Learning and Adaptivity

Basic Concepts - Lecture II

Course Outline

Basic Concepts

- Parametric Method,
- Bayesian Learning and Nonparametrics Methods
- Clustering and Mixture of Gaussians

Supervised Learning, Classification Approaches

- Ensemble Methods and Boosting
- Randomized Trees, Forest

Unsupervised Learning

- Dimensionality Reduction and Manifold Learning (PCA, SNE/t-SNE, MDS, umap)
- Uncertainty Estimation

Reinforcement Learning

- Classical Reinforcement Learning
- Deep Reinforcement Learning

Today's topics

Bayes Decision Theory

- Basic concepts
- Minimizing the misclassification rate
- Minimizing the expected loss

Probability Density Estimation

- General concepts
- Gaussian distribution

Parametric Methods

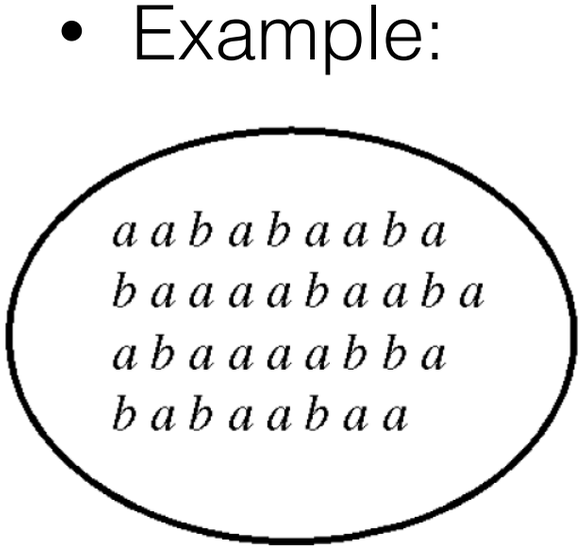
- Maximum Likelihood approach
- Bayesian vs. Frequentist view on probability
- Bayesian Learning

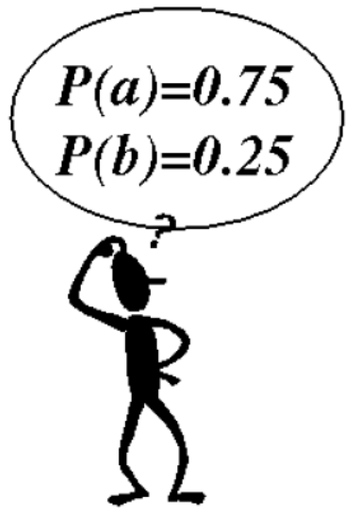
Recap: Bayes Decision Theory

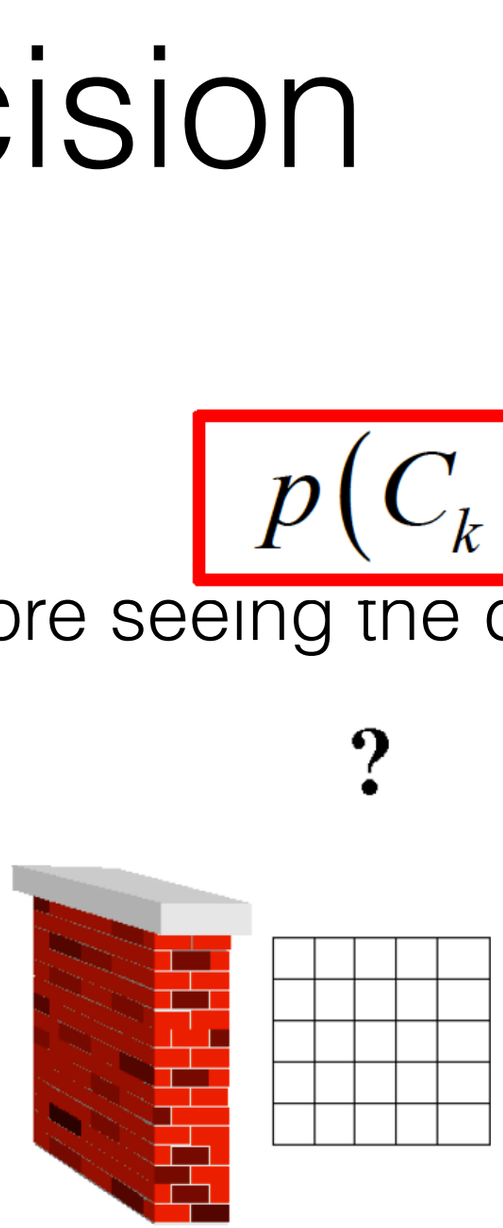
$$p(C_k)$$

Concept 1: **Priors** (a priori probabilities)

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







$$C_1 = a \qquad p(C_1) = 0.75$$

$$C_2 = b \qquad p(C_2) = 0.25$$

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

Recap: Bayes Decision Theory

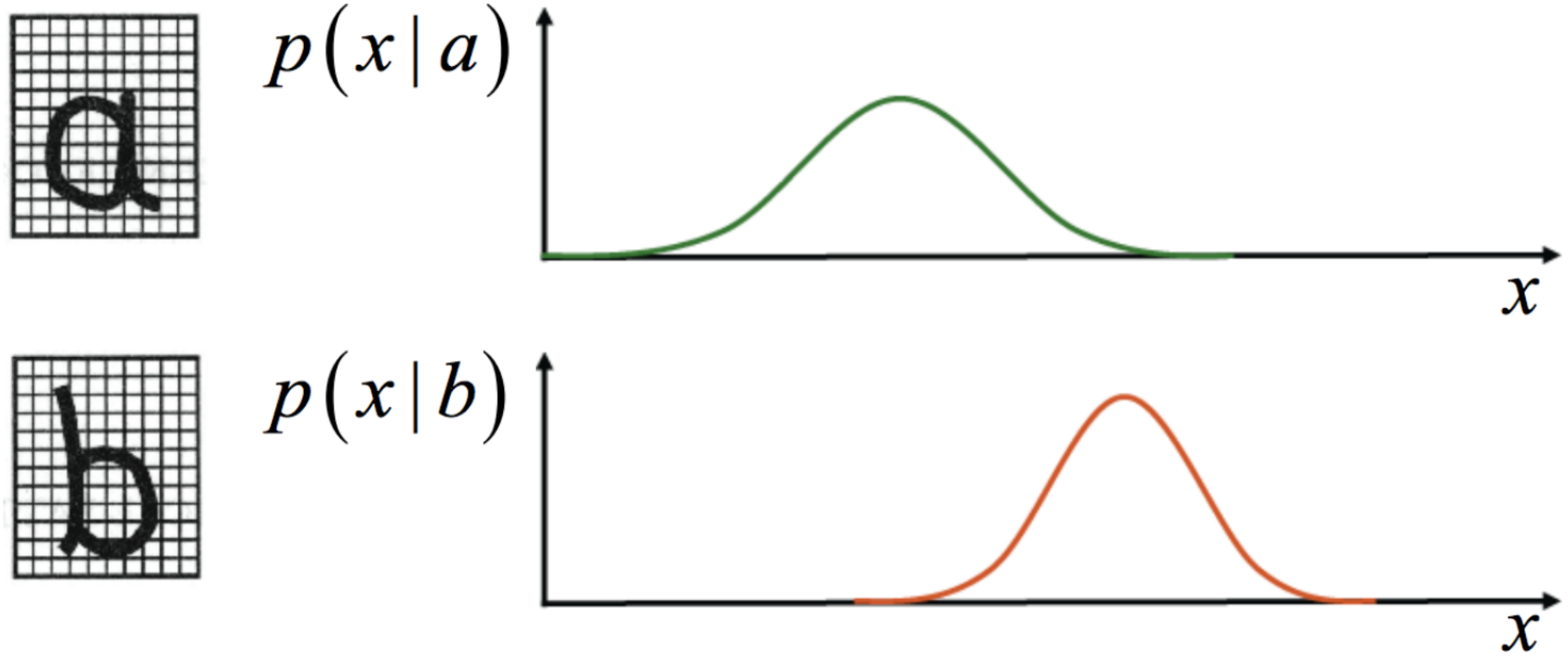
Concept 2: **Conditional probabilities**

$p(x|C_k)$

- 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.



Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

Concept 3: **Posterior probabilities**

$$p(C_k \mid x)$$

- We are typically interested in the a *posteriori* probability, i.e. the probability of class C_k given the measurement vector x

Bayes' Theorem:

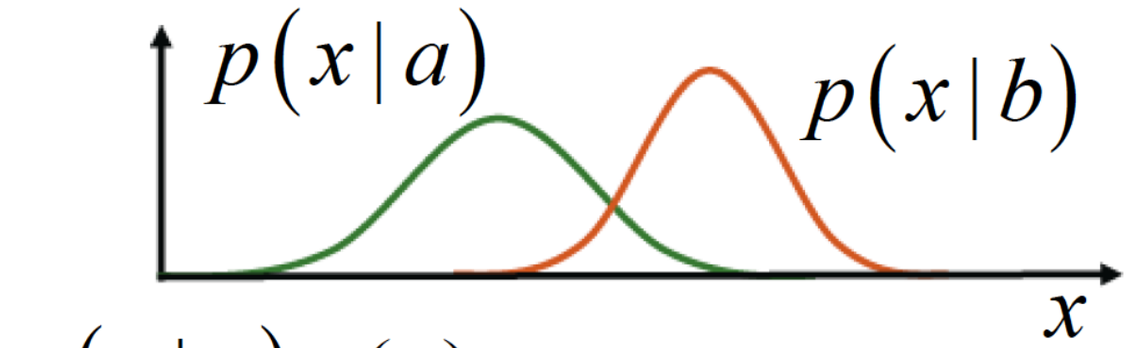
$$p(C_k \mid x) = \frac{p(x \mid C_k)\, p(C_k)}{p(x)} = \frac{p(x \mid C_k)\, p(C_k)}{\sum_i p(x \mid C_i)\, p(C_i)}$$

Interpretation

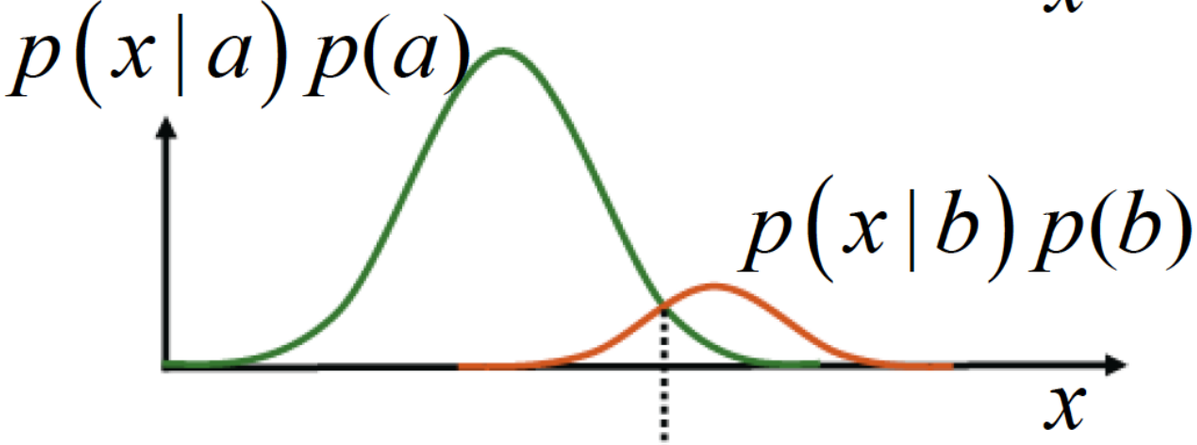
$$Posterior = \frac{Likelihood \times Prior}{Normalization\ Factor}$$

Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

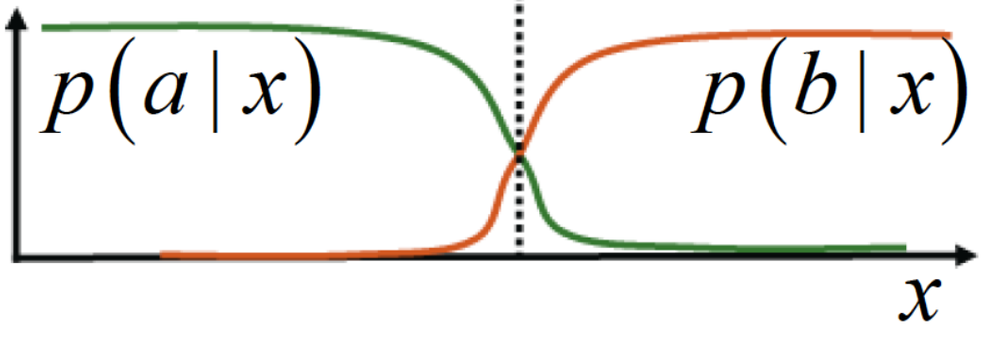


Likelihood



Likelihood $\times$ *Prior*

Decision boundary

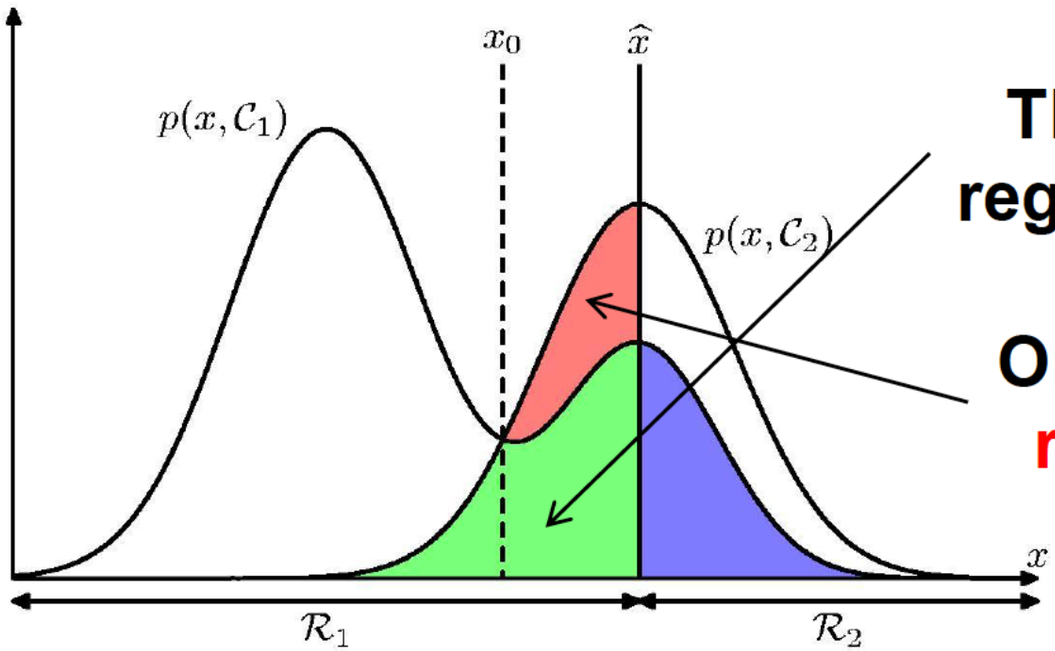


$$Posterior = \frac{Likelihood \times Prior}{NormalizationFactor}$$

Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

Goal: Minimize the probability of a misclassification



The green and blue regions stay constant.

Only the size of the red region varies!

$$
\begin{aligned}
p(\text{mistake}) &= p(\mathbf{x} \in \mathcal{R}_1, \mathcal{C}_2) + p(\mathbf{x} \in \mathcal{R}_2, \mathcal{C}_1) \\
&= \int_{\mathcal{R}_1} p(\mathbf{x}, \mathcal{C}_2)\, \mathrm{d}\mathbf{x} + \int_{\mathcal{R}_2} p(\mathbf{x}, \mathcal{C}_1)\, \mathrm{d}\mathbf{x} \\
&= \int_{\mathcal{R}_1} p(\mathcal{C}_2|\mathbf{x}) p(\mathbf{x}) d\mathbf{x} + \int_{\mathcal{R}_2} p(\mathcal{C}_1|\mathbf{x}) p(\mathbf{x}) d\mathbf{x}
\end{aligned}
$$

Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

Optimal decision rule

Decide for C_1 if

$$p(\mathcal{C}_1|x) > p(\mathcal{C}_2|x)$$

This is equivalent to

$$p(x|\mathcal{C}_1)p(\mathcal{C}_1) > p(x|\mathcal{C}_2)p(\mathcal{C}_2)$$

Which is again equivalent to (Likelihood-Ratio test)

$$\frac{p(x|\mathcal{C}_1)}{p(x|\mathcal{C}_2)} > \underbrace{\frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}}_{\text{Decision threshold } \theta}$$

Slide source: B. Leibe Slides ML, 2015

Generalization to More Than Two Classes

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

$$p(\mathcal{C}_k|x) > p(\mathcal{C}_j|x) \quad \forall j \neq k$$

$$p(x|\mathcal{C}_k)p(\mathcal{C}_k) > p(x|\mathcal{C}_j)p(\mathcal{C}_j) \quad \forall j \neq k$$

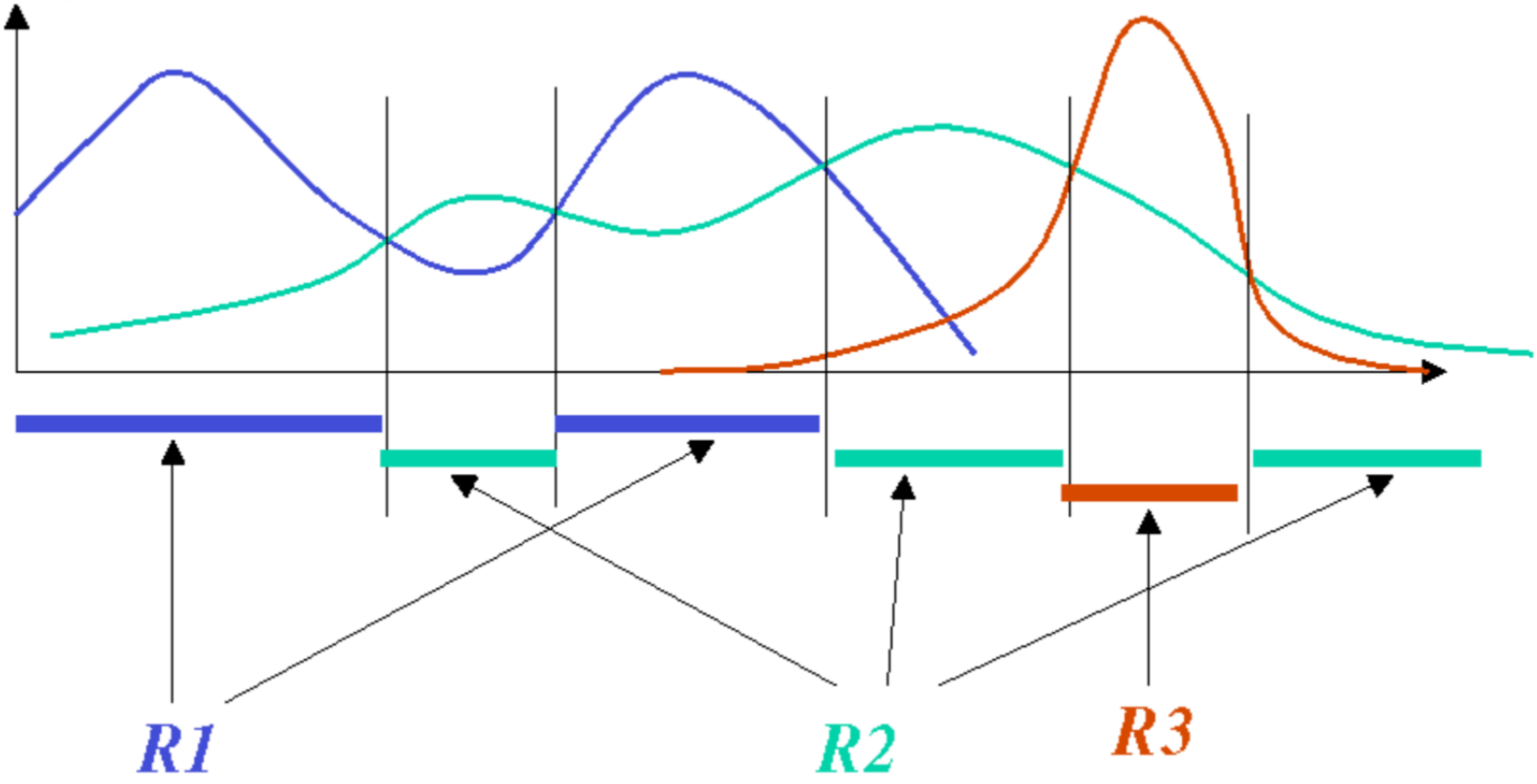
Likelihood-Ratio test

$$\frac{p(x|\mathcal{C}_k)}{p(x|\mathcal{C}_j)} > \frac{p(\mathcal{C}_j)}{p(\mathcal{C}_k)} \quad \forall j \neq k$$

Slide source: B. Leibe Slides ML, 2015

Bayes Decision Theory

Decision regions: $R1, R2, R3, \ldots$



Slide source: B. Leibe Slides ML, 2015

Generalization with Loss Function

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:

$$loss(decision = healthy|patient = sick) >>$$

$$loss(decision = sick|patient = healthy)$$

Slide source: B. Leibe Slides ML, 2015

Classifying with Loss Function

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

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

Example cancer diagnosis:

$$L_{cancer\ \ diagnosis} = \begin{array}{c} \\ \textbf{Truth} \\ \begin{array}{l} \text{cancer} \\ \text{normal} \end{array} \end{array} \begin{array}{c} \textbf{Decision} \\ \begin{array}{cc} \text{cancer} & \text{normal} \end{array} \\ \begin{pmatrix} 0 & 1000 \\ 1 & 0 \end{pmatrix} \end{array}$$

Slide source: B. Leibe Slides ML, 2015

Classifying with Loss Function

Loss function can be different for different actors

Example:

"invest" "don't invest"

$$L_{stocktrader}(subprime) = \begin{pmatrix} -\frac{1}{2}c_{gain} & 0 \\ 0 & 0 \end{pmatrix}$$

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

Different loss functions may lead to different Bayes optimal strategies

Slide source: B. Leibe Slides ML, 2015

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**

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

This can be done by choosing the regions R_j such that

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

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

Slide source: B. Leibe Slides ML, 2015

Minimizing the Expected Loss

Example

2 Classes: C_1, C_2

2 Decision: α_1, α_2

Loss function: $L(\alpha_j|\mathcal{C}_k) = L_{kj}$

Expected loss (= risk R) for the two decisions:

$$\mathbb{E}_{\alpha_1}[L] = R(\alpha_1|\mathbf{x}) = L_{11}p(\mathcal{C}_1|\mathbf{x}) + L_{21}p(\mathcal{C}_2|\mathbf{x})$$

$$\mathbb{E}_{\alpha_2}[L] = R(\alpha_2|\mathbf{x}) = L_{12}p(\mathcal{C}_1|\mathbf{x}) + L_{22}p(\mathcal{C}_2|\mathbf{x})$$

Goal: Decide such that expected loss is minimized

I.e. decide α_1 if $R(\alpha_2|\mathbf{x}) > R(\alpha_1|\mathbf{x})$

Slide source: B. Leibe Slides ML, 2015

Minimizing the Expected Loss

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

$$L_{12}p(\mathcal{C}_1|\mathbf{x}) + L_{22}p(\mathcal{C}_2|\mathbf{x}) > L_{11}p(\mathcal{C}_1|\mathbf{x}) + L_{21}p(\mathcal{C}_2|\mathbf{x})$$

$$(L_{12} - L_{11})p(\mathcal{C}_1|\mathbf{x}) > (L_{21} - L_{22})p(\mathcal{C}_2|\mathbf{x})$$

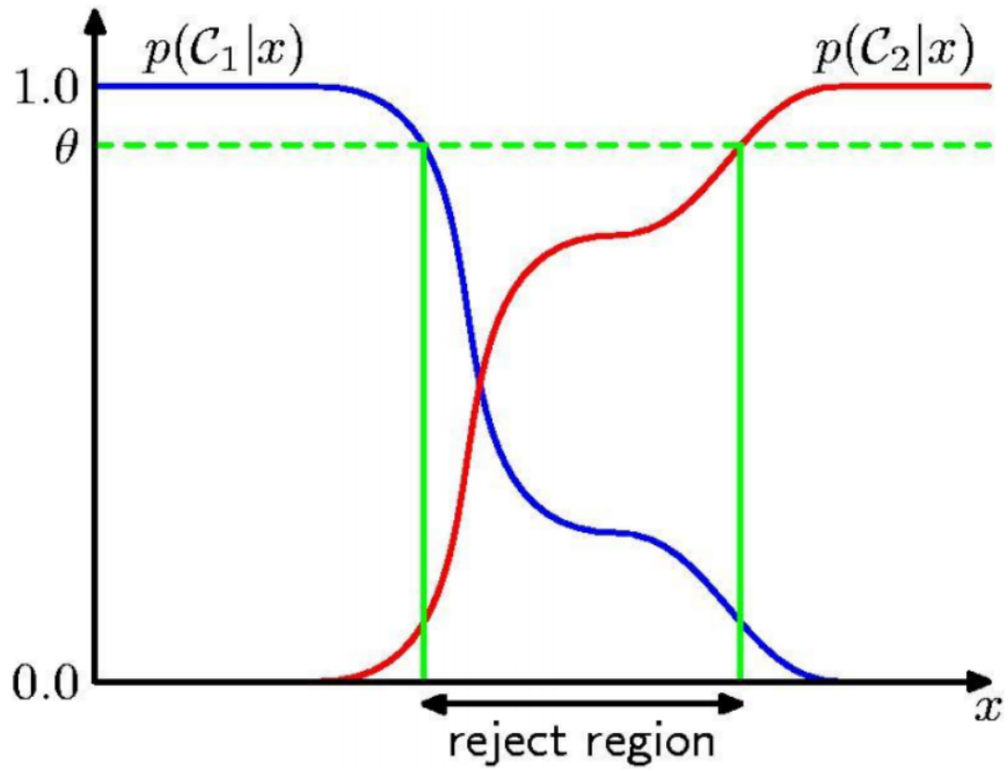
$$\frac{(L_{12} - L_{11})}{(L_{21} - L_{22})} > \frac{p(\mathcal{C}_2|\mathbf{x})}{p(\mathcal{C}_1|\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)}$$

$$\frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} > \frac{(L_{21} - L_{22})}{(L_{12} - L_{11})}\frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}$$

=> Adapted decision rule taking into account the loss

Slide source: B. Leibe Slides ML, 2015

The Reject Option



Classification errors arise from regions where the largest posterior probability $p(C_k|\mathrm{x})$ 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

Slide source: B. Leibe Slides ML, 2015

Discriminant Functions

Formulate classification in terms of comparisons

- Discriminant functions

$$y_1(x), ..., y_K(x)$$

- Classify x as class Ck if

$$y_k(x) > y_j(x) \; \forall j \neq j$$

Examples (Bayes Decision Theory)

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

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

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

Slide source: B. Leibe Slides ML, 2015

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

Slide source: B. Leibe Slides ML, 2015

Today's topics

Bayes Decision Theory

- Basic concepts
- Minimizing the misclassification rate
- Minimizing the expected loss

Probability Density Estimation

- General concepts
- Gaussian distribution

Parametric Methods

- Maximum Likelihood approach
- Bayesian vs. Frequentist view on probability
- Bayesian Learning

Probability Density Estimation

Up to now

- Bayes optimal classification
- Based on the probabilities

$$p(\mathrm{x}|C_k)p(C_k)$$

How can we estimate (=learn) those probability densities?

- Supervised training case: data and class labels are known.
- Estimate the probability density for each class separately:

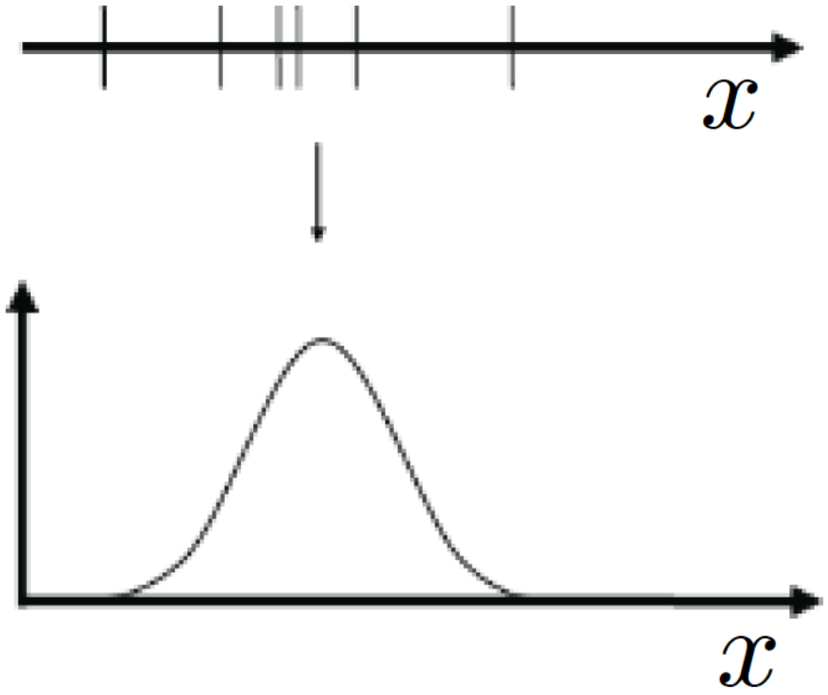
$$p(\mathrm{x}|C_k)$$

- (For simplicity of notation, we will drop the class label in the following.)

Slide source: B. Leibe Slides ML, 2015

Probability Density Estimation

Data: $x_1, x_2, x_3, x_4, \ldots$



Estimate: $p(x)$

Methods

- Parametric representations
- Non-parametric representations
- Mixture models

Slide source: B. Leibe Slides ML, 2015

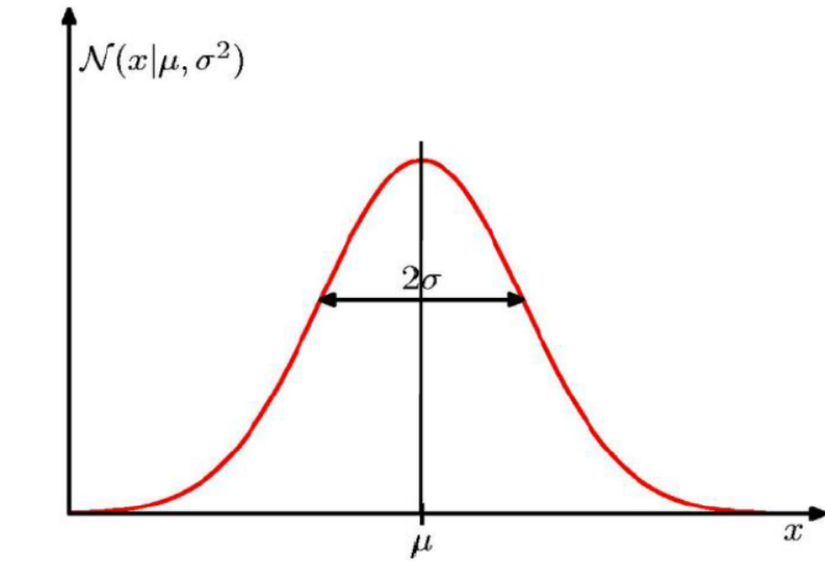
The Gaussian (or Normal) Distribution

One-dimensional case

Mean μ

Variance σ^2

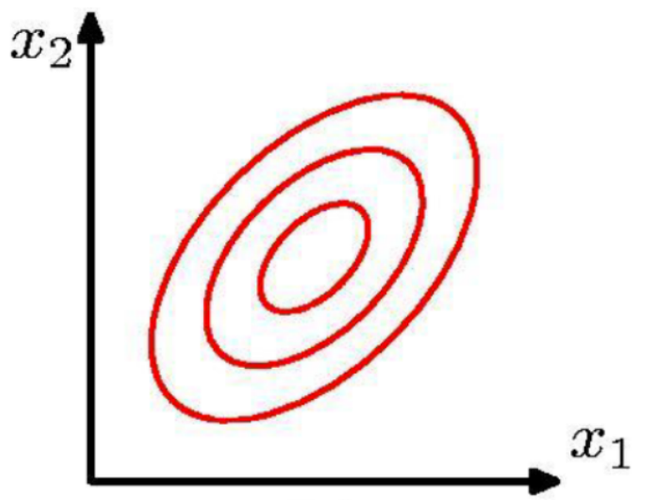
$$\mathcal{N}(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$



Multi-dimensional case

Mean μ

Covariance Σ



$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu},\boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}}\exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

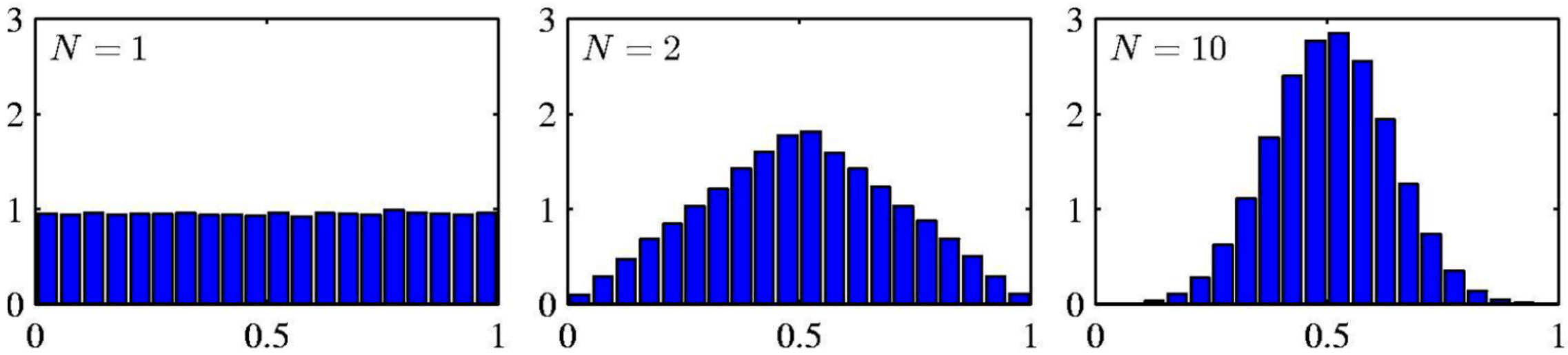
Slide source: B. Leibe Slides ML, 2015

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



Slide source: B. Leibe Slides ML, 2015

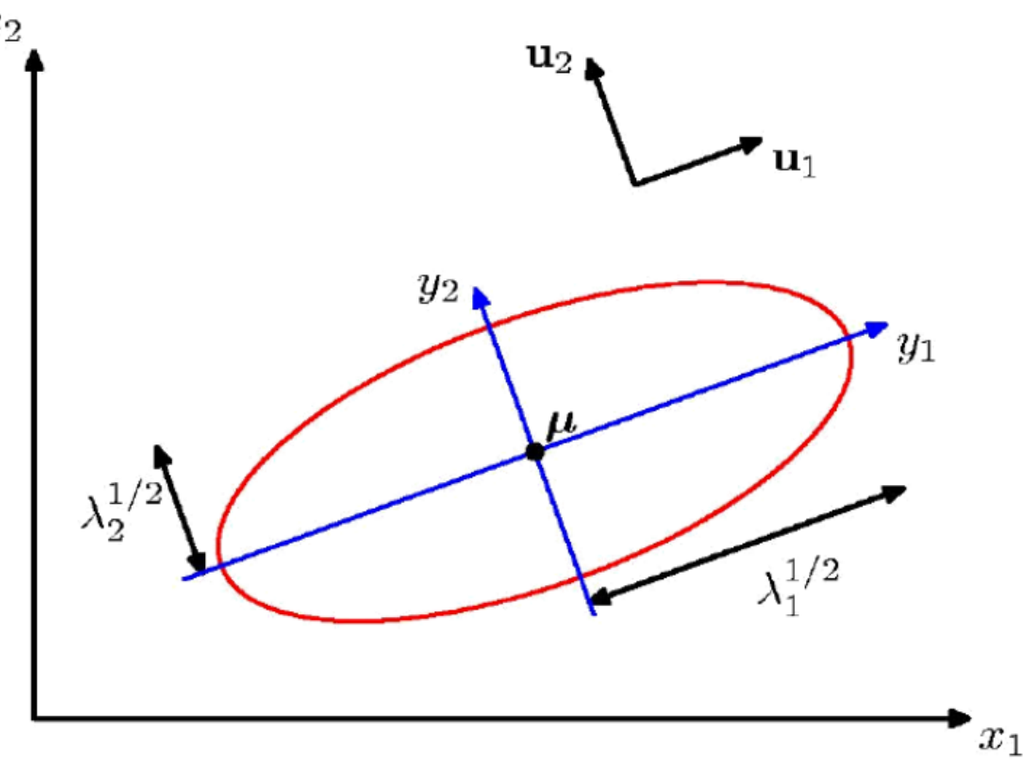
Gaussian Distribution – Properties

Quadratic Form

$\mathcal{N}$ depends on $\mathbf{x}$ through the exponent

$$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$$

Here, $\triangle$ is often called the Mahalanobis distance from $\mathbf{x}$ to μ.



Shape of the Gaussian

$\boldsymbol{\Sigma}$ is a real, symmetric matrix.

We can therefore decompose it into its eigenvectors

$$\boldsymbol{\Sigma} = \sum_{i=1}^{D} \lambda_i \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}} \qquad \boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

and thus obtain $\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$ with $y_i = \mathbf{u}_i^{\mathrm{T}} (\mathbf{x} - \boldsymbol{\mu})$

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

Slide source: B. Leibe Slides ML, 2015

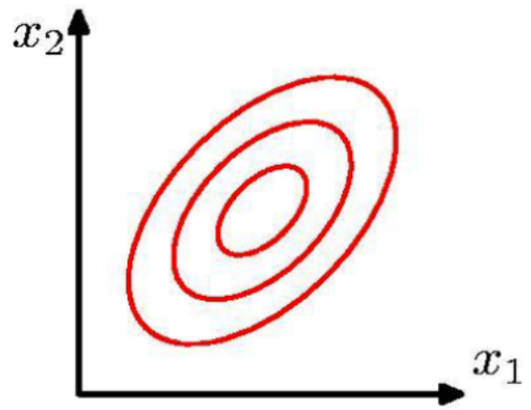
Gaussian Distribution – Properties

Special cases

Full covariance matrix

$$\boldsymbol{\Sigma} = [\sigma_{ij}]$$

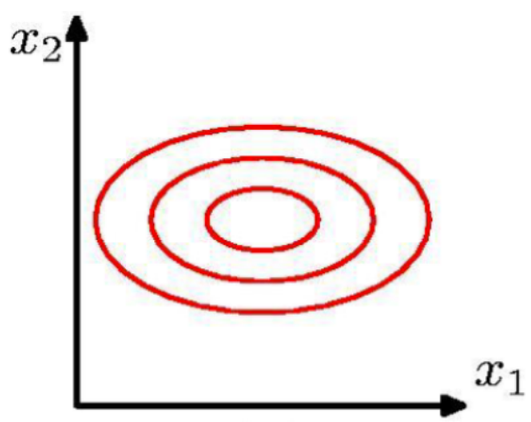
=> General ellipsoid shape



Diagonal covariance matrix

$$\boldsymbol{\Sigma} = diag\{\sigma_i\}$$

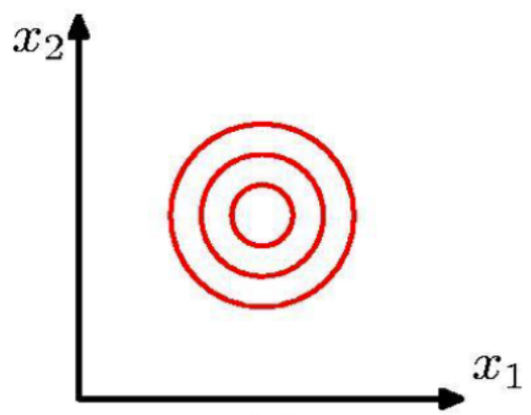
=> Axis-aligned ellipsoid



Uniform variance

$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$$

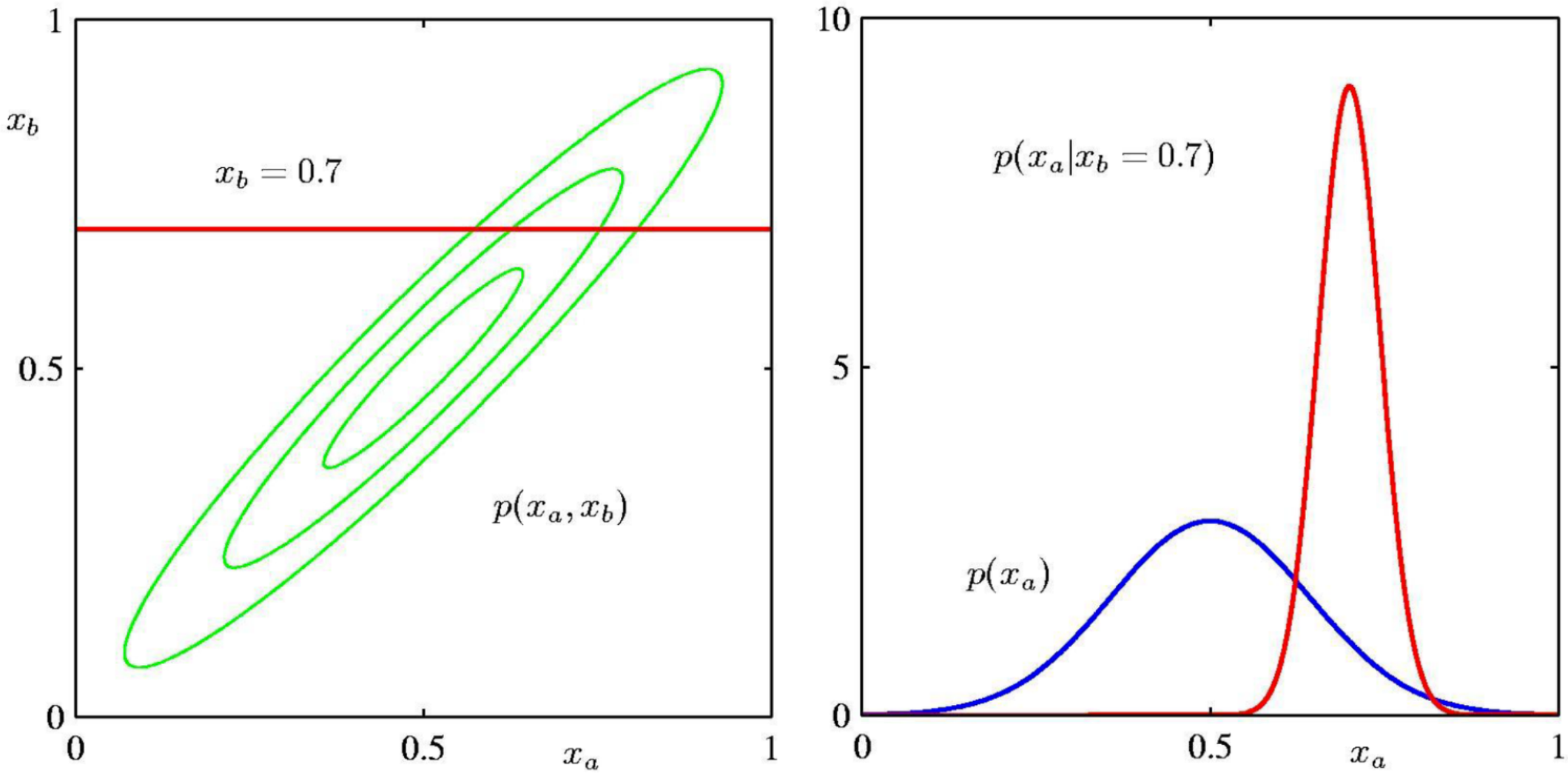
=> Hypersphere



Slide source: B. Leibe Slides ML, 2015

Gaussian Distribution – Properties

The marginals of a Gaussian are again Gaussians:



Slide source: B. Leibe Slides ML, 2015

Today's topics

Bayes Decision Theory

- Basic concepts
- Minimizing the misclassification rate
- Minimizing the expected loss

Probability Density Estimation

- General concepts
- Gaussian distribution

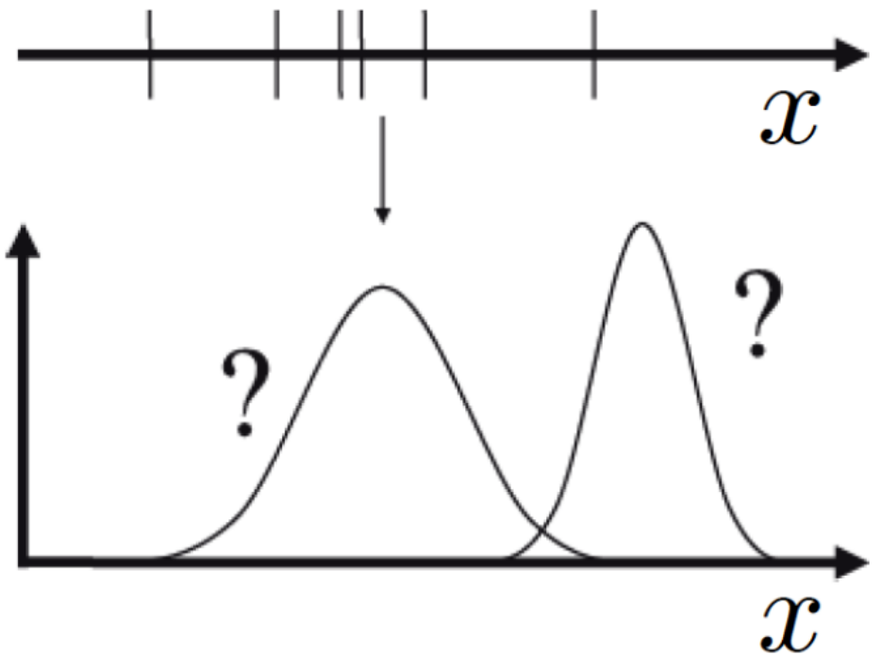
Parametric Methods

- Maximum Likelihood approach
- Bayesian vs. Frequentist view on probability
- Bayesian Learning

Parametric Methods

Given

- Data $X = \{x_1, x_2, \ldots, x_N\}$
- 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

$$L(\theta) = p(X|\theta)$$

Slide source: B. Leibe Slides ML, 2015

Maximum Likelihood Approach

Computation of the likelihood

- Single data point: $p(x_n|\theta)$

- Assumption: all data points are independent

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

- Log-likelihood

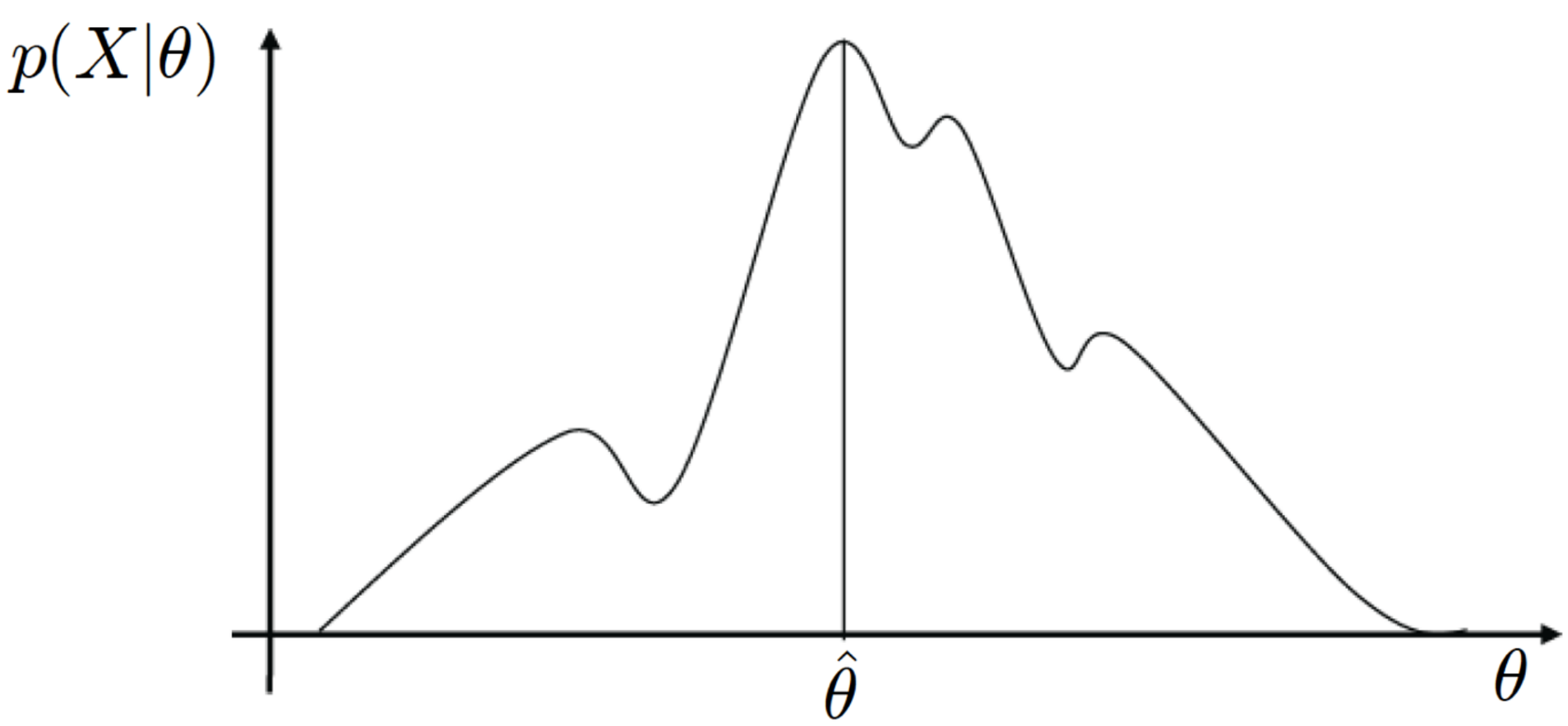
$$E(\theta) = -lnL(\theta) = -\sum_{n=1}^{N} \ln p(x_n|\theta)$$

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

Slide source: B. Leibe Slides ML, 2015

Maximum Likelihood Approach

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



Slide source: B. Leibe Slides ML, 2015

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)

$$\begin{aligned} E(\theta) &= -\sum_{n=1}^{N}\ln p(x_n|\mu,\sigma) \\ &= -\sum_{n=1}^{N}\ln\left(\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{||x_n-\mu||^2}{2\sigma^2}\right\}\right) \end{aligned}$$

Slide source: B. Leibe Slides ML, 2015

Maximum Likelihood Approach

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

Minimizing the log-likelihood

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

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

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

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

Slide source: B. Leibe Slides ML, 2015

Maximum Likelihood Approach

- We thus obtain

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

"sample mean"

- In 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…

Slide source: B. Leibe Slides ML, 2015

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

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

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

=> The ML estimate will underestimate the true variance

- Corrected estimate

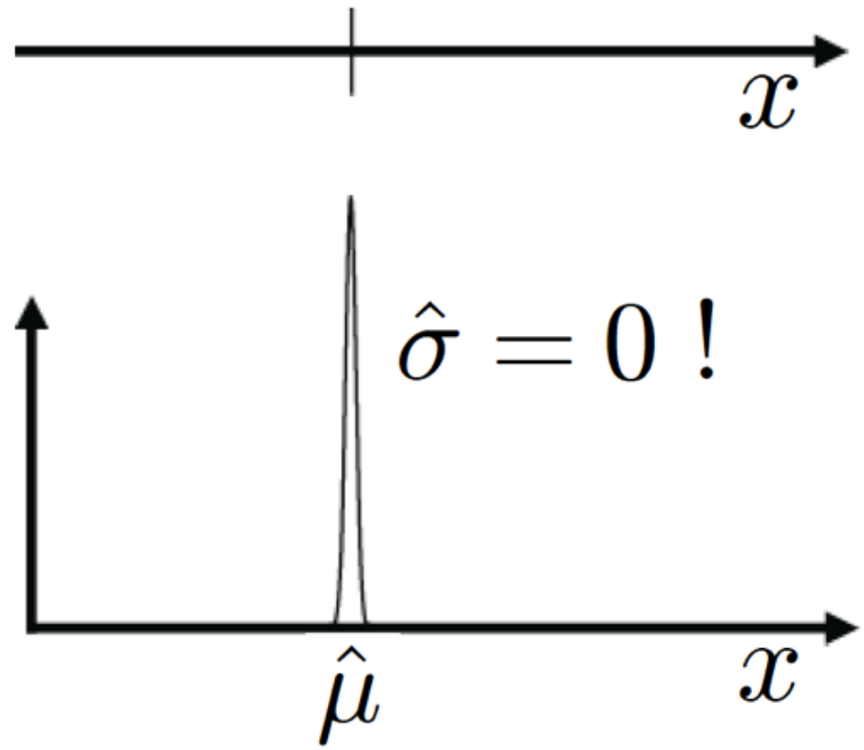
$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

Slide source: B. Leibe Slides ML, 2015

Maximum Likelihood Limitations

- Maximum Likelihood has several significant limitations
  - It systematically underestimates the variance of the distribution!
  - E.g. consider the case

$$N = 1, \mathrm{X} = \{x_1\}$$



=> 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.

Slide source: B. Leibe Slides ML, 2015

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…

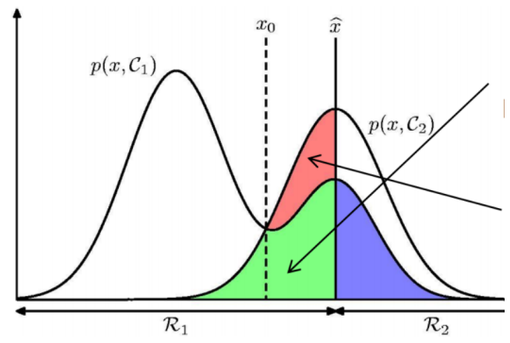
Slide source: B. Leibe Slides ML, 2015

Summary

Bayes Decision Theory

Classifying with Loss Function

Minimizing the Expected Loss

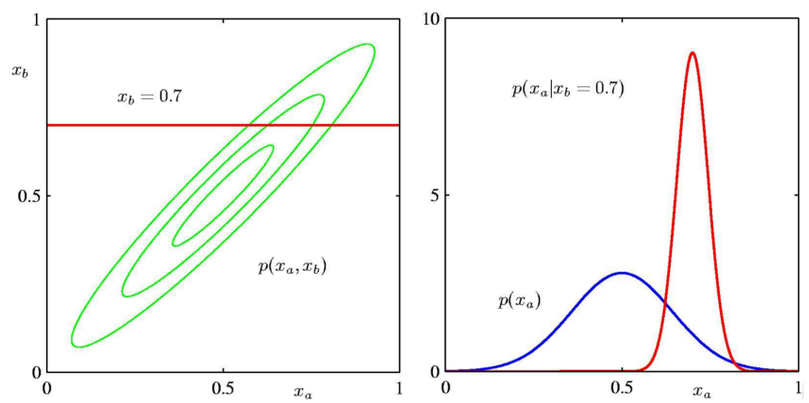


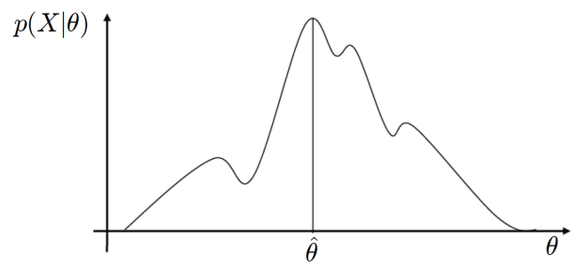
Probability Density Estimation

Gaussian Distribution

Parametric Methods

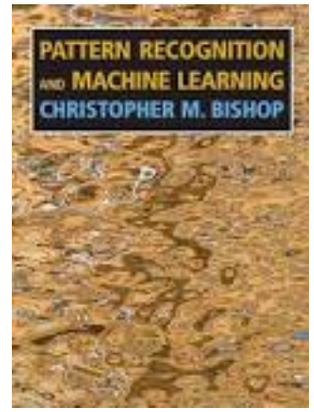
Maximum Likelihood





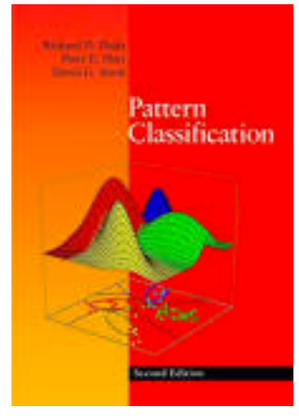
Readings

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.

Duda & Hart



- ML estimation Ch. 3.2.
- Bayesian Learning: Ch. 3.3-3.5
- Nonparametric methods: Ch. 4.1.-4.5.