Announcements: Reminders

- Moodle electronic learning room
  - Slides, exercises, and supplementary material will be made available here
  - Lecture recordings will be uploaded 2-3 days after the lecture
  - Moodle access should now be fixed for all registered participants!

- Course webpage
  - [http://www.vision.rwth-aachen.de/courses/](http://www.vision.rwth-aachen.de/courses/)
  - Slides will also be made available on the webpage

- Please subscribe to the lecture on rwth online!
  - Important to get email announcements and moodle access!

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

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

Recap: The Rules of Probability

- We have shown in the last lecture

  **Sum Rule**
  \[ p(X) = \sum_Y p(X,Y) \]

  **Product Rule**
  \[ p(X,Y) = p(Y|X)p(X) \]

- From those, we can derive

  **Bayes’ Theorem**
  \[ p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)} \]

  where
  \[ p(X) = \sum_Y p(X|Y)p(Y) \]

Probability Densities

- Probabilities over continuous variables are defined over their probability density function (pdf) \( p(x) \)

  \[ p(x \in (a,b)) = \int_a^b p(x) \, dx \]

- The probability that \( x \) lies in the interval \((-\infty, z]\) is given by the cumulative distribution function

  \[ P(z) = \int_{-\infty}^z p(x) \, dx \]
Expectations

- The average value of some function $f(x)$ under a probability distribution $p(x)$ is called its expectation

$$\mathbb{E}[f] = \sum_x p(x) f(x) \quad \text{discrete case}$$

$$\mathbb{E}[f] = \int p(x) f(x) \, dx \quad \text{continuous case}$$

- If we have a finite number $N$ of samples drawn from a pdf, then the expectation can be approximated by

$$\mathbb{E}[f] \simeq \frac{1}{N} \sum_{n=1}^{N} f(x_n)$$

- We can also consider a conditional expectation

$$\mathbb{E}_y[f|y] = \sum_y p(x|y) f(x)$$

Variances and Covariances

- The variance provides a measure how much variability there is in $f(x)$ around its mean value

$$\text{var}[f] = \mathbb{E} \left[ (f(x) - \mathbb{E}[f(x)])^2 \right] = \mathbb{E}[f(x)^2] - \mathbb{E}[f(x)]^2$$

- For two random variables $x$ and $y$, the covariance is defined by

$$\text{cov}[x,y] = \mathbb{E}_{xy} \left[ (x - \mathbb{E}[x])(y - \mathbb{E}[y]) \right]$$

$$= \mathbb{E}_x[y|x] - \mathbb{E}[x] \mathbb{E}[y]$$

- If $x$ and $y$ are vectors, the result is a covariance matrix

$$\text{cov}[x,y] = \mathbb{E}_{xy} \left[ (x - \mathbb{E}[x])(y^T - \mathbb{E}[y]^T) \right]$$

$$= \mathbb{E}_x[y|x] - \mathbb{E}[x] \mathbb{E}[y^T]$$

Bayes Decision Theory

- Example: handwritten character recognition

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

R.A. Fisher, 1925

- Goal:

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

Bayes Decision Theory

- Concept 1: Priors (a priori probabilities)

  - What we can tell about the probability before seeing the data.

  - Example:

    ![Prior probabilities](image)

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

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

- 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 $C_k$.  

  ![Conditional probabilities](image)
Bayes Decision Theory

Example:

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

Question:

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

\[ x = 15 \]

Bayes Decision Theory

Example:

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

Question:

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

\[ x = 25 \]

Bayes Decision Theory

Example:

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

Question:

- Which class?
- Remember that \( p(a) = 0.75 \) and \( p(b) = 0.25 \)...
  - I.e., the decision should be again 'a'.

\[ x = 20 \]

Bayesian Decision Theory

Goal: Minimize the probability of a misclassification

Decision rule:

\[ x < \delta \Rightarrow C_1 \]
\[ x \geq \delta \Rightarrow C_2 \]

How does \( p(\text{mistake}) \) change when we move \( \delta \)?

\[ p(\text{mistake}) = p(x \in R_1, C_2) + p(x \in R_2, C_1) \]
\[ = \int_{R_2} p(x, C_2) \, dx + \int_{R_2} p(x, C_1) \, dx \]
\[ = \int_{R_1} p(C_2|x)p(x) \, dx + \int_{R_2} p(C_1|x)p(x) \, dx \]

The green and blue regions stay constant.

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

- Optimal decision rule
  - Decide for C1, 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)
    \[ \frac{p(x|C_1)}{p(x|C_2)} > \frac{p(C_1)}{p(C_2)} \]
  - Decision threshold \( \theta \)

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

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:
    \[ \text{loss}(\text{decision} = \text{sick} | \text{patient} = \text{sick}) > \text{loss}(\text{decision} = \text{healthy} | \text{patient} = \text{sick}) \]

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
  \[ L_{\text{cancer diagnosis}} = \begin{bmatrix} 0 & 1000 \\ 1 & 0 \end{bmatrix} \]

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: \( L(\alpha|C) = L_k \)
  - Expected loss (= risk \( R \)) for the two decisions:
    \[
    E_{\alpha_1}[L] = R(\alpha_1|x) = L_{11}p(C_1|x) + L_{21}p(C_2|x)
    \]
    \[
    E_{\alpha_2}[L] = R(\alpha_2|x) = L_{12}p(C_1|x) + L_{22}p(C_2|x)
    \]
  - Goal: Decide such that expected loss is minimized
    - i.e. decide \( \alpha \), if \( R(\alpha_2|x) > R(\alpha_1|x) \)

## 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)
    \[
    y_k(x) = p(C_k|x)
    \]
    \[
    y_k(x) = log p(x|C_k)p(C_k)
    \]
    \[
    y_k(x) = log p(x|C_k) + log p(C_k)
    \]

## 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.
      - **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.
      - **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:
    \[
    p(x|C_k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
    \]
  - (For simplicity of notation, we will drop the class label \( C_k \) in the following.)

The Gaussian (or Normal) Distribution

- One-dimensional case
  - Mean \( \mu \)
  - Variance \( \sigma^2 \)
    \[
    N(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}
    \]

- Multi-dimensional case
  - Mean \( \mu \)
  - Covariance \( \Sigma \)
    \[
    N(x|\mu,\Sigma) = \frac{1}{|\Sigma|^{\frac{1}{2}}(2\pi)^{\frac{d}{2}}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}
    \]

Gaussian Distribution – Properties

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

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

Gaussian Distribution – Properties

- Special cases
  - Full covariance matrix
    \[
    \Sigma = [\sigma_{ij}]
    \]
    \( \Rightarrow \) General ellipsoid shape
  - Diagonal covariance matrix
    \[
    \Sigma = \text{diag}\{\sigma_i\}
    \]
    \( \Rightarrow \) Axis-aligned ellipsoid
  - Uniform variance
    \[
    \Sigma = \sigma^2 I
    \]
    \( \Rightarrow \) Hypersphere
Gaussian Distribution – Properties

• The marginals of a Gaussian are again Gaussians:

![Gaussian Distribution Graph]

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

Probability Densities

• Probabilities over continuous variables are defined over their probability density function (pdf) \( p(x) \)

\[ p(x \in (a, b)) = \int_a^b p(x) \, dx \]

• The probability that \( x \) lies in the interval \((-\infty, z)\) is given by the cumulative distribution function

\[ P(z) = \int_{-\infty}^z p(x) \, dx \]

Expectations

• The average value of some function \( f(x) \) under a probability distribution \( p(x) \) is called its expectation

\[ E[f] = \sum p(x) f(x) \]

continuous case

• If we have a finite number \( N \) of samples drawn from a pdf, then the expectation can be approximated by

\[ E[f] \approx \frac{1}{N} \sum_{n=1}^N f(x_n) \]

• We can also consider a conditional expectation

\[ E_x[f|y] = \sum p(x|y) f(x) \]

Variances and Covariances

• The variance provides a measure how much variability there is in \( f(x) \) around its mean value \( E[f(x)] \)

\[ \text{var}[f] = E\left[ (f(x) - E[f(x)])^2 \right] = E[f(x)^2] - E[f(x)]^2 \]

• For two random variables \( x \) and \( y \), the covariance is defined by

\[ \text{cov}[x,y] = E_{x,y} \{(x - E[x]) (y - E[y])\} = E_{x,y} [xy] - E[x] E[y] \]

• If \( x \) and \( y \) are vectors, the result is a covariance matrix

\[ \text{cov}[x,y] = E_{x,y} \{(x - E[x]) (y - E[y])\} = E_{x,y} [xy^T] - E[x] E[y^T] \]

Parametric Methods

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

• Learning
  - Estimation of the parameters \( \theta \)

• Likelihood of \( \theta \)
  - Probability that the data \( X \) have indeed been generated from a probability density with parameters \( \theta \)

\[ L(\theta) = p(X|\theta) \]
This is a very important result.

Corrected estimate:

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

"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…

We want to obtain \( \theta \) such that \( L(\theta) \) is maximized.

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

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

Estimation of the parameters \( \theta \) (Learning)

- Maximize the likelihood
- Minimize the negative log-likelihood

Minimizing the log-likelihood

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

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

\[ \frac{\partial}{\partial \mu} E(\mu, \sigma) \]

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

\[ \frac{\partial}{\partial \sigma^2} E(\mu, \sigma) = 0 \quad \Leftrightarrow \quad \hat{\sigma}^2 = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \hat{\mu})^2 \]

We can now compute the expectations of the ML estimates with respect to the data set values. It can be shown that

\[ E(\hat{\mu}_{ML}) = \mu \]

\[ E(\hat{\sigma}^2_{ML}) = \left( \frac{N-1}{N} \right) \sigma^2 \]

\( \hat{\theta} \) will underestimate the true variance.

Corrected estimate:

\[ \hat{\sigma}^2 = \frac{N}{N-1} \hat{\sigma}^2_{ML} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \hat{\mu})^2 \]
Maximum Likelihood – Limitations

• Maximum Likelihood has several significant limitations
  - It systematically underestimates the variance of the distribution!
  - E.g. consider the case
    \[ N = 1, X = \{ x_1 \} \]
    \[ \Rightarrow \text{Maximum-likelihood estimate:} \]
    \[ \hat{\sigma} = 0 ! \]
  - We say ML overfits 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 \( \propto \) 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.

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

• Bayesian Learning is an important concept
  - However, it would lead to far here.
  - \( \Rightarrow \) I will introduce it in more detail in the Advanced ML lecture.

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