Machine Learning – Lecture 2

Probability Density Estimation

15.04.2016

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

leibe@vision.rwth-aachen.de

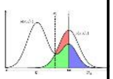
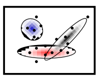
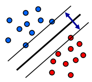
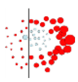
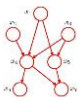
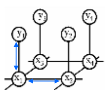
Many slides adapted from B. Schiele

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!

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

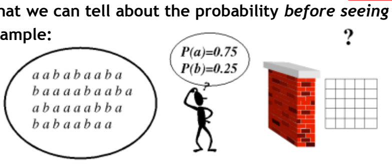







Topics of This Lecture

- 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) $p(C_k)$
  - What we can tell about the probability *before seeing the data*.
  - Example:



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

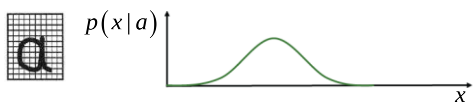
$C_2 = b$ $\qquad p(C_2) = 0.25$

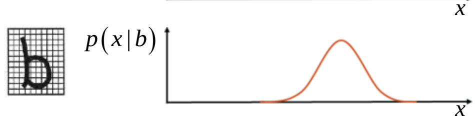
- In general: $\sum_k p(C_k) = 1$

Slide credit: Bernt Schiele

Recap: Bayes Decision Theory Concepts

- 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 credit: Bernt Schiele

Bayes Decision Theory Concepts

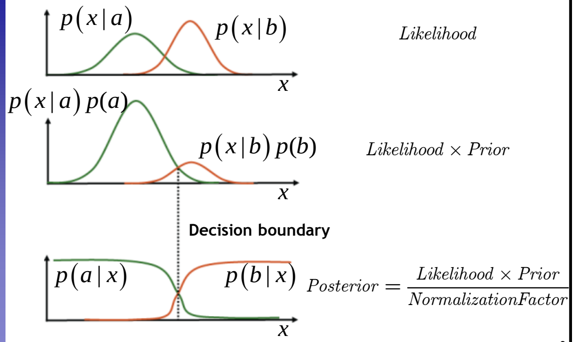
- Concept 3: Posterior probabilities $p(C_k \mid x)$
  - We are typically interested in the *a posteriori* probability, i.e. the probability of class C_k given the measurement vector x.
- Bayes' Theorem:

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

- Interpretation

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

Recap: Bayes Decision Theory



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

Recap: 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

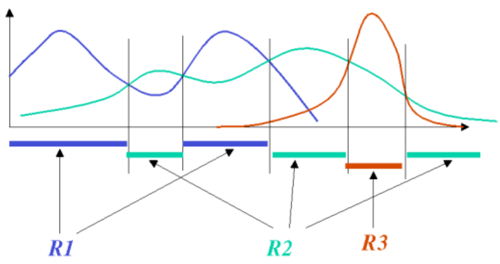
$$p(x|\mathcal{C}_1)p(\mathcal{C}_1) > p(x|\mathcal{C}_2)p(\mathcal{C}_2)$$

  - Which is again equivalent to (Likelihood-Ratio test)

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

Bayes Decision Theory

- Decision regions: $\mathcal{R}_1$, $\mathcal{R}_2$, $\mathcal{R}_3$, …



Recap: 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:

$$\begin{aligned}\mathbb{E}_{\alpha_1}[L] = R(\alpha_1|\mathbf{x}) &= L_{11}p(\mathcal{C}_1|\mathbf{x}) + L_{21}p(\mathcal{C}_2|\mathbf{x}) \\ \mathbb{E}_{\alpha_2}[L] = R(\alpha_2|\mathbf{x}) &= L_{12}p(\mathcal{C}_1|\mathbf{x}) + L_{22}p(\mathcal{C}_2|\mathbf{x})\end{aligned}$$

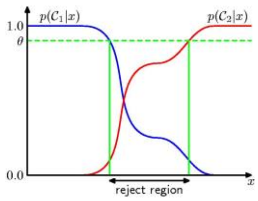
- Goal: Decide such that expected loss is minimized
  - I.e. decide α_1 if $R(\alpha_2|\mathbf{x}) > R(\alpha_1|\mathbf{x})$

Recap: Minimizing the Expected Loss

$$\begin{aligned}R(\alpha_2|\mathbf{x}) &> R(\alpha_1|\mathbf{x}) \\ L_{12}p(\mathcal{C}_1|\mathbf{x}) + L_{22}p(\mathcal{C}_2|\mathbf{x}) &> L_{11}p(\mathcal{C}_1|\mathbf{x}) + L_{21}p(\mathcal{C}_2|\mathbf{x}) \\ (L_{12} - L_{11})p(\mathcal{C}_1|\mathbf{x}) &> (L_{21} - L_{22})p(\mathcal{C}_2|\mathbf{x}) \\ \frac{(L_{12} - L_{11})}{(L_{21} - L_{22})} &> \frac{p(\mathcal{C}_2|\mathbf{x})}{p(\mathcal{C}_1|\mathbf{x})} = \frac{p(\mathbf{x}|\mathcal{C}_2)p(\mathcal{C}_2)}{p(\mathbf{x}|\mathcal{C}_1)p(\mathcal{C}_1)} \\ \frac{p(\mathbf{x}|\mathcal{C}_1)}{p(\mathbf{x}|\mathcal{C}_2)} &> \frac{(L_{21} - L_{22})}{(L_{12} - L_{11})}\frac{p(\mathcal{C}_2)}{p(\mathcal{C}_1)}\end{aligned}$$

⇒ Adapted decision rule taking into account the loss.

The Reject Option



- Classification errors arise from regions where the largest posterior probability $p(\mathcal{C}_k|\mathbf{x})$ 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.

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)

$$\begin{aligned} y_k(x) &= p(\mathcal{C}_k|x) \\ y_k(x) &= p(x|\mathcal{C}_k)p(\mathcal{C}_k) \\ y_k(x) &= \log p(x|\mathcal{C}_k) + \log p(\mathcal{C}_k) \end{aligned}$$

Different Views on the Decision Problem

- $y_k(x) \propto p(x|\mathcal{C}_k)p(\mathcal{C}_k)$
  - 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*
- $y_k(x) = p(\mathcal{C}_k|x)$
  - 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

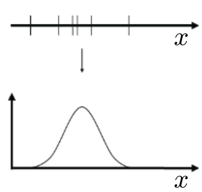
- Up to now
  - Bayes optimal classification
  - Based on the probabilities $p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$
- How can we estimate (=learn) those probability densities?
  - Supervised training case: data and class labels are known.
  - Estimate the probability density for each class $\mathcal{C}_k$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

$$p(\mathbf{x}|\mathcal{C}_k)$$

  - (For simplicity of notation, we will drop the class label $\mathcal{C}_k$ in the following.)

Probability Density Estimation

- Data: x_1, x_2, x_3, x_4, …
- Estimate: $p(x)$

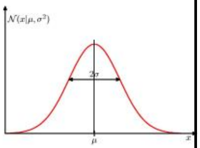


- Methods
  - Parametric representations (today)
  - Non-parametric representations (lecture 3)
  - Mixture models (lecture 4)

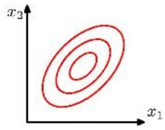
The Gaussian (or Normal) Distribution

- One-dimensional case
  - Mean μ
  - Variance σ^2

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



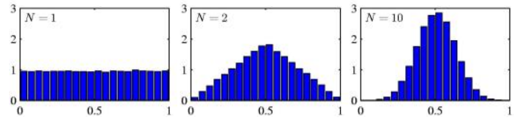
- Multi-dimensional case
  - Mean μ
  - Covariance Σ



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

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.

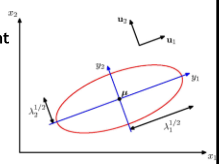


Gaussian Distribution - Properties

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

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

  - Here, Δ 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 $y_i = \mathbf{u}_i^{\mathrm{T}}(\mathbf{x}-\boldsymbol{\mu})$.

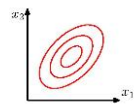
⇒ Constant density on ellipsoids with main directions along the eigenvectors $\mathbf{u}_i$ and scaling factors $\sqrt{\lambda_i}$.

Gaussian Distribution - Properties

- Special cases
  - Full covariance matrix

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

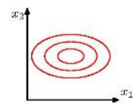
⇒ General ellipsoid shape



  - Diagonal covariance matrix

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

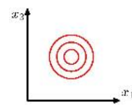
⇒ Axis-aligned ellipsoid



  - Uniform variance

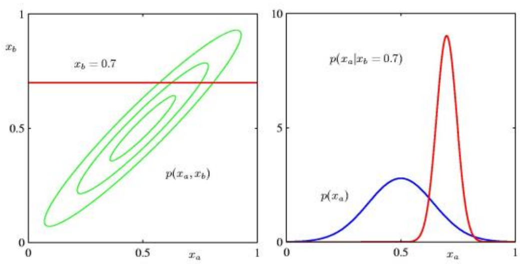
$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$$

⇒ Hypersphere



Gaussian Distribution - Properties

- The marginals of a Gaussian are again Gaussians:

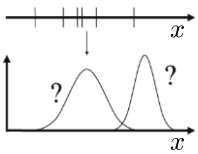


Topics of This Lecture

- 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 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

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

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

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

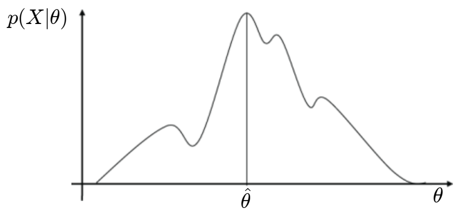
  - Log-likelihood

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

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

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.

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

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

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

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

Maximum Likelihood Approach

- Minimizing the log-likelihood

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

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

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

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

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

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

Maximum Likelihood Approach

- We thus obtain

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

- In a similar fashion, we get

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

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

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 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

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

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

⇒ The ML estimate will underestimate the true variance.

- Corrected estimate:

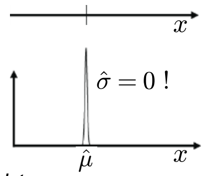
$$\tilde{\sigma}^2 = \frac{N}{N-1}\sigma^2_{\mathrm{ML}} = \frac{1}{N-1}\sum_{n=1}^{N}(x_n - \hat{\mu})^2$$

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.

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

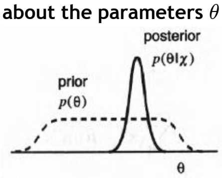
- To see the difference…
  - Suppose we want to estimate the uncertainty whether the Arctic ice cap will have disappeared by the end of the century.
  - This question makes no sense in a Frequentist view, since the event cannot be repeated numerous times.
  - In the Bayesian view, we generally have a 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.

$$\mathit{Posterior} \propto \mathit{Likelihood} \times \mathit{Prior}$$

  - 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 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.

Bayesian Learning Approach

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

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

Assumption: given θ, this doesn't depend on X anymore

$$p(x, \theta|X) = p(x|\theta, \cancel{X})p(\theta|X)$$

$$p(x|X) = \int \underbrace{p(x|\theta)}p(\theta|X) d\theta$$

This is entirely determined by the parameter θ (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

Likelihood of the parametric form θ given the data set X.

Estimate for x based on parametric form θ

Prior for the parameters θ

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

Normalization: integrate over all possible values of θ

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

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

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

⇒ The more uncertain we are about θ, the more we average over all parameter values.

Bayesian Density Estimation

- Problem
  - In the general case, the integration over θ is not possible (or only possible stochastically).
- Example where an analytical solution is possible
  - Normal distribution for the data, σ^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 μ,

$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$$

Bayesian Learning Approach

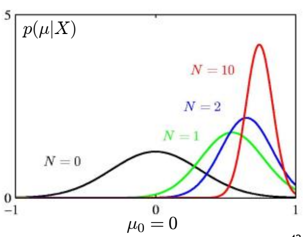
- Sample mean: $\bar{x} = \frac{1}{N}\sum_{n=1}^{N} x_n$
- Bayes estimate:

$$\mu_N = \frac{\sigma^2\mu_0 + N\sigma_0^2\bar{x}}{\sigma^2 + N\sigma_0^2}$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}$$

- Note:

	$N = 0$	$N \to \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0



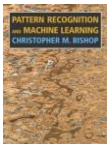
Summary: ML vs. Bayesian Learning

- 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 θ 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



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