

RWTH AACHEN
UNIVERSITY

Machine Learning – Lecture 2

Probability Density Estimation

16.10.2017

Bastian Leibe
RWTH Aachen
<http://www.vision.rwth-aachen.de>
leibe@vision.rwth-aachen.de

Machine Learning Winter '17

RWTH AACHEN
UNIVERSITY

Announcements

- Exceptional number of lecture participants this year
 - Current count: 449 participants
 - This is very nice, but it stretches our resources to their limits
- Monday lecture slot
 - Shifted to 8:30 – 10:00 in AH IV (276 seats)
 - We will monitor the situation and take further action if the space is not sufficient
- Thursday lecture slot
 - Will stay at 14:15 – 15:45 in H02 (C.A.R.L, 786 seats)
- Exercises (non-mandatory)
 - We will try to offer corrections, but we will have to see how to handle those numbers...

B. Leibe 2

RWTH AACHEN
UNIVERSITY

Announcements

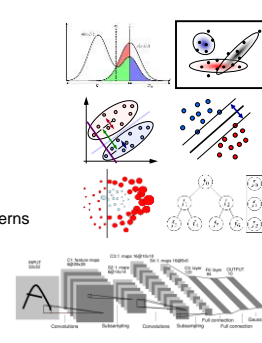
- L2P electronic repository
 - Slides, exercises, and supplementary material will be made available here
 - Lecture recordings will be uploaded 2-3 days after the lecture
- Course webpage
 - <http://www.vision.rwth-aachen.de/courses/>
 - Slides will also be made available on the webpage
- Please subscribe to the lecture on the Campus system!
 - Important to get email announcements and L2P access!

B. Leibe 3

RWTH AACHEN
UNIVERSITY

Course Outline

- Fundamentals
 - Bayes Decision Theory
 - Probability Density Estimation
- Classification Approaches
 - Linear Discriminants
 - Support Vector Machines
 - Ensemble Methods & Boosting
 - Randomized Trees, Forests & Ferns
- Deep Learning
 - Foundations
 - Convolutional Neural Networks
 - Recurrent Neural Networks



B. Leibe 4

RWTH AACHEN
UNIVERSITY


Topics of This Lecture

- Bayes Decision Theory
 - Basic concepts
 - Minimizing the misclassification rate
 - Minimizing the expected loss
 - Discriminant functions
- Probability Density Estimation
 - General concepts
 - Gaussian distribution
- Parametric Methods
 - Maximum Likelihood approach
 - Bayesian vs. Frequentist views on probability

B. Leibe 5

RWTH AACHEN
UNIVERSITY

Bayes Decision Theory



Thomas Bayes, 1701-1761

"The theory of inverse probability is founded upon an error, and must be wholly rejected."

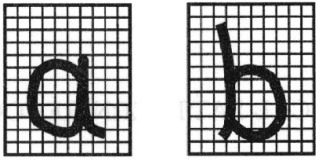
R.A. Fisher, 1925

B. Leibe 14
Image source: Wikipedia

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

- Example: handwritten character recognition



- Goal:
 - Classify a new letter such that the probability of misclassification is minimized.

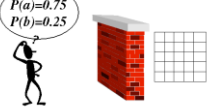
Machine Learning Winter '17 | Slide credit: Bernt Schiele | B. Leibe | Image source: C.M. Bishop, 2004 | 15

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

- Concept 1: **Priors** (a priori probabilities) $p(C_k)$
 - What we can tell about the probability *before seeing the data*.
 - Example:

$p(a)=0.75$
 $p(b)=0.25$



?

$C_1 = a$
 $C_2 = b$

$p(C_1) = 0.75$
 $p(C_2) = 0.25$


- In general: $\sum_k p(C_k) = 1$

Machine Learning Winter '17 | Slide credit: Bernt Schiele | B. Leibe | 16

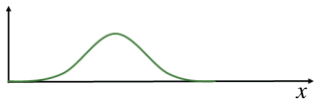
RWTH AACHEN UNIVERSITY


Bayes Decision Theory

- Concept 2: **Conditional probabilities** $p(x|C_k)$
 - Let x be a feature vector.
 - x measures/describes certain properties of the input.
 - E.g. number of black pixels, aspect ratio, ...
 - $p(x|C_k)$ describes its **likelihood** for class C_k .

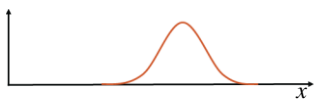


$p(x|a)$





$p(x|b)$

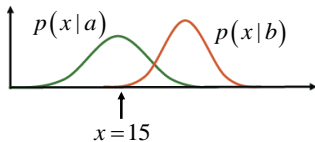


Machine Learning Winter '17 | Slide credit: Bernt Schiele | B. Leibe | 17

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

- Example:

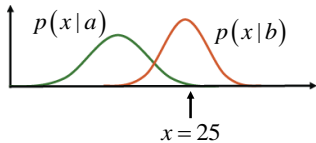

- Question:
 - Which class?
 - Since $p(x|b)$ is much smaller than $p(x|a)$, the decision should be 'a' here.

Machine Learning Winter '17 | Slide credit: Bernt Schiele | B. Leibe | 18

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

- Example:

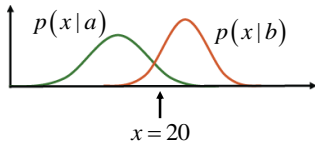

- Question:
 - Which class?
 - Since $p(x|a)$ is much smaller than $p(x|b)$, the decision should be 'b' here.

Machine Learning Winter '17 | Slide credit: Bernt Schiele | B. Leibe | 19

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

- Example:


- Question:
 - Which class?
 - Remember that $p(a) = 0.75$ and $p(b) = 0.25$...
 - I.e., the decision should be again 'a'.
 - ⇒ How can we formalize this?

Machine Learning Winter '17 | Slide credit: Bernt Schiele | B. Leibe | 20

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

- Concept 3: **Posterior probabilities** $p(C_k | x)$
 - We are typically interested in the *a posteriori* probability, i.e. the probability of class C_k given the measurement vector x .
- Bayes' Theorem:

$$p(C_k | x) = \frac{p(x | C_k) p(C_k)}{p(x)} = \frac{p(x | C_k) p(C_k)}{\sum_i p(x | C_i) p(C_i)}$$
- Interpretation

$$\text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalization Factor}}$$

Machine Learning Winter '17 21

Slide credit: Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

Machine Learning Winter '17 22

Slide credit: Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

Bayesian Decision Theory

- Goal: **Minimize the probability of a misclassification**

Machine Learning Winter '17 23

Image source: C.M. Bishop, 2006 B. Leibe

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

- Optimal decision rule
 - Decide for C_1 if

$$p(C_1 | x) > p(C_2 | x)$$
 - This is equivalent to

$$p(x | C_1) p(C_1) > p(x | C_2) p(C_2)$$
 - Which is again equivalent to (**Likelihood-Ratio test**)

$$\frac{p(x | C_1)}{p(x | C_2)} > \underbrace{\frac{p(C_2)}{p(C_1)}}_{\text{Decision threshold } \theta}$$

Machine Learning Winter '17 24

Slide credit: Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

Generalization to More Than 2 Classes

- Decide for class k whenever it has the greatest posterior probability of all classes:

$$p(C_k | x) > p(C_j | x) \quad \forall j \neq k$$

$$p(x | C_k) p(C_k) > p(x | C_j) p(C_j) \quad \forall j \neq k$$
- Likelihood-ratio test

$$\frac{p(x | C_k)}{p(x | C_j)} > \frac{p(C_j)}{p(C_k)} \quad \forall j \neq k$$

Machine Learning Winter '17 25

Slide credit: Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

Classifying with Loss Functions

- Generalization to decisions with a **loss function**
 - Differentiate between the possible decisions and the possible true classes.
 - Example: medical diagnosis
 - Decisions: *sick or healthy* (or: *further examination necessary*)
 - Classes: *patient is sick or healthy*
 - The cost may be asymmetric:

$$\text{loss}(\text{decision} = \text{healthy} | \text{patient} = \text{sick}) \gg \text{loss}(\text{decision} = \text{sick} | \text{patient} = \text{healthy})$$

Machine Learning Winter '17 26

Slide credit: Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

Classifying with Loss Functions

- In general, we can formalize this by introducing a loss matrix L_{kj}

$$L_{kj} = \text{loss for decision } C_j \text{ if truth is } C_k.$$

- Example: cancer diagnosis

		Decision	
		cancer	normal
Truth	cancer	0	1000
	normal	1	0

$$L_{\text{cancer diagnosis}} = \begin{matrix} & \text{Decision} \\ & \text{cancer} & \text{normal} \\ \text{Truth} & \text{cancer} & \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \\ & \text{normal} & \end{matrix}$$

27

RWTH AACHEN UNIVERSITY

Classifying with Loss Functions

- Loss functions may be different for different actors.

- Example:

	"invest"	"don't invest"	
$L_{\text{stocktrader}}(\text{subprime}) =$	$\begin{pmatrix} -\frac{1}{2}C_{\text{gain}} & 0 \\ 0 & 0 \end{pmatrix}$		
- | | | | |
|--------------------------------------|---|--|--|
| $L_{\text{bank}}(\text{subprime}) =$ | $\begin{pmatrix} -\frac{1}{2}C_{\text{gain}} & 0 \\ \text{skull} & 0 \end{pmatrix}$ | | |
|--------------------------------------|---|--|--|

⇒ Different loss functions may lead to different Bayes optimal strategies.

28

RWTH AACHEN UNIVERSITY

Minimizing the Expected Loss

- Optimal solution is the one that minimizes the loss.
 - But: loss function depends on the true class, which is unknown.
- Solution: **Minimize the expected loss**

$$\mathbb{E}[L] = \sum_k \sum_j \int_{\mathcal{R}_j} L_{kj} p(\mathbf{x}, C_k) d\mathbf{x}$$

- This can be done by choosing the regions \mathcal{R}_j such that

$$\mathbb{E}[L] = \sum_k L_{kj} p(C_k | \mathbf{x})$$

which is easy to do once we know the posterior class probabilities $p(C_k | \mathbf{x})$

29

RWTH AACHEN UNIVERSITY

Minimizing the Expected Loss

- Example:
 - 2 Classes: C_1, C_2
 - 2 Decision: α_1, α_2
 - Loss function: $L(\alpha_j | C_k) = L_{kj}$
- Expected loss (= risk R) for the two decisions:

$$\mathbb{E}_{\alpha_1}[L] = R(\alpha_1 | \mathbf{x}) = L_{11}p(C_1 | \mathbf{x}) + L_{21}p(C_2 | \mathbf{x})$$

$$\mathbb{E}_{\alpha_2}[L] = R(\alpha_2 | \mathbf{x}) = L_{12}p(C_1 | \mathbf{x}) + L_{22}p(C_2 | \mathbf{x})$$
- Goal: Decide such that expected loss is minimized
 - I.e. decide α_1 if $R(\alpha_2 | \mathbf{x}) > R(\alpha_1 | \mathbf{x})$

30

RWTH AACHEN UNIVERSITY

Minimizing the Expected Loss

$$R(\alpha_2 | \mathbf{x}) > R(\alpha_1 | \mathbf{x})$$

$$L_{12}p(C_1 | \mathbf{x}) + L_{22}p(C_2 | \mathbf{x}) > L_{11}p(C_1 | \mathbf{x}) + L_{21}p(C_2 | \mathbf{x})$$

$$(L_{12} - L_{11})p(C_1 | \mathbf{x}) > (L_{21} - L_{22})p(C_2 | \mathbf{x})$$

$$\frac{(L_{12} - L_{11})}{(L_{21} - L_{22})} > \frac{p(C_2 | \mathbf{x})}{p(C_1 | \mathbf{x})} = \frac{p(\mathbf{x} | C_2)p(C_2)}{p(\mathbf{x} | C_1)p(C_1)}$$

$$\frac{p(\mathbf{x} | C_1)}{p(\mathbf{x} | C_2)} > \frac{(L_{21} - L_{22})p(C_2)}{(L_{12} - L_{11})p(C_1)}$$

⇒ Adapted decision rule taking into account the loss.

31

RWTH AACHEN UNIVERSITY

The Reject Option

- Classification errors arise from regions where the largest posterior probability $p(C_k | \mathbf{x})$ is significantly less than 1.
 - These are the regions where we are relatively uncertain about class membership.
 - For some applications, it may be better to reject the automatic decision entirely in such a case and e.g. consult a human expert.

32

RWTH AACHEN UNIVERSITY

Discriminant Functions

- Formulate classification in terms of comparisons
 - Discriminant functions

$$y_1(x), \dots, y_K(x)$$
 - Classify x as class C_k if

$$y_k(x) > y_j(x) \quad \forall j \neq k$$
- Examples (Bayes Decision Theory)

$$y_k(x) = p(C_k|x)$$

$$y_k(x) = p(x|C_k)p(C_k)$$

$$y_k(x) = \log p(x|C_k) + \log p(C_k)$$

Machine Learning Winter '17 33

Slide credit: Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

Different Views on the Decision Problem

- $y_k(x) \propto p(x|C_k)p(C_k)$
 - First determine the class-conditional densities for each class individually and separately infer the prior class probabilities.
 - Then use Bayes' theorem to determine class membership.
 \Rightarrow *Generative methods*
- $y_k(x) = p(C_k|x)$
 - First solve the inference problem of determining the posterior class probabilities.
 - Then use decision theory to assign each new x to its class.
 \Rightarrow *Discriminative methods*
- Alternative
 - Directly find a discriminant function $y_k(x)$ which maps each input x directly onto a class label.

Machine Learning Winter '17 34

B. Leibe

RWTH AACHEN UNIVERSITY

Topics of This Lecture

- Bayes Decision Theory
 - Basic concepts
 - Minimizing the misclassification rate
 - Minimizing the expected loss
 - Discriminant functions
- Probability Density Estimation
 - General concepts
 - Gaussian distribution
- Parametric Methods
 - Maximum Likelihood approach
 - Bayesian vs. Frequentist views on probability
 - Bayesian Learning

Machine Learning Winter '17 35

B. Leibe

RWTH AACHEN UNIVERSITY

Probability Density Estimation

- Up to now
 - Bayes optimal classification
 - Based on the probabilities $p(\mathbf{x}|C_k)p(C_k)$
- How can we estimate (=learn) those probability densities?
 - Supervised training case: data and class labels are known.
 - Estimate the probability density for each class C_k separately:

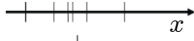
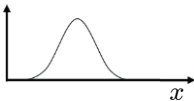
$$p(\mathbf{x}|C_k)$$
 - (For simplicity of notation, we will drop the class label C_k in the following.)

Machine Learning Winter '17 36

Slide credit: Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

Probability Density Estimation

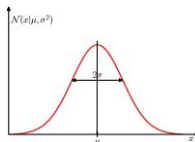
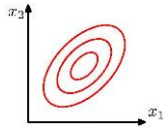
- Data: $x_1, x_2, x_3, x_4, \dots$

- Estimate: $p(x)$

- Methods
 - Parametric representations (today)
 - Non-parametric representations (lecture 3)
 - Mixture models (lecture 4)

Machine Learning Winter '17 37

Slide credit: Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

The Gaussian (or Normal) Distribution

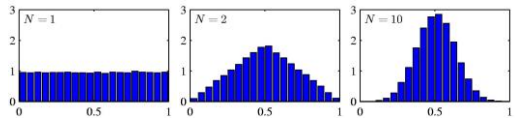
- One-dimensional case
 - Mean μ
 - Variance σ^2
$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

- Multi-dimensional case
 - Mean μ
 - Covariance Σ
$$\mathcal{N}(\mathbf{x}|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\mu)^T \Sigma^{-1}(\mathbf{x}-\mu)\right\}$$


Machine Learning Winter '17 38

B. Leibe Image source: C.M. Bishop, 2006

RWTH AACHEN UNIVERSITY

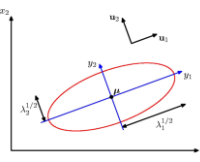
Gaussian Distribution – Properties

- **Central Limit Theorem**
 - “The distribution of the sum of N i.i.d. random variables becomes increasingly Gaussian as N grows.”
 - In practice, the convergence to a Gaussian can be very rapid.
 - This makes the Gaussian interesting for many applications.
- **Example: N uniform [0,1] random variables.**


39
B. Leibe Image source: C.M. Bishop, 2006

RWTH AACHEN UNIVERSITY

Gaussian Distribution – Properties

- **Quadratic Form**
 - \mathcal{N} depends on \mathbf{x} through the exponent
 - $$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$
 - Here, Δ is often called the **Mahalanobis distance** from \mathbf{x} to $\boldsymbol{\mu}$.
- **Shape of the Gaussian**
 - $\boldsymbol{\Sigma}$ is a real, symmetric matrix.
 - We can therefore decompose it into its eigenvectors
 - $$\boldsymbol{\Sigma} = \sum_{i=1}^D \lambda_i \mathbf{u}_i \mathbf{u}_i^T \quad \boldsymbol{\Sigma}^{-1} = \sum_{i=1}^D \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$
 - and thus obtain $\Delta^2 = \sum_{i=1}^D \frac{y_i^2}{\lambda_i}$ with $y_i = \mathbf{u}_i^T (\mathbf{x} - \boldsymbol{\mu})$

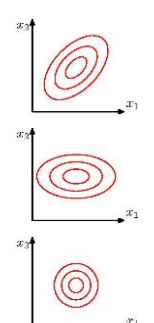
⇒ **Constant density on ellipsoids** with main directions along the eigenvectors \mathbf{u}_i , and scaling factors $\sqrt{\lambda_i}$

40
Image source: C.M. Bishop, 2006

RWTH AACHEN UNIVERSITY

Gaussian Distribution – Properties

- **Special cases**
 - Full covariance matrix
 $\boldsymbol{\Sigma} = [\sigma_{ij}]$
⇒ General ellipsoid shape
 - Diagonal covariance matrix
 $\boldsymbol{\Sigma} = \text{diag}\{\sigma_i\}$
⇒ Axis-aligned ellipsoid
 - Uniform variance
 $\boldsymbol{\Sigma} = \sigma^2 \mathbf{I}$
⇒ Hypersphere

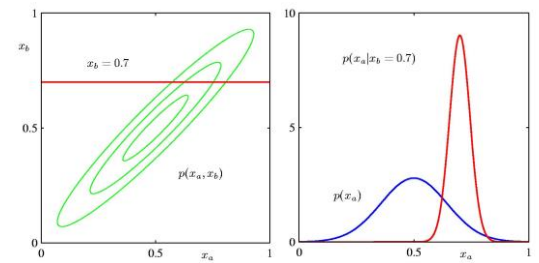


41
B. Leibe Image source: C.M. Bishop, 2006

RWTH AACHEN UNIVERSITY

Gaussian Distribution – Properties

- The marginals of a Gaussian are again Gaussians:



42
B. Leibe Image source: C.M. Bishop, 2006

RWTH AACHEN UNIVERSITY

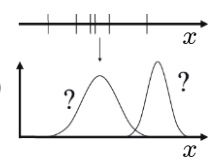
Topics of This Lecture

- **Bayes Decision Theory**
 - Basic concepts
 - Minimizing the misclassification rate
 - Minimizing the expected loss
 - Discriminant functions
- **Probability Density Estimation**
 - General concepts
 - Gaussian distribution
- **Parametric Methods**
 - Maximum Likelihood approach
 - Bayesian vs. Frequentist views on probability
 - Bayesian Learning

43
B. Leibe

RWTH AACHEN UNIVERSITY

Parametric Methods

- **Given**
 - Data $X = \{x_1, x_2, \dots, x_N\}$
 - Parametric form of the distribution with parameters θ
 - E.g. for Gaussian distrib.: $\theta = (\mu, \sigma)$
- **Learning**
 - Estimation of the parameters θ
- **Likelihood of θ**
 - Probability that the data X have indeed been generated from a probability density with parameters θ
 - $$L(\theta) = p(X|\theta)$$

44
Slide adapted from Bernt Schiele B. Leibe

RWTH AACHEN UNIVERSITY

Maximum Likelihood Approach

- Computation of the likelihood
 - Single data point: $p(x_n|\theta) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$
 - Assumption: all data points are independent

$$L(\theta) = p(X|\theta) = \prod_{n=1}^N p(x_n|\theta)$$

- Log-likelihood

$$E(\theta) = -\ln L(\theta) = -\sum_{n=1}^N \ln p(x_n|\theta)$$

- Estimation of the parameters θ (Learning)
 - Maximize the likelihood
 - Minimize the negative log-likelihood

Machine Learning Winter '17 | Slide credit: Bernt Schiele | B. Leibe | 45

RWTH AACHEN UNIVERSITY

Maximum Likelihood Approach

- Likelihood: $L(\theta) = p(X|\theta) = \prod_{n=1}^N p(x_n|\theta)$
- We want to obtain $\hat{\theta}$ such that $L(\hat{\theta})$ is maximized.

Machine Learning Winter '17 | Slide credit: Bernt Schiele | B. Leibe | 46

RWTH AACHEN UNIVERSITY

Maximum Likelihood Approach

- Minimizing the log-likelihood
 - How do we minimize a function?
 - ⇒ Take the derivative and set it to zero.

$$\frac{\partial}{\partial \theta} E(\theta) = -\frac{\partial}{\partial \theta} \sum_{n=1}^N \ln p(x_n|\theta) = -\sum_{n=1}^N \frac{\frac{\partial}{\partial \theta} p(x_n|\theta)}{p(x_n|\theta)} \stackrel{!}{=} 0$$

- Log-likelihood for Normal distribution (1D case)

$$E(\theta) = -\sum_{n=1}^N \ln p(x_n|\mu, \sigma)$$

$$= -\sum_{n=1}^N \ln \left(\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{\|x_n - \mu\|^2}{2\sigma^2}\right\} \right)$$

Machine Learning Winter '17 | B. Leibe | 47

RWTH AACHEN UNIVERSITY

Maximum Likelihood Approach

- Minimizing the log-likelihood

$$\frac{\partial}{\partial \mu} E(\mu, \sigma) = -\sum_{n=1}^N \frac{\frac{\partial}{\partial \mu} p(x_n|\mu, \sigma)}{p(x_n|\mu, \sigma)}$$

$$= -\sum_{n=1}^N \frac{2(x_n - \mu)}{2\sigma^2}$$

$$= \frac{1}{\sigma^2} \sum_{n=1}^N (x_n - \mu)$$

$$= \frac{1}{\sigma^2} \left(\sum_{n=1}^N x_n - N\mu \right)$$

$$\frac{\partial}{\partial \mu} E(\mu, \sigma) \stackrel{!}{=} 0 \quad \Leftrightarrow \quad \hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n$$

$$p(x_n|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\|x_n - \mu\|^2}{2\sigma^2}}$$

Machine Learning Winter '17 | B. Leibe | 48

RWTH AACHEN UNIVERSITY

Maximum Likelihood Approach

- We thus obtain

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^N x_n \quad \text{"sample mean"}$$

- In a similar fashion, we get

$$\hat{\sigma}^2 = \frac{1}{N} \sum_{n=1}^N (x_n - \hat{\mu})^2 \quad \text{"sample variance"}$$

- $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ is the **Maximum Likelihood estimate** for the parameters of a Gaussian distribution.
- This is a very important result.
- Unfortunately, it is wrong...

Machine Learning Winter '17 | B. Leibe | 49

RWTH AACHEN UNIVERSITY

Maximum Likelihood Approach

- Or not wrong, but rather **biased**...
- Assume the samples x_1, x_2, \dots, x_N come from a true Gaussian distribution with mean μ and variance σ^2
 - We can now compute the expectations of the ML estimates with respect to the data set values. It can be shown that

$$\mathbb{E}(\mu_{ML}) = \mu$$

$$\mathbb{E}(\sigma_{ML}^2) = \left(\frac{N-1}{N}\right) \sigma^2$$

⇒ The ML estimate will underestimate the true variance.

- Corrected estimate:

$$\hat{\sigma}^2 = \frac{N}{N-1} \sigma_{ML}^2 = \frac{1}{N-1} \sum_{n=1}^N (x_n - \hat{\mu})^2$$

Machine Learning Winter '17 | B. Leibe | 50

RWTH AACHEN UNIVERSITY

Maximum Likelihood – Limitations

- Maximum Likelihood has several significant limitations
 - It systematically underestimates the variance of the distribution!
 - E.g. consider the case $N = 1, X = \{x_1\}$

⇒ Maximum-likelihood estimate:

- We say ML *overfits to the observed data*.
- We will still often use ML, but it is important to know about this effect.

Machine Learning Winter '17 | Slide adapted from Bernd Schiele | B. Leibe | 51

RWTH AACHEN UNIVERSITY

Deeper Reason

- Maximum Likelihood is a **Frequentist** concept
 - In the **Frequentist view**, probabilities are the frequencies of random, repeatable events.
 - These frequencies are fixed, but can be estimated more precisely when more data is available.
- This is in contrast to the **Bayesian** interpretation
 - In the **Bayesian view**, probabilities quantify the uncertainty about certain states or events.
 - This uncertainty can be revised in the light of new evidence.
- Bayesians and Frequentists do not like each other too well...

Machine Learning Winter '17 | B. Leibe | 52

RWTH AACHEN UNIVERSITY

Bayesian vs. Frequentist View

- To see the difference...
 - Suppose we want to estimate the uncertainty whether the Arctic ice cap will have disappeared by the end of the century.
 - This question makes no sense in a Frequentist view, since the event cannot be repeated numerous times.
 - In the Bayesian view, we generally have a prior, e.g. from calculations how fast the polar ice is melting.
 - If we now get fresh evidence, e.g. from a new satellite, we may revise our opinion and update the uncertainty from the prior.

$Posterior \propto Likelihood \times Prior$

- This generally allows to get better uncertainty estimates for many situations.
- Main Frequentist criticism**
 - The prior has to come from somewhere and if it is wrong, the result will be worse.

Machine Learning Winter '17 | B. Leibe | 53

RWTH AACHEN UNIVERSITY

Bayesian Approach to Parameter Learning

- Conceptual shift
 - Maximum Likelihood views the true parameter vector θ to be unknown, but fixed.
 - In Bayesian learning, we consider θ to be a random variable.
- This allows us to use knowledge about the parameters θ
 - i.e. to use a prior for θ
 - Training data then converts this prior distribution on θ into a posterior probability density.

- The prior thus encodes knowledge we have about the type of distribution we expect to see for θ .

Machine Learning Winter '17 | Slide adapted from Bernd Schiele | B. Leibe | 54

RWTH AACHEN UNIVERSITY

Bayesian Learning

- Bayesian Learning is an important concept
 - However, it would lead to far here.
 - ⇒ I will introduce it in more detail in the **Advanced ML lecture**.

Machine Learning Winter '17 | B. Leibe | 55

RWTH AACHEN UNIVERSITY

References and Further Reading

- More information in Bishop's book
 - Gaussian distribution and ML: Ch. 1.2.4 and 2.3.1-2.3.4.
 - Bayesian Learning: Ch. 1.2.3 and 2.3.6.
 - Nonparametric methods: Ch. 2.5.

Christopher M. Bishop
Pattern Recognition and Machine Learning
Springer, 2006

Machine Learning Winter '17 | B. Leibe | 84