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

- Recap: 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)

  \[ p(C_1) \]
  \[ p(C_2) \]

  \[ C_1 = a \]
  \[ p(C_1) = 0.75 \]
  \[ C_2 = b \]
  \[ p(C_2) = 0.25 \]

  In general: \( \sum p(C_k) = 1 \)

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!
Bayes Decision Theory Concepts

- Concept 3: Posterior probabilities
  - We are typically interested in the posterior probability, i.e. the probability of class $C_i$ given the measurement vector $x$.
- Bayes' Theorem:
  $$p(C_i|x) = \frac{p(x|C_i)p(C_i)}{\sum_j p(x|C_j)p(C_j)}$$
- Interpretation
  $$Posterior = \frac{Likelihood \times Prior}{Normalization Factor}$$

Recap: Bayes Decision Theory

- Optimal decision rule
  - Decide for $C_i$ 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)
    $$\frac{p(x|C_1)}{p(x|C_2)} > \frac{p(C_2)}{p(C_1)}$$
    Decision threshold $\theta$

Recap: 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 ($= risk \, \bar{R}$) 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 $\alpha_j$ if $R(\alpha_2|x) > R(\alpha_1|x)$

Recap: Minimizing the Expected Loss

- Decision regions: $R_1, R_2, R_3, \ldots$
  $$R(\alpha_2|x) > R(\alpha_1|x)$$
  $$L_{12}p(C_1|x) + L_{22}p(C_2|x) > L_{11}p(C_1|x) + L_{21}p(C_2|x)$$
  $$\frac{(L_{12} - L_{11})}{(L_{21} - L_{22})} > \frac{p(C_2)}{p(C_1)}$$

$\Rightarrow$ Adapted decision rule taking into account the loss.
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 $y_k(x) > y_j(x) \ \forall j \neq k$
- 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.

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.)

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

Methods
- Parametric representations (today)
- Non-parametric representations (lecture 3)
- Mixture models (lecture 4)
The Gaussian (or Normal) Distribution

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

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

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.

- 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
  $$\Delta^2 = \sum_{i=1}^{D} \frac{y_i}{\lambda_i} (x - \mu, u_i)$$
  - Constant density on ellipsoids with main directions along the eigenvectors $u_i$ and scaling factors $\sqrt{\lambda_i}$.

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

- 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 dist.: $\theta = (\mu, \sigma)$

Learning
- Estimation of the parameters $\theta$

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

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.

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

Maximum Likelihood

- Computation of the likelihood
  - Single data point: $p(x_n|\theta) = \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{(x_n - \mu)^2}{2\sigma^2} \right\}$
  - 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 $\theta$ (Learning)
    - Maximize the likelihood
    - Minimize the negative log-likelihood

Maximum Likelihood Approach

- Minimizing the log-likelihood
  - How do we minimize a function?
    $\Rightarrow$ Take the derivative and set it to zero.
    $\frac{\partial}{\partial \mu} E(\mu, \sigma) = - \sum_{n=1}^{N} \frac{2(x_n - \mu)}{2\sigma^2} = 0$
  - Log-likelihood for Normal distribution (1D case)
    $E(\theta) = -\sum_{n=1}^{N} \ln p(x_n|\mu, \sigma)$
    $= -\sum_{n=1}^{N} \ln \left( \frac{1}{\sqrt{2\pi\sigma}} \exp \left\{ -\frac{||x_n - \mu||^2}{2\sigma^2} \right\} \right)$

- 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
    \[
    E(\hat{\mu}_{ML}) = \mu \\
    E(\hat{\sigma}^2_{ML}) = \left( \frac{N-1}{N} \right) \sigma^2
    \]
  - The ML estimate will underestimate the true variance.
- Corrected estimate:
  \[
  \hat{\sigma}^2 = \frac{N}{N-1} \hat{\sigma}^2_{ML} = \frac{1}{N-1} \sum_{n=1}^{N} (x_n - \hat{\mu})^2
  \]

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.
    \[
    \text{Posterior} \propto \text{Likelihood} \times \text{Prior}
    \]
  - 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
    \[
    p(x|X) = \int p(x, \theta|X)d\theta
    \]
    \[
    p(x, \theta|X) = p(x|\theta, X)p(\theta|X)
    \]
    \[
    p(x|X) = \int p(x|\theta)p(\theta|X)d\theta
    \]
  - This is entirely determined by the parameter \( \theta \) (i.e. by the parametric form of the pdf).
Bayesian Learning Approach

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

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

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

- Inserting this above, we obtain

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

Discussion

- 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

Discussion

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

- 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.

Summary: ML vs. Bayesian Learning

- Maximum Likelihood
  - Simple approach, often analytically possible.
  - Problem: estimation bias, tends to overfit to the data.
  - \( \Rightarrow \) 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 Adv. ML).

(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