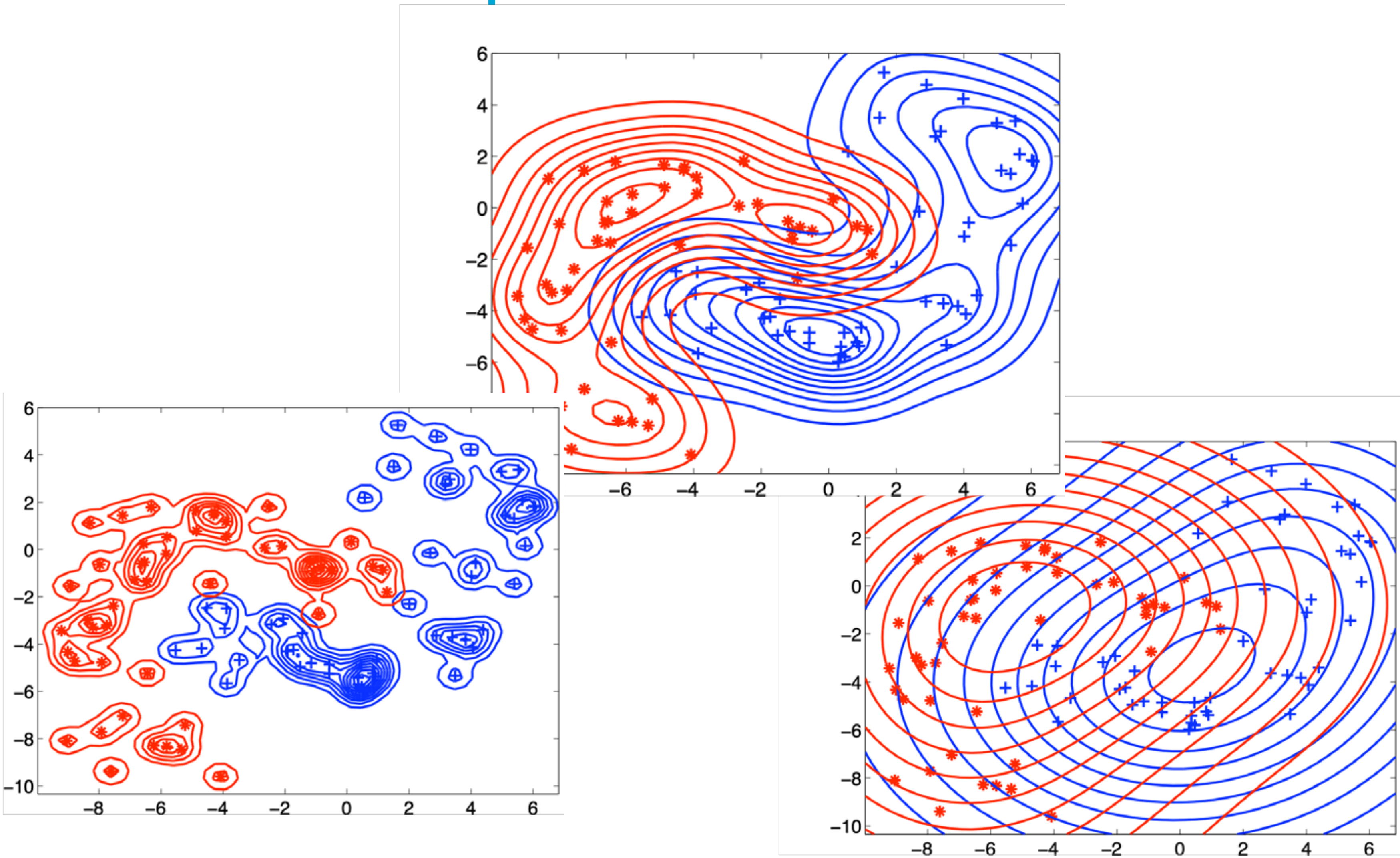
- Non-parametric density estimation
- K-nn and Parzen



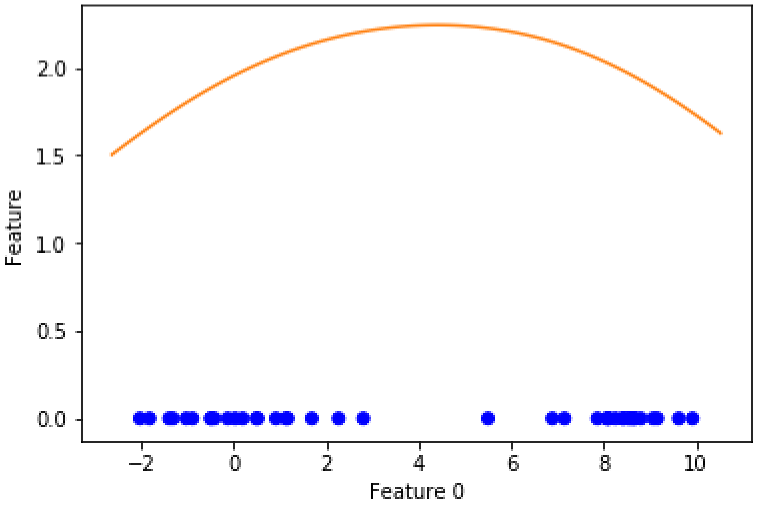


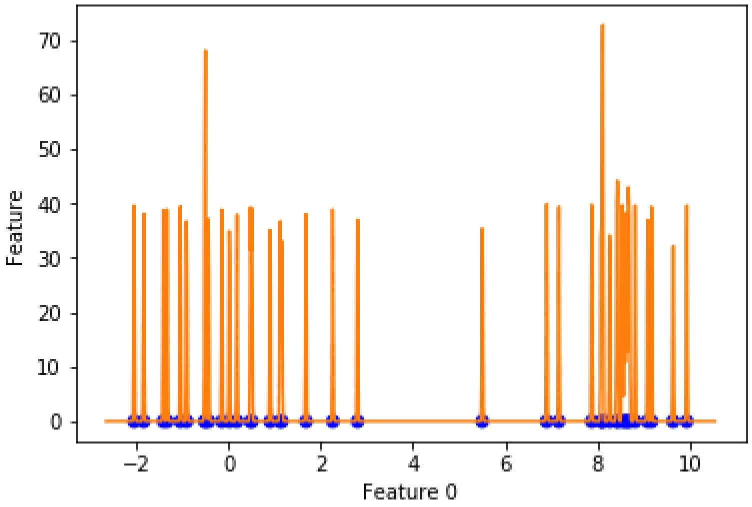
- Lab: optimize k for k-nn
- Now: optimize h for Parzen density estimation

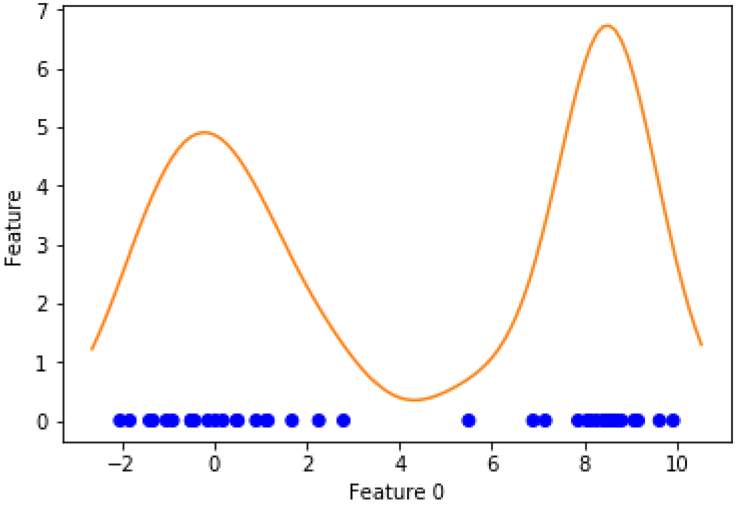
Parzen width parameter



Parzen densities for different h







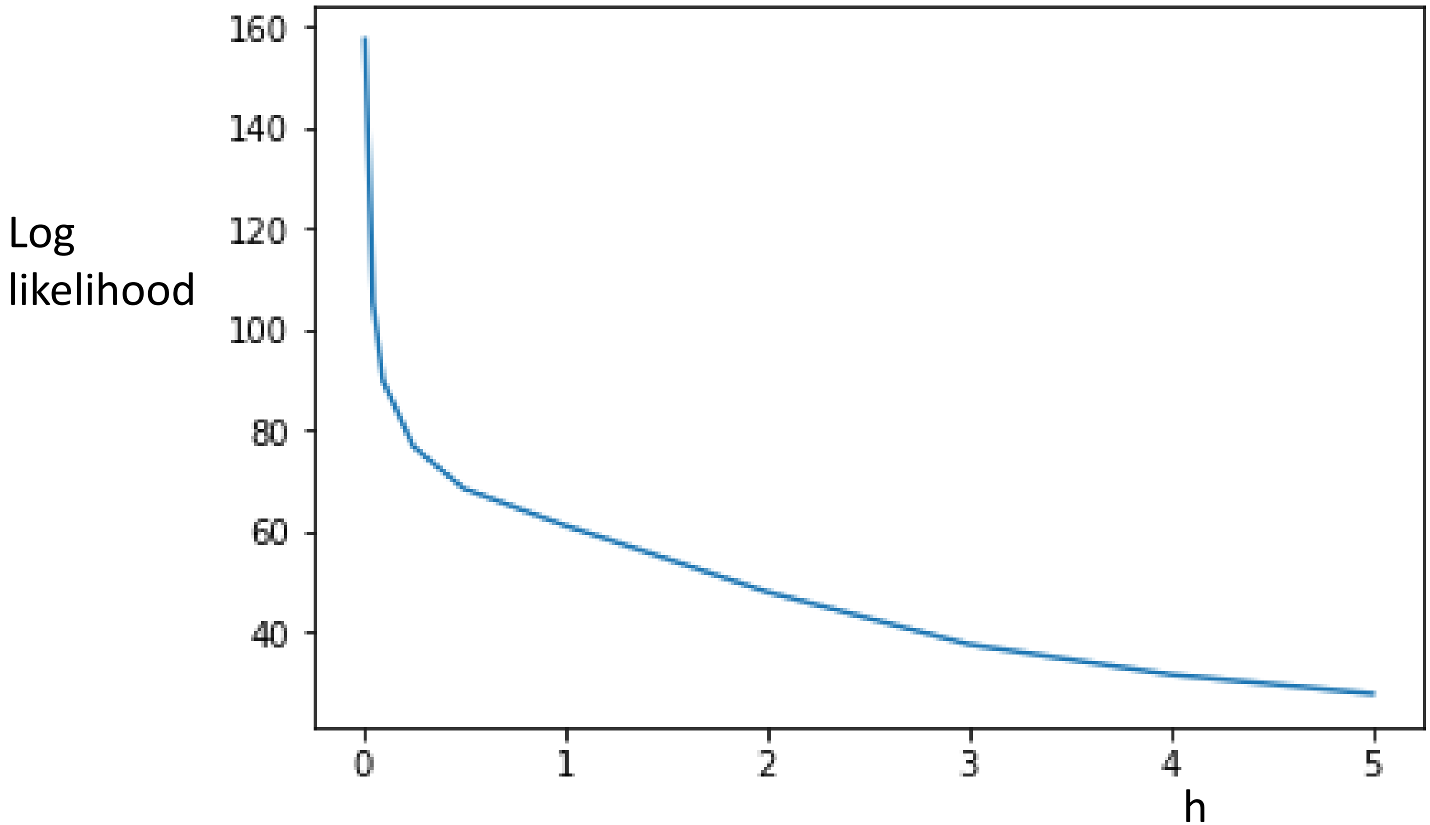
TUDelft

Log-likelihood

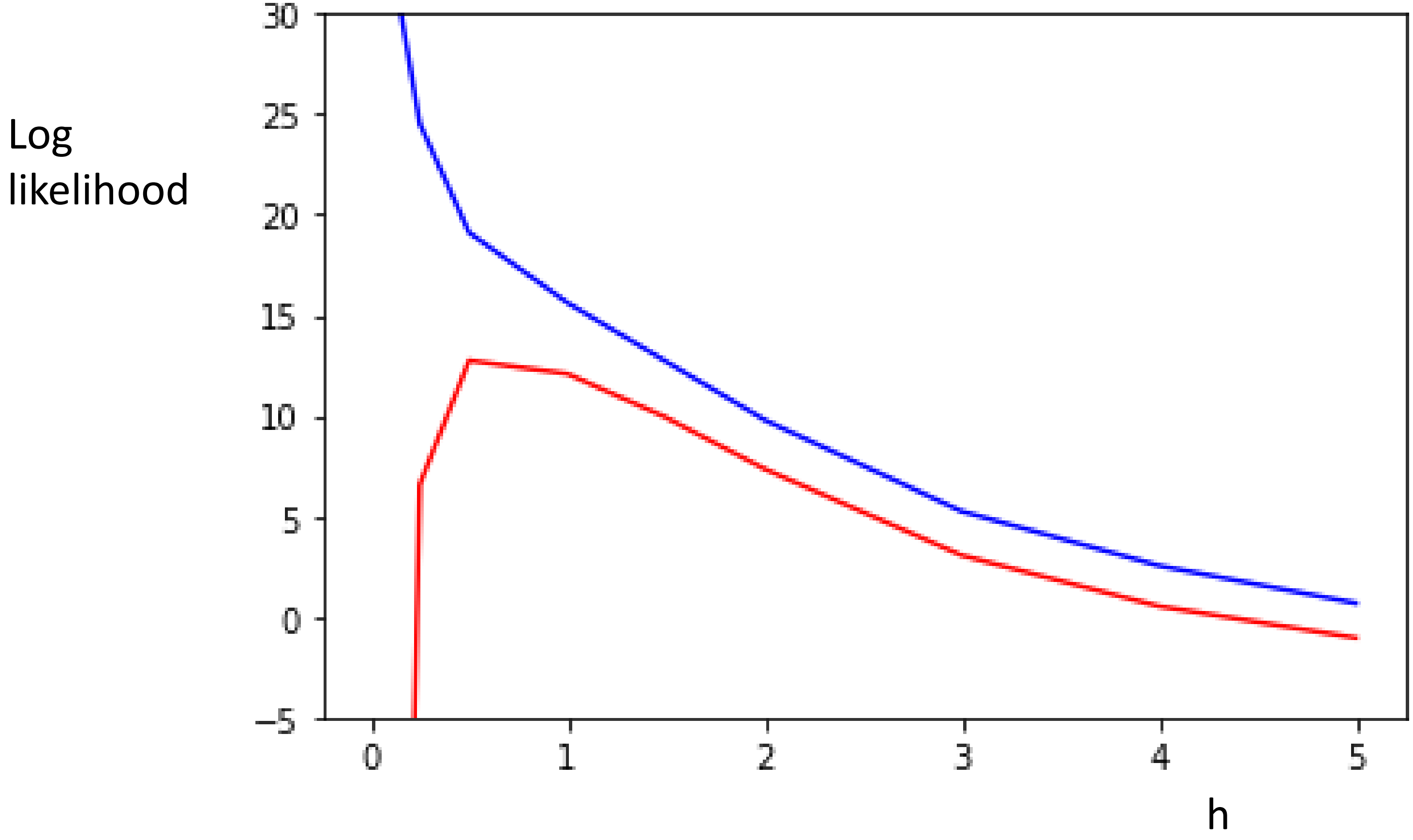
- To evaluate a fit of a density model to some data, -> define an error.
- Eg. use the log-likelihood:

$$LL(X) = \log\left(\prod_i \hat{p}(x_i)\right) = \sum_i log(\hat{p}(x_i))$$

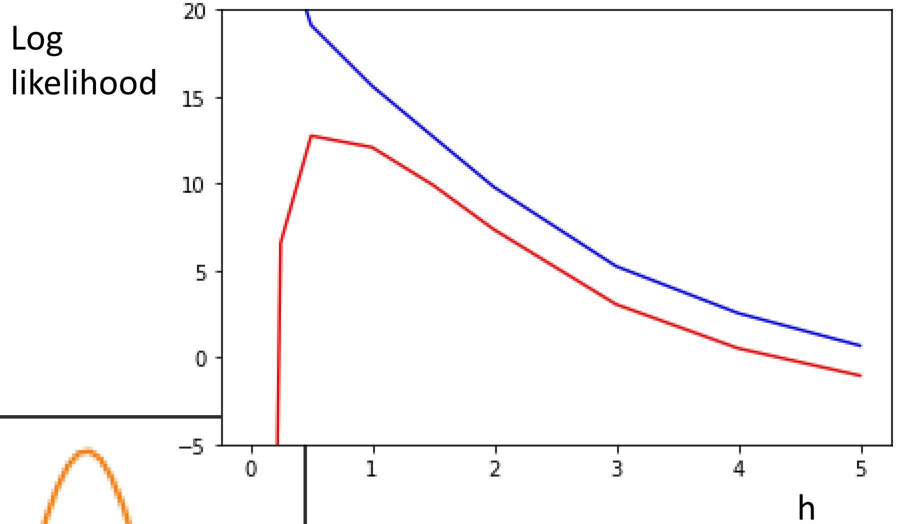
Loglikelihood vs. h for training set

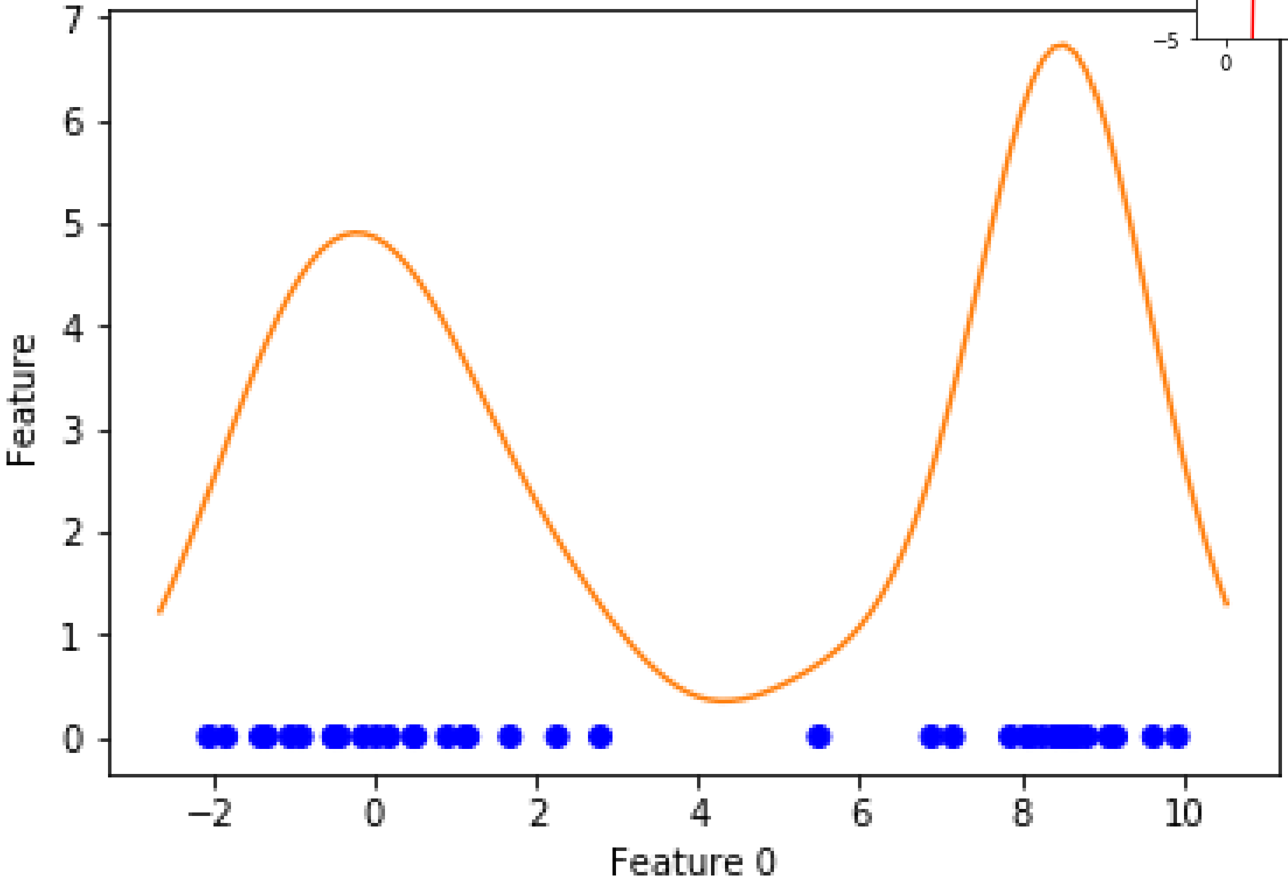


Loglikelihood vs. h for training set (blue) and test set (red)



h=0.9


Log
likelihood
20
15
10
5
0
−5
0
1
2
3
4
5
h


7
6
5
4
3
2
1
0
Feature
−2
0
2
4
6
8
10
Feature 0

TU Delft

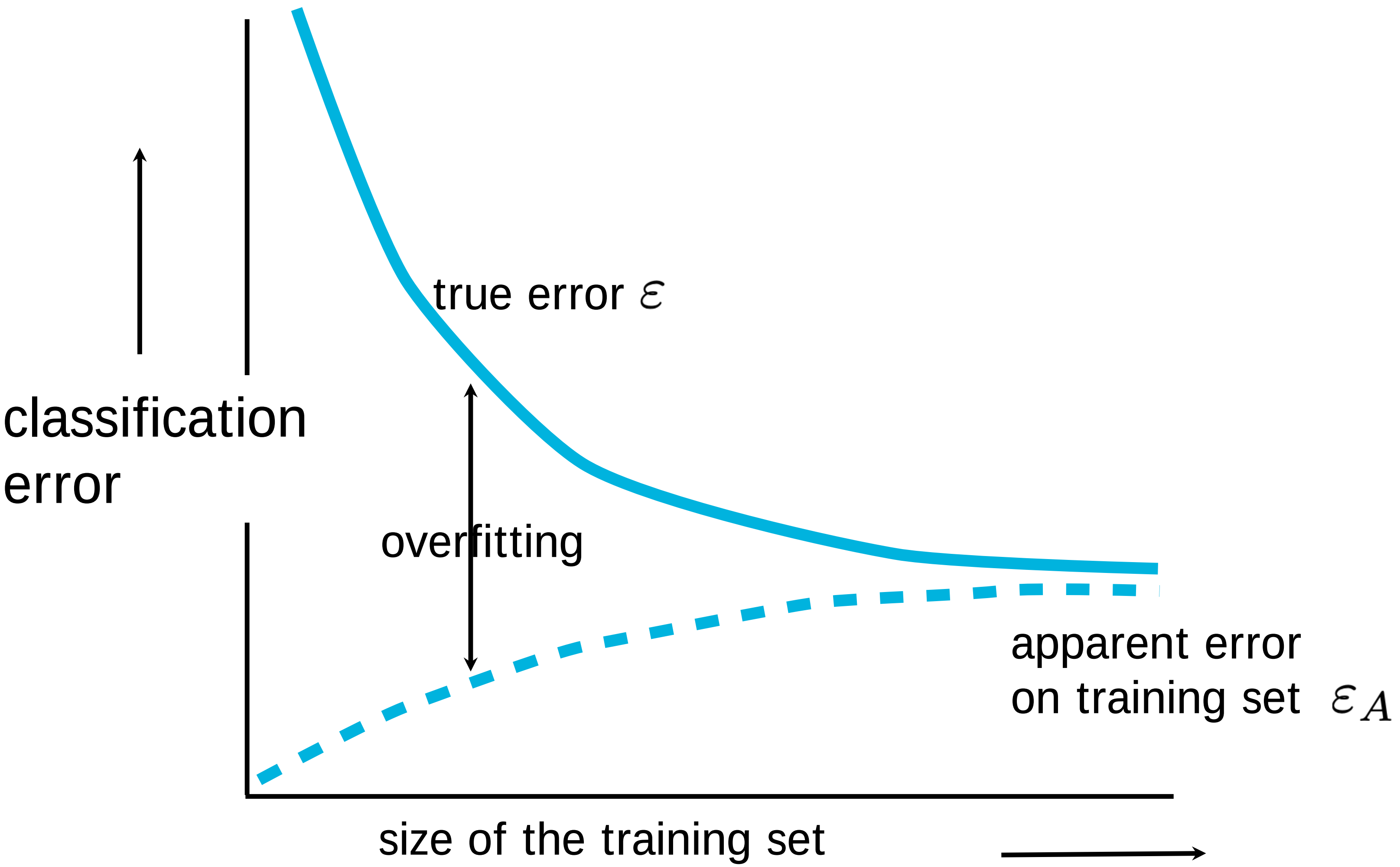
Learning curves

Learning curves

- Curves that plot [estimated] classification errors against the number of samples in training set
- Usually plot error both on training and on test set
- Gives insight in, e.g.
  – Amount of overtraining
  – Usefulness of additional data
  – How different classifiers compare

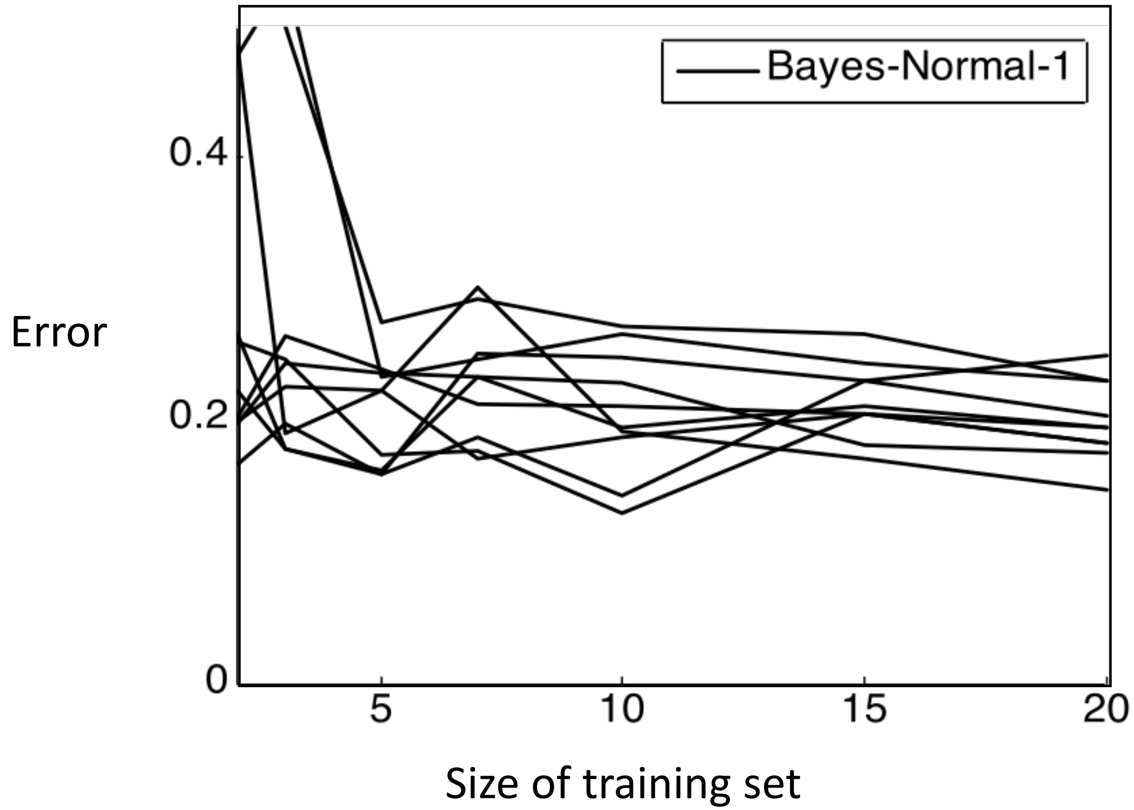
TU Delft

Apparent classification error



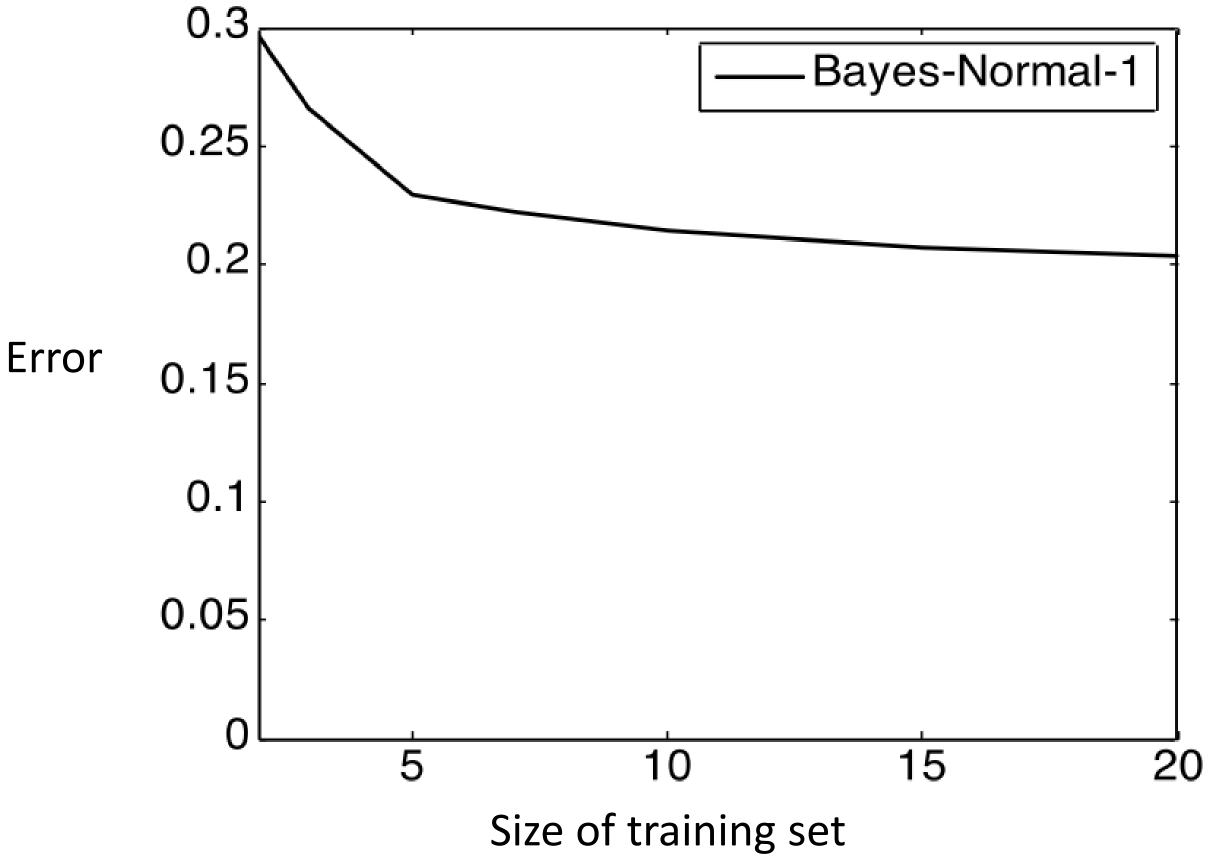
Repeated learning curves

- Small sample sizes have a very large variability

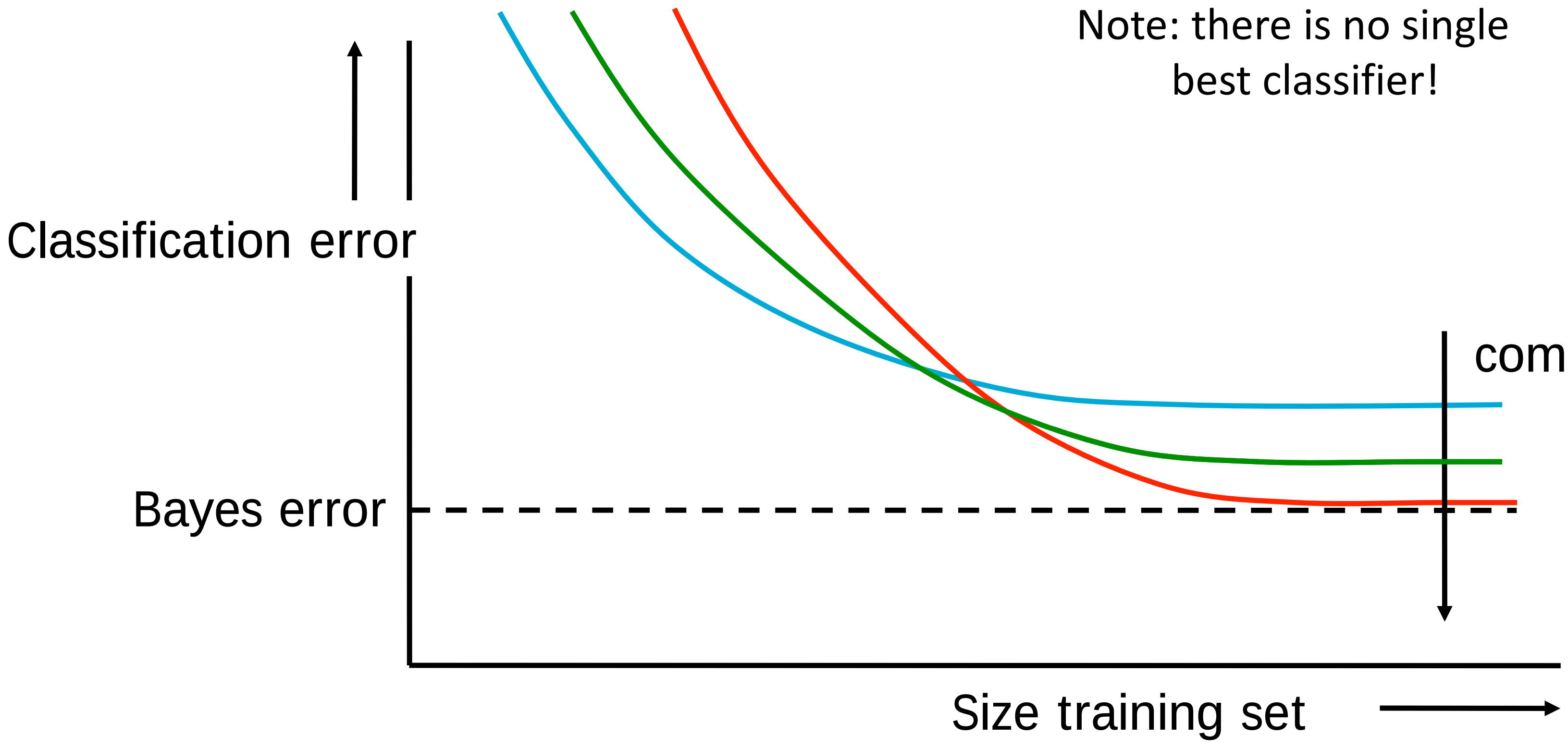


TU Delft

Averaged learning curve



Different classifier complexity



Fill in short evaluation

- One positive comment about the course
- One point of improvement
- Other remarks

- https://forms.gle/YAQtzDynSubZnvn28

Naïve Bayes classifier

Recap Bayes classifer

- For classification we need $p(y|x)$
- We can use Bayes' theorem if we can estimate $p(y)$ and $p(x|y)$

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

Recap Bayes classifier

- Assigning an object to the class with the maximum posterior probability gives the Bayes' classifier

$$p(x|y_1)p(y_1) > p(x|y_2)p(y_2)$$

- The Bayes' classifier is the optimal classifier
- The Bayes' error is the smallest error attainable

$$\varepsilon^* > \varepsilon$$

Warming up question

- Suppose we have trained a generative model and now get a new test example x. Our model tells us that: $p(x|y_o) = 0.01$, $p(x|y_1) = 0.03$ and $p(y_0) = p(y_1) = 0.5$
- What is $p(y_1|x)$?
  A. 0.015
  B. 0.25
  C. 0.75
  D. Insufficient information to compute. We also need to know the p(x).

Solution

- $p(y_1|x) = \frac{p(x|y_1)p(y_1)}{p(x)}$
- $p(x) = p(x|y_1)p(y_1) + p(x|y_0)p(y_0)$

- $p(y_1|x) = \frac{0.03*0.5}{0.03*0.5+0.01*0.5} = 0.75$

Density estimation

- So, we want to estimate a class probability density function:
$$p(x|y)$$
- Typically, each feature vector **x** has many features:
$$p(x|y) = p(x_1, x_2, x_3, x_4, \dots, x_d|y)$$

- To estimate this joint pdf (conditional on the class), we need LOTS of data... (curse of dimensionality)

Naive Bayes: Independence assumption

- Now assume, that all features are independent
- We assume conditional independence given y
- We just estimate $p(x_i|y)$ per feature and multiply them.

$$p(x|y) = p(x_1, x_2, x_3, x_4, \dots, x_d|y) = \prod_{i=1}^{d} p(x_i|y)$$

$$= p(x_1|y)p(x_2|y) \dots p(x_d)$$

- No curse of dimensionality!

Conditional independence example

- We assume conditional independence of two variables given a third variable.

- Probabilities of going to the beach and getting a heat stroke may be independent if we know the wheather is hot

$$p(B,S|H) = p(B|H)p(S|H)$$

- Hot weather “explains” all the dependence between beach and heartstroke
- In classification: class value explains all the dependence between attributes

Naive Bayes: Independence assumption

- Now assume, that all features are independent
- We assume conditional independence given y
- We just estimate $p(x_i|y)$ per feature and multiply them.

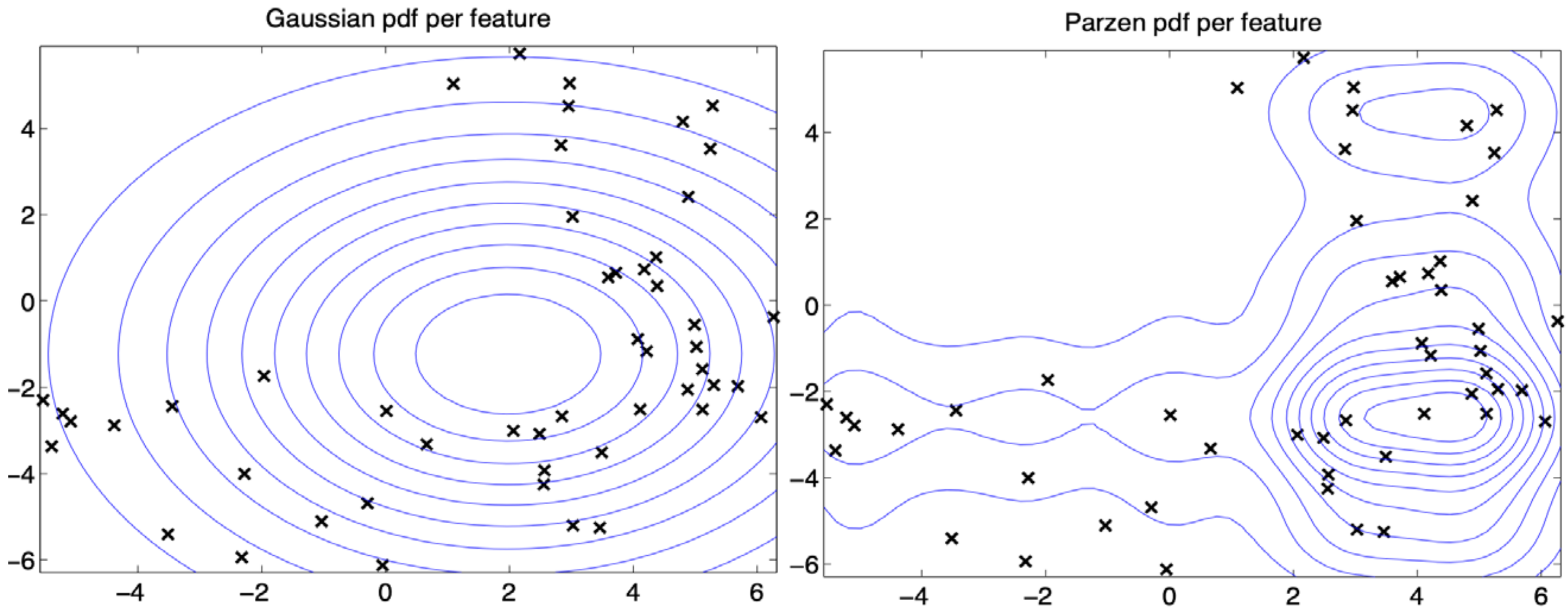
$$p(x|y) = p(x_1, x_2, x_3, x_4, \dots, x_d|y) = \prod_{i=1}^{d} p(x_i|y)$$

$$= p(x_1|y)p(x_2|y)\dots p(x_d|y)$$

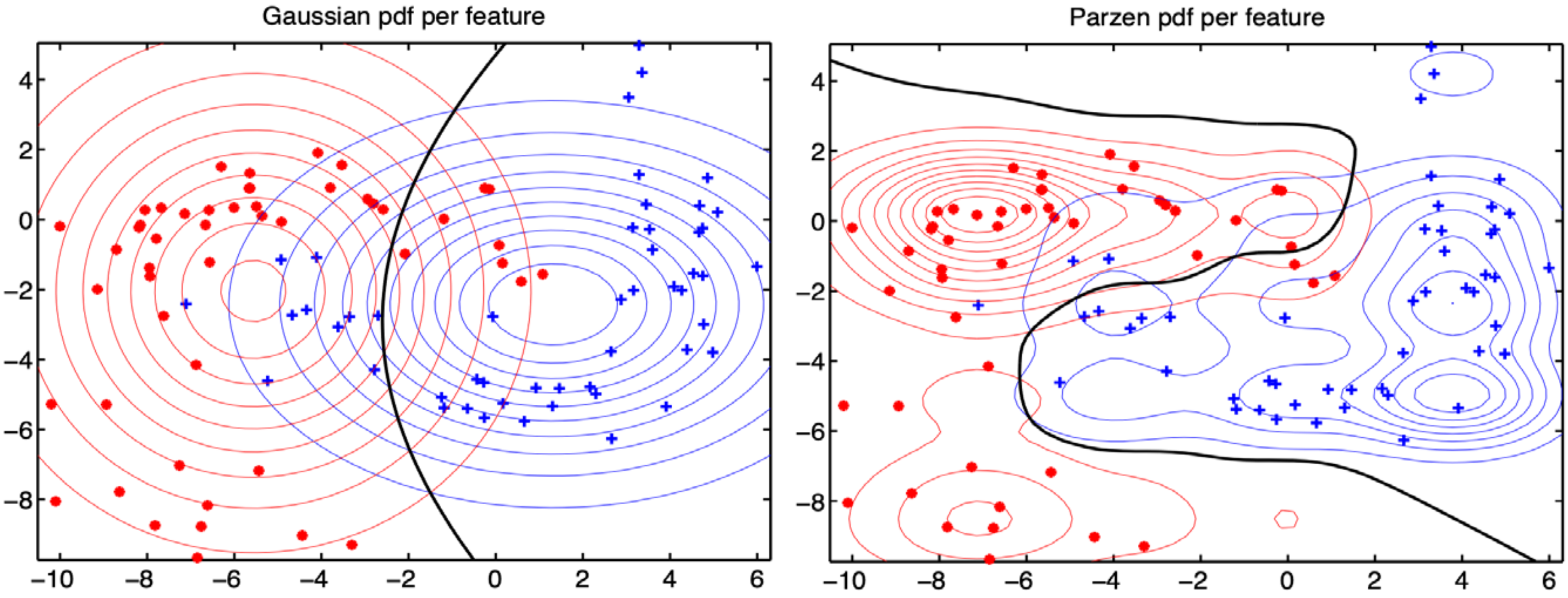
- No curse of dimensionality!

Parametric vs. non-parametric

- You still have to choose a model for $p(x_i|y)$



Naive Bayes classifier



Continuous data example

- Distinguish children from adults based on size
  - Classes: y = {a, c}, features: x = {height (cm), weight (kg)}
  - Training examples: 4 adults, 12 children
- Class probabailities $p(a) = \frac{4}{4+12} = 0.25$, $p(c) = 0.75$
- Model for adults:
  - Assume height and weight are independent
  - Height, estimate Gaussian with mean, variance

$$\begin{cases} \mu_{h,a} = \frac{1}{4} \sum_{i:y_i=a} h_i \\ \sigma_{h,a}^2 = \frac{1}{4} \sum_{i:y_i=a} (h_i - \mu_{h,a})^2 \end{cases}$$

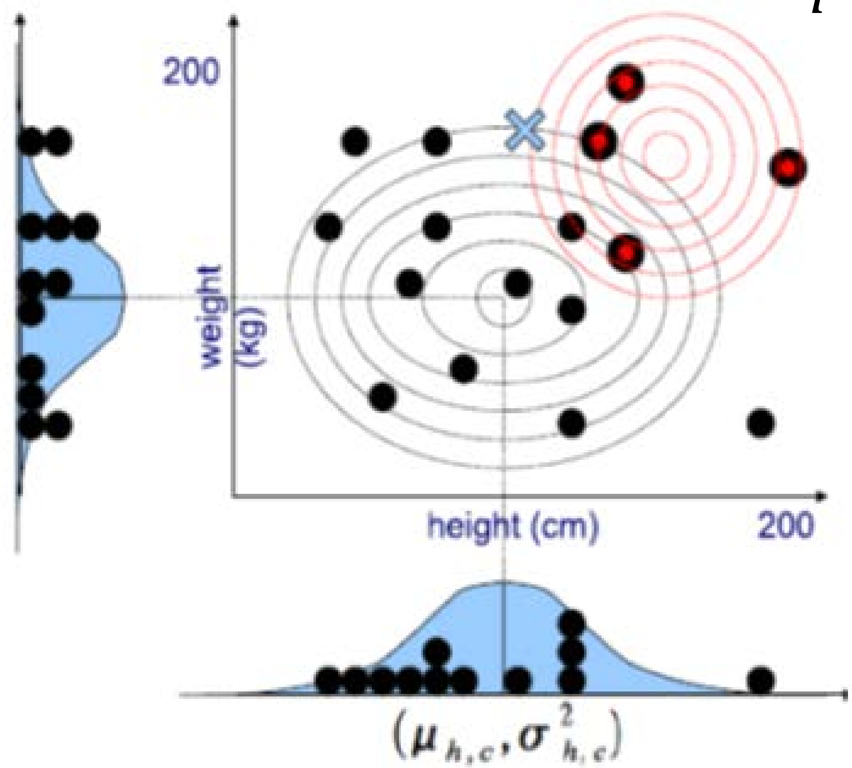
  - Weight, estimate Gaussian $(\mu_{w,a}, \sigma_{w,a}^2)$
- Model for children: use $(\mu_{h,c}, \sigma_{h,c}^2), (\mu_{w,c}, \sigma_{w,c}^2)$

TUDelft

Continuous example

$$p(w|a) = \frac{1}{\sqrt{2\pi\sigma_{w,a}^2}} exp - (\frac{w - \mu_{w,a}{}^2}{2\sigma_{w,a}^2})$$

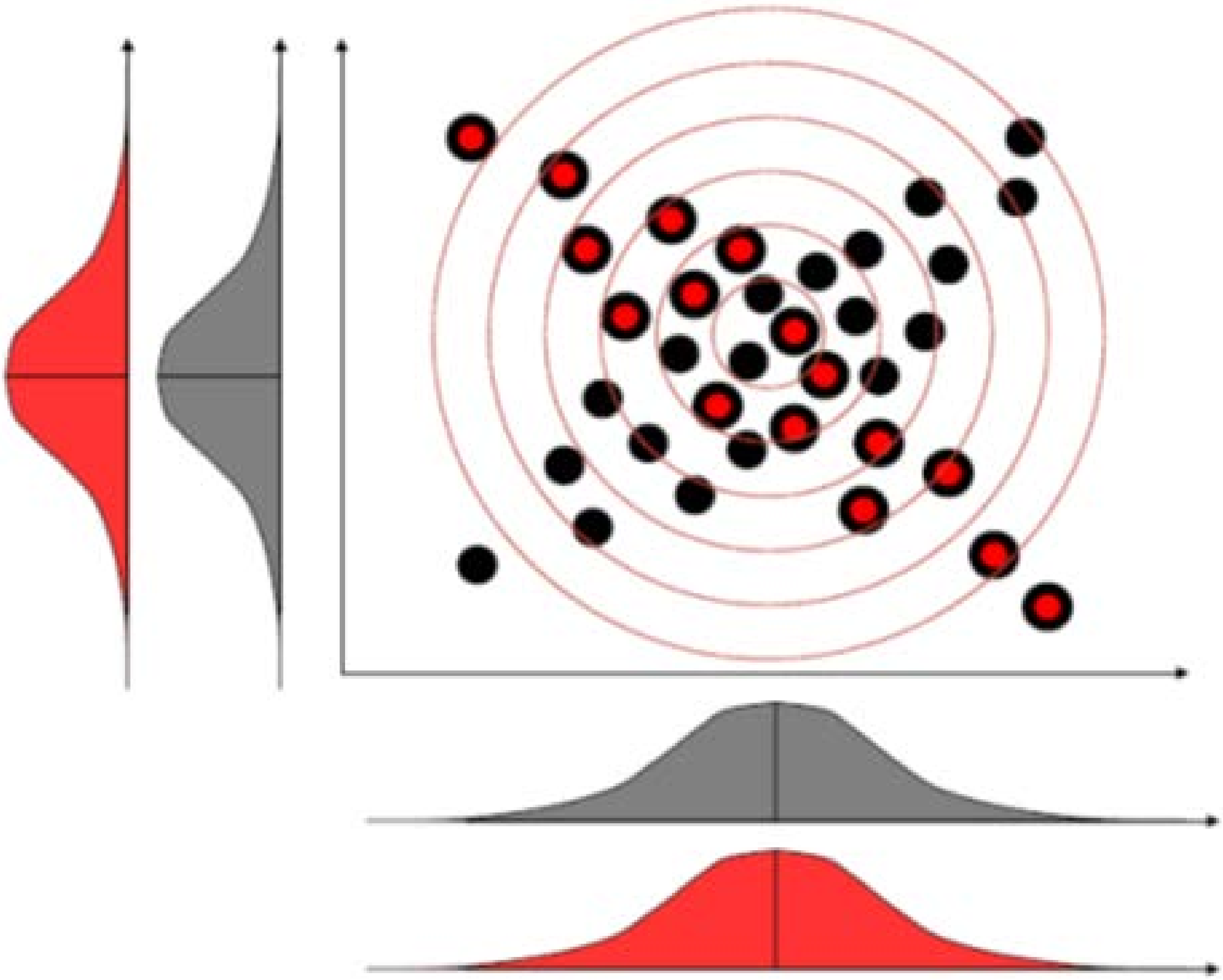
$$p(h|a) = \frac{1}{\sqrt{2\pi\sigma_{h,a}^2}} exp - (\frac{h - \mu_{h,a}{}^2}{2\sigma_{h,a}^2})$$



$$p(x|a) = p(w|a)p(h|a)$$

$$p(a|x) = \frac{p(x|a)p(a)}{p(x)}$$

Problems with Naive Bayes



Discrete example

- Separate spam from valid email (features = words)

D1: “send us your password”	spam
D2: “send us your review”	valid
D3: “review your password”	valid
D4: “review us”	spam
D5: “send your password”	spam
D6: “send us your account”	spam

p(spam) = 4/6 p(valid) = 2/6

	spam	valid
Password	2/4	½
Review	1/4	2/2
Send	3/4	½
Us	3/4	½
Your	3/4	½
Account	1/4	0/2

- New email “review us now”

TUDelft

Discrete example

p(spam) = 4/6 p(valid) = 2/6		
	spam	valid
Password	2/4	½
Review	1/4	2/2
Send	3/4	½
Us	3/4	½
Your	3/4	½
Account	1/4	0/2

- New email: “review us now”
- p(“review us”|spam) =
  p([0, 1, 0, 1, 0, 0]|spam) =
  $(1-\frac{2}{4})(\frac{1}{4})(1-\frac{3}{4})(\frac{3}{4})(1-\frac{3}{4})(1-\frac{1}{4})$ = 0.0044
- p(“review us”|valid) =
  p([0, 1, 0, 1, 0, 0]|valid) =
  $\left(1-\frac{1}{2}\right)\left(\frac{2}{2}\right)\left(1-\frac{1}{2}\right)\left(\frac{1}{2}\right)\left(1-\frac{1}{2}\right)\left(1-\frac{0}{2}\right)$ = 0.0625

TU Delft

Solution

p(spam) = 4/6 p(valid) = 2/6		
	spam	valid
Password	2/4	½
Review	1/4	2/2
Send	3/4	½
Us	3/4	½
Your	3/4	½
Account	1/4	0/2

- p("review us"|spam) = 0.0044
- p("review us"|valid) = 0.0625

- p("review us"|spam)p(spam) = 0.0044 * 4/6 = 0.0029
- p("review us"|valid)p(valid) = 0.0625 * 2/6 = 0.02

- Note: identical example!

Zero frequency problem

p(spam) = 4/6 p(valid) = 2/6		
	spam	valid
Password	2/4	½
Review	1/4	2/2
Send	3/4	½
Us	3/4	½
Your	3/4	½
Account	1/4	0/2

- Any email containing "account" is spam
  - p("account"|valid) = 0/2
- Solution: never allow zero probabilities
  - Laplace smoothing: add a small positive number to the counts (K-> number of classes)

$$\mathrm{p}(w|c) = \frac{num(w,c) + \varepsilon}{num(c) + K\varepsilon}$$

  - May use global statistics in place of ε: num(w)/num
  - Very common problem (50% of words occure once)

Fooling Naive Bayes

- Every word contributes independently to p(spam|email)
- Add lots of valid words into spam email.

Missing data

- Suppose we don't have value for some attribute x_j
  - Eg. some medical test not performed on patient
- How to compute $p(x_1, \ldots, x_j, \ldots x_d | y)$
- Easy with Naive Bayes
  - Ignore attribute instance where it's missing a value
  - Compute likelihood based on observed values
  - No need to fill in or explicitly model missing values
  - Based on conditional independence between attributes

$$P\left(x_1, \ldots, x_j, \ldots, x_d\right) = \prod_{i \neq j}^{d} p(x_i | y)$$

TUDelft

Missing data example

- Three coin tosses: event = $\{x_1 = H, x_2 =?, x_3 = T\}$
  - Event: head, unknown (either tail ot head), tail
  - event = {H, H, T} + {H, T, T}
  - P(event) = P(H, H, T) + P(H, T, T)
- General case: x_j has missing value

  $$p(x_1, \dots, \boxed{x_j}, \dots x_d | y) = p(x_1|y) \dots \boxed{p(x_j|y)} \dots p(x_d|y)$$

- $$\textstyle\sum_{x_j} p(x_1, \dots, \boxed{x_j}, \dots x_d | y) =$$

  $$\textstyle\sum_{x_j} p(x_1|y) \dots \boxed{p(x_j|y)} \dots p(x_d|y) =$$

  $$p(x_1|y) \dots \boxed{\left[\textstyle\sum_{x_j} p(x_j|y)\right]} \dots p(x_d|y) =$$

  $$p(x_1|y) \dots \boxed{[1]} \dots p(x_d|y)$$

TU Delft

Naive Bayes pros and cons

- Can handle high dimensional feature spaces
- Fast training time
- Can handle missing values
- Transparent

- Can’t deal with correlated features

After practicing with the concept of this lecture you should be able to:

- Explain what are and how to use the learning curves
- Explain the Naive Bayes classifier, including the following:
  - components and their function
  - independence assumption
  - dealing with missing data
  - Continuous example
  - Discrete example
  - Pros and cons

TU Delft

Questions to think about

- Is feature scaling an issue for Naive Bayes?
- How would the learning curve look like for a very simple classifier, like nearest mean?
- Which classifier doesn't make 0 training error when we have 1 object per class? K-nn, Parzen, Nearest mean, LDA, QDA, Naive Bayes?

Exercise Naive Bayes

- Predict if Bob will default his loan

<u>Bob:</u>

Homeowner: no

Maritial status: married

Job experience: 3

Home owner	Maritial status	Job experience	Deafulted
Yes	Single	3	No
No	Married	4	No
No	Single	5	No
Yes	Married	4	No
No	Divorced	2	Yes
No	Married	4	No
Yes	Divorced	2	No
No	Married	3	Yes
No	Married	3	No
Yes	Single	2	Yes

TUDelft