Machine Learning - Lecture 2

Probability Density Estimation

16.04.2015

Bastian Leibe
RWTH Aachen
http://www.vision.rwth-aachen.de

leibe@vision.rwth-aachen.de

Many slides adapted from B. Schiele
Announcements

• Course webpage
  - http://www.vision.rwth-aachen.de/teaching/
  - Slides will be made available on the webpage

• L2P electronic repository
  - Exercises and supplementary materials will be posted on the L2P

• Please subscribe to the lecture on the Campus system!
  - Important to get email announcements and L2P access!
Announcements

• Exercise sheet 1 is now available on L2P
  - Bayes decision theory
  - Maximum Likelihood
  - Kernel density estimation / k-NN
  ⇒ Submit your results to Ishrat/Michael until evening of 29.04.

• Work in teams (of up to 3 people) is encouraged
  - Who is not part of an exercise team yet?
Course Outline

• Fundamentals (2 weeks)
  - Bayes Decision Theory
  - Probability Density Estimation

• Discriminative Approaches (5 weeks)
  - Linear Discriminant Functions
  - Support Vector Machines
  - Ensemble Methods & Boosting
  - Randomized Trees, Forests & Ferns

• Generative Models (4 weeks)
  - Bayesian Networks
  - Markov Random Fields
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
Recap: Bayes Decision Theory Concepts

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

    - Given data:
      - $C_1 = a$
      - $C_2 = b$
      - $p(C_1) = 0.75$
      - $p(C_2) = 0.25$

- In general: $\sum_k p(C_k) = 1$
Recap: Bayes Decision Theory Concepts

- **Concept 2: Conditional probabilities**
  - 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 $C_k$. 

Slide credit: Bernt Schiele
Bayes Decision Theory Concepts

- **Concept 3: Posterior probabilities**
  - We are typically interested in the \( a \text{ posteriori} \) probability, i.e. the probability of class \( C_k \) given the measurement vector \( x \).

- **Bayes’ Theorem:**
  
  \[
  p(C_k \mid x) = \frac{p(x \mid C_k) \, p(C_k)}{p(x)} = \frac{p(x \mid C_k) \, p(C_k)}{\sum_i p(x \mid C_i) \, p(C_i)}
  \]

- **Interpretation**
  
  \[
  \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{Normalization Factor}}
  \]

Slide credit: Bernt Schiele
Bayes Decision Theory

\[ \text{Posterior} = \frac{\text{Likelihood} \times \text{Prior}}{\text{NormalizationFactor}} \]

\[ p(x|a) \]

\[ p(x|b) \]

\[ p(x|a)p(a) \]

\[ p(x|b)p(b) \]

Decision boundary

Slide credit: Bernt Schiele
Bayesian Decision Theory

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

\[
p(\text{mistake}) = p(x \in \mathcal{R}_1, C_2) + p(x \in \mathcal{R}_2, C_1)
\]

\[
= \int_{\mathcal{R}_1} p(x, C_2) \, dx + \int_{\mathcal{R}_2} p(x, C_1) \, dx.
\]

\[
= \int_{\mathcal{R}_1} p(C_2 | x)p(x) \, dx + \int_{\mathcal{R}_2} p(C_1 | x)p(x) \, dx
\]

The green and blue regions stay constant.

Only the size of the red region varies!
Bayes Decision Theory

- Optimal decision rule
  - Decide for $C_1$ if
    $$p(C_1 \mid x) > p(C_2 \mid x)$$
  - This is equivalent to
    $$p(x \mid C_1)p(C_1) > p(x \mid C_2)p(C_2)$$
  - Which is again equivalent to (Likelihood-Ratio test)
    $$\frac{p(x \mid C_1)}{p(x \mid C_2)} > \frac{p(C_2)}{p(C_1)}$$
    Decision threshold $\theta$

Slide credit: Bernt Schiele
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

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

Slide credit: Bernt Schiele
Bayes Decision Theory

- Decision regions: $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$, ...
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: diagnosis is sick or healthy
      (or: further examination necessary)
    - Classes: patient is sick or healthy
  ➢ The cost may be asymmetric:
    \[
    \text{loss}(\text{decision} = \text{healthy} | \text{patient} = \text{sick}) \gg \text{loss}(\text{decision} = \text{sick} | \text{patient} = \text{healthy})
    \]
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

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

B. Leibe
Classifying with Loss Functions

- Loss functions may be different for different actors.

  Example:

  \[
  L_{stock\text{trader}}(subprime) = \begin{pmatrix}
  -\frac{1}{2}c_{gain} & 0 \\
  0 & 0 
  \end{pmatrix}
  \]

  \[
  L_{bank}(subprime) = \begin{pmatrix}
  -\frac{1}{2}c_{gain} & 0 \\
  \text{skull} & 0 
  \end{pmatrix}
  \]

  \[
  \Rightarrow \text{Different loss functions may lead to different Bayes optimal strategies.}
  \]
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

\[ \mathbb{E}[L] = \sum_{k} \sum_{j} \int_{R_j} L_{kj} p(x, C_k) \, dx \]

- This can be done by choosing the regions \( R_j \) such that

\[ \mathbb{E}[L] = \sum_{k} L_{kj} p(C_k | x) \]

which is easy to do once we know the posterior class probabilities \( p(C_k | x) \).
Minimizing the Expected Loss

- **Example:**
  - 2 Classes: \( C_1, C_2 \)
  - 2 Decision: \( \alpha_1, \alpha_2 \)
  - Loss function: \( L(\alpha_j|C_k) = L_{kj} \)

- Expected loss (= risk \( R \)) for the two decisions:
  
  \[
  \mathbb{E}_{\alpha_1}[L] = R(\alpha_1|x) = L_{11}p(C_1|x) + L_{21}p(C_2|x)
  \]
  
  \[
  \mathbb{E}_{\alpha_2}[L] = R(\alpha_2|x) = L_{12}p(C_1|x) + L_{22}p(C_2|x)
  \]

- **Goal:** Decide such that expected loss is minimized
  - I.e. decide \( \alpha_1 \) if \( R(\alpha_2|x) > R(\alpha_1|x) \)
Minimizing the Expected Loss

\[ R(\alpha_2 | x) > R(\alpha_1 | x) \]
\[ L_{12}p(C_1 | x) + L_{22}p(C_2 | x) > L_{11}p(C_1 | x) + L_{21}p(C_2 | x) \]
\[ (L_{12} - L_{11})p(C_1 | x) > (L_{21} - L_{22})p(C_2 | x) \]
\[ \frac{(L_{12} - L_{11})}{(L_{21} - L_{22})} > \frac{p(C_2 | x)}{p(C_1 | x)} = \frac{p(x | C_2)p(C_2)}{p(x | C_1)p(C_1)} \]
\[ \frac{p(x | C_1)}{p(x | C_2)} > \frac{(L_{21} - L_{22})}{(L_{12} - L_{11})} \frac{p(C_2)}{p(C_1)} \]

⇒ Adapted decision rule taking into account the loss.
The Reject Option

- Classification errors arise from regions where the largest posterior probability \( p(C_k|x) \) 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.
Discriminant Functions

- Formulate classification in terms of comparisons
  - Discriminant functions
    \[ y_1(x), \ldots, y_K(x) \]
  - Classify \( x \) as class \( C_k \) if
    \[ y_k(x) > y_j(x) \quad \forall j \neq k \]

- Examples (Bayes Decision Theory)
  \[
  y_k(x) = p(C_k|x) \\
  y_k(x) = p(x|C_k)p(C_k) \\
  y_k(x) = \log p(x|C_k) + \log p(C_k)
  \]
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.
  \( \Rightarrow \text{Generative methods} \)

- \( y_k(x) = p(C_k|x) \)
  - First solve the inference problem of determining the posterior class probabilities.
  - Then use decision theory to assign each new \( x \) to its class.
  \( \Rightarrow \text{Discriminative methods} \)

- Alternative
  - Directly find a discriminant function \( y_k(x) \) which maps each input \( x \) directly onto a class label.
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
Probability Density Estimation

• Up to now
  ➢ Bayes optimal classification
  ➢ Based on the probabilities \( p(x|C_k)p(C_k) \)

• 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:
    \[ p(x|C_k) \]
  ➢ (For simplicity of notation, we will drop the class label \( C_k \) in the following.)
Probability Density Estimation

• **Data:** \( x_1, x_2, x_3, x_4, \ldots \)

• **Estimate:** \( p(x) \)

• **Methods**
  - Parametric representations
  - Non-parametric representations
  - Mixture models (next lecture)
The Gaussian (or Normal) Distribution

- **One-dimensional case**
  - Mean $\mu$
  - Variance $\sigma^2$

\[
\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left\{ -\frac{(x - \mu)^2}{2\sigma^2} \right\}
\]

- **Multi-dimensional case**
  - Mean $\mu$
  - Covariance $\Sigma$

\[
\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x - \mu)^T\Sigma^{-1}(x - \mu) \right\}
\]
**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.
Gaussian Distribution - Properties

• Quadratic Form
  - $\mathcal{N}$ depends on $\mathbf{x}$ through the exponent
  $$\Delta^2 = (\mathbf{x} - \mu)^T \Sigma^{-1} (\mathbf{x} - \mu)$$
  - Here, $\Delta$ is often called the Mahalanobis distance from $\mu$ to $\mathbf{x}$.

• Shape of the Gaussian
  - $\Sigma$ is a real, symmetric matrix.
  - We can therefore decompose it into its eigenvectors
  $$\Sigma = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^T$$
  $$\Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$$
  and thus obtain
  $$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$
  with
  $$y_i = \mathbf{u}_i^T (\mathbf{x} - \mu).$$
  $$\Rightarrow$$ Constant density on ellipsoids with main directions along the eigenvectors $\mathbf{u}_i$ and scaling factors $\sqrt{\lambda_i}$.

B. Leibe

Image source: C.M. Bishop, 2006
Gaussian Distribution - Properties

• Special cases
  - Full covariance matrix
    \[ \Sigma = \begin{bmatrix} \sigma_{ij} \end{bmatrix} \]
    \Rightarrow General ellipsoid shape
  - Diagonal covariance matrix
    \[ \Sigma = \text{diag}\{\sigma_i\} \]
    \Rightarrow Axis-aligned ellipsoid
  - Uniform variance
    \[ \Sigma = \sigma^2 \mathbf{I} \]
    \Rightarrow Hypersphere

Image source: C.M. Bishop, 2006
Gaussian Distribution - Properties

- The marginals of a Gaussian are again Gaussians:
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
Parametric Methods

- **Given**
  - Data $X = \{x_1, x_2, \ldots, x_N\}$
  - Parametric form of the distribution with parameters $\theta$
  - E.g. for Gaussian distrib.: $\theta = (\mu, \sigma)$

- **Learning**
  - Estimation of the parameters $\theta$

- **Likelihood of $\theta$**
  - Probability that the data $X$ have indeed been generated from a probability density with parameters $\theta$
  $$L(\theta) = p(X|\theta)$$

Slide adapted from Bernt Schiele
Maximum Likelihood Approach

- Computation of the likelihood
  - Single data point: \( p(x_n | \theta) \)
  - 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 \( \theta \) (Learning)
  - Maximize the likelihood
  - Minimize the negative log-likelihood

Slide credit: Bernt Schiele
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.
Maximum Likelihood Approach

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

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

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

\[
E(\theta) = - \sum_{n=1}^{N} \ln p(x_n|\mu, \sigma)
= - \sum_{n=1}^{N} \ln \left( \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ - \frac{\|x_n - \mu\|^2}{2\sigma^2} \right\} \right)
\]
Maximum Likelihood Approach

- Minimizing the log-likelihood

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

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

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

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

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

\[
p(x_n | \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{||x_n - \mu||^2}{2\sigma^2}}
\]
Maximum Likelihood Approach

• We thus obtain

\[ \hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

“sample mean”

• In a similar fashion, we get

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

“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...
Maximum Likelihood Approach

- Or not wrong, but rather *biased*...

- Assume the samples $x_1, x_2, \ldots, x_N$ come from a true Gaussian distribution with mean $\mu$ and variance $\sigma^2$
  
  - We can now compute the expectations of the ML estimates with respect to the data set values. It can be shown that
    
    $$
    \mathbb{E}(\mu_{ML}) = \mu \\
    \mathbb{E}(\sigma^2_{ML}) = \left( \frac{N - 1}{N} \right) \sigma^2
    $$

  - The ML estimate will underestimate the true variance.

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

B. Leibe
Maximum Likelihood - Limitations

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

\[ \implies \text{Maximum-likelihood estimate:} \]

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

\[
Posterior \propto \text{Likelihood} \times \text{Prior}
\]

➢ 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.
Bayesian Approach to Parameter Learning

• Conceptual shift
  - Maximum Likelihood views the true parameter vector $\theta$ to be unknown, but fixed.
  - In Bayesian learning, we consider $\theta$ to be a random variable.

• This allows us to use knowledge about the parameters $\theta$
  - i.e. to use a prior for $\theta$
  - Training data then converts this prior distribution on $\theta$ into a posterior probability density.

- The prior thus encodes knowledge we have about the type of distribution we expect to see for $\theta$.
Bayesian Learning Approach

- Bayesian view:
  - Consider the parameter vector $\theta$ as a random variable.
  - When estimating the parameters, what we compute is

\[
p(x|X) = \int p(x, \theta|X) d\theta
\]

\[
p(x, \theta|X) = p(x|\theta, X)p(\theta|X)
\]

\[
p(x|X) = \int p(x|\theta)p(\theta|X) d\theta
\]

Assumption: given $\theta$, this doesn’t depend on $X$ anymore

This is entirely determined by the parameter $\theta$ (i.e. by the parametric form of the pdf).
Bayesian Learning Approach

\[ p(x | X) = \int p(x | \theta) p(\theta | X) d\theta \]

\[ p(\theta | X) = \frac{p(X | \theta) p(\theta)}{p(X)} = \frac{p(\theta)}{p(X)} L(\theta) \]

\[ p(X) = \int p(X | \theta) p(\theta) d\theta = \int L(\theta) p(\theta) d\theta \]

- Inserting this above, we obtain

\[ p(x | X) = \int \frac{p(x | \theta) L(\theta) p(\theta)}{p(X)} d\theta = \int \frac{p(x | \theta) L(\theta) p(\theta)}{\int L(\theta) p(\theta) d\theta} d\theta \]
Bayesian Learning Approach

- Discussion

If we now plug in a (suitable) prior $p(\theta)$, we can estimate $p(x|X)$ from the data set $X$. 

\[ p(x|X) = \int \frac{p(x|\theta)L(\theta)p(\theta)}{\int L(\theta)p(\theta)d\theta} d\theta \]

Likelihood of the parametric form $\theta$ given the data set $X$.

Estimate for $x$ based on parametric form $\theta$.

Prior for the parameters $\theta$.

Normalization: integrate over all possible values of $\theta$.

B. Leibe
Bayesian Density Estimation

- **Discussion**

\[
p(x|X) = \int p(x|\theta)p(\theta|X)\,d\theta = \int \frac{p(x|\theta)L(\theta)p(\theta)}{\int L(\theta)p(\theta)\,d\theta} \,d\theta
\]

- The probability \( p(\theta|X) \) makes the dependency of the estimate on the data explicit.

- If \( p(\theta|X) \) is very small everywhere, but is large for one \( \hat{\theta} \), then

\[
p(x|X) \approx p(x|\hat{\theta})
\]

\( \Rightarrow \) The more uncertain we are about \( \theta \), the more we average over all parameter values.
Bayesian Density Estimation

- **Problem**
  - In the general case, the integration over $\theta$ is not possible (or only possible stochastically).

- **Example where an analytical solution is possible**
  - Normal distribution for the data, $\sigma^2$ assumed known and fixed.
  - Estimate the distribution of the mean:
    \[
    p(\mu|X) = \frac{p(X|\mu)p(\mu)}{p(X)}
    \]
  - Prior: We assume a Gaussian prior over $\mu$,
    \[
    p(\mu) = \mathcal{N} (\mu|\mu_0, \sigma_0^2) .
    \]
Bayesian Learning Approach

- **Sample mean:**
  \[ \bar{x} = \frac{1}{N} \sum_{n=1}^{N} x_n \]

- **Bayes estimate:**
  \[ \mu_N = \frac{\sigma^2 \mu_0 + N \sigma_0^2 \bar{x}}{\sigma^2 + N \sigma_0^2} \]

  \[ \frac{1}{\sigma^2_N} = \frac{1}{\sigma^2_0} + \frac{N}{\sigma^2} \]

- **Note:**

<table>
<thead>
<tr>
<th></th>
<th>( N = 0 )</th>
<th>( N \to \infty )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \mu_N )</td>
<td>( \mu_0 )</td>
<td>( \mu_{ML} )</td>
</tr>
<tr>
<td>( \sigma^2_N )</td>
<td>( \sigma^2_0 )</td>
<td>0</td>
</tr>
</tbody>
</table>

Slide adapted from Bernt Schiele

Image source: C.M. Bishop, 2006
Summary: ML vs. Bayesian Learning

• Maximum Likelihood
  - Simple approach, often analytically possible.
  - Problem: estimation is biased, tends to overfit to the data.
    ⇒ Often needs some correction or regularization.
  - But:
    - Approximation gets accurate for $N \to \infty$.

• Bayesian Learning
  - General approach, avoids the estimation bias through a prior.
  - Problems:
    - Need to choose a suitable prior (not always obvious).
    - Integral over $\theta$ often not analytically feasible anymore.
  - But:
    - Efficient stochastic sampling techniques available (see Lecture 15).

(In this lecture, we’ll use both concepts wherever appropriate)
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.

• Additional information can be found in Duda & Hart
  - ML estimation: Ch. 3.2
  - Bayesian Learning: Ch. 3.3-3.5
  - Nonparametric methods: Ch. 4.1-4.5

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

R.O. Duda, P.E. Hart, D.G. Stork
Pattern Classification
2nd Ed., Wiley-Interscience, 2000