Machine Learning – Lecture 2

Probability Density Estimation

10.10.2019

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

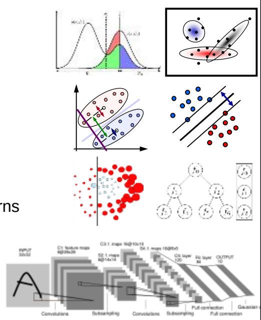
leibe@vision.rwth-aachen.de

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/
  - 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

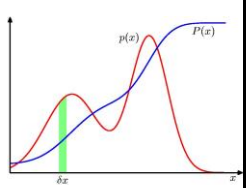
Bayes' Theorem $p(Y|X) = \frac{p(X|Y)p(Y)}{p(X)}$

where $p(X) = \sum_Y p(X|Y)p(Y)$

Probability Densities

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

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$



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

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

Image source: C.M. Bishop, 2006

Expectations

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

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] = \int p(x) f(x)\,\mathrm{d}x$$

discrete case — 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

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

RWTH AACHEN UNIVERSITY

Variances and Covariances

- The variance provides a measure how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$.

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

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

$$\begin{aligned}\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

- If $\mathbf{x}$ and $\mathbf{y}$ are vectors, the result is a covariance matrix

$$\begin{aligned}\mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\end{aligned}$$

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

Thomas Bayes, 1701-1761

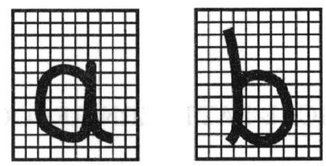
"The theory of inverse probability is founded upon an error, and must be wholly rejected."

R.A. Fisher, 1925

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

- Example: handwritten character recognition

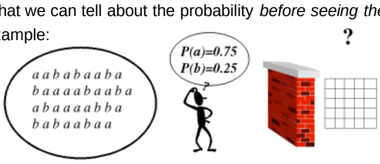


- Goal:
  - Classify a new letter such that the probability of misclassification is minimized.

RWTH AACHEN UNIVERSITY

Bayes Decision Theory

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



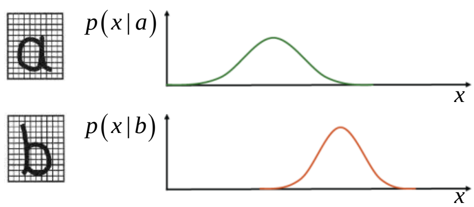
$$C_1 = a \qquad p(C_1) = 0.75$$
$$C_2 = b \qquad p(C_2) = 0.25$$

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

RWTH AACHEN UNIVERSITY

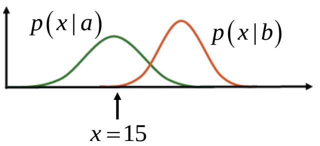
Bayes Decision Theory

- 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.



Bayes Decision Theory

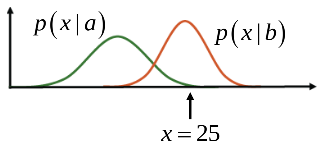
- Example:



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

Bayes Decision Theory

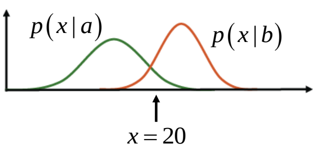
- Example:



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

Bayes Decision Theory

- Example:



- Question:
  - Which class?
  - Remember that $p(a) = 0.75$ and $p(b) = 0.25$…
  - I.e., the decision should be again 'a'.
  - ⇒ How can we formalize this?

Bayes Decision Theory

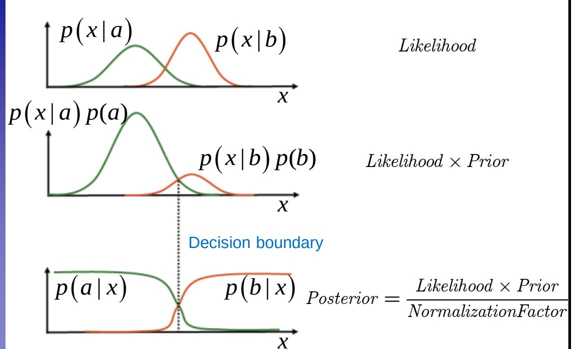
- Concept 3: Posterior probabilities $p(C_k|x)$
  - We are typically interested in the *a posteriori* probability, i.e., the probability of class C_k given the measurement vector x.
- Bayes' Theorem:

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

- Interpretation

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

Bayes Decision Theory

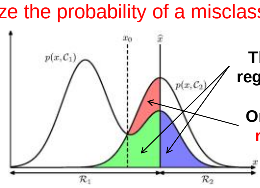


Bayesian Decision Theory

- Goal: Minimize the probability of a misclassification

Decision rule:
$x < \hat{x} \Rightarrow C_1$
$x \geq \hat{x} \Rightarrow C_2$

How does p(mistake) change when we move $\hat{x}$?



The green and blue regions stay constant.

Only the size of the red region varies!

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

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)

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

Machine Learning Winter '19

Slide credit: Bernt Schiele

B. Leibe

27

Generalization to More Than 2 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

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

Machine Learning Winter '19

Slide credit: Bernt Schiele

B. Leibe

28

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:

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

Machine Learning Winter '19

Slide credit: Bernt Schiele

B. Leibe

29

Classifying with Loss Functions

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

$$L_{kj} = loss\ for\ decision\ \mathcal{C}_j\ if\ truth\ is\ \mathcal{C}_k.$$

- Example: cancer diagnosis

Decision

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

Machine Learning Winter '19

B. Leibe

30

Classifying with Loss Functions

- Loss functions may be different for different actors.
  - Example:

"invest" *"don't invest"*

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

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

⇒ Different loss functions may lead to different Bayes optimal strategies.

Machine Learning Winter '19

B. Leibe

31

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

Machine Learning Winter '19

B. Leibe

32

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:

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

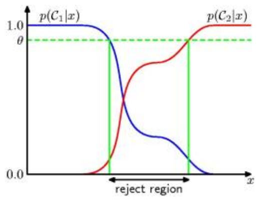
- Goal: Decide such that expected loss is minimized
  - I.e. decide α_1 if $R(\alpha_2|\mathbf{x}) > R(\alpha_1|\mathbf{x})$

Minimizing the Expected Loss

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

$\Rightarrow$ 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.
  - $\Rightarrow$ *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.
  - $\Rightarrow$ *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

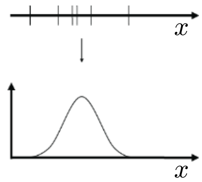
- Up to now
  - Bayes optimal classification
  - Based on the probabilities $p(\mathbf{x}|\mathcal{C}_k)p(\mathcal{C}_k)$
- How can we estimate (= learn) those probability densities?
  - Supervised training case: data and class labels are known.
  - Estimate the probability density for each class $\mathcal{C}_k$ 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: $x_1, x_2, x_3, x_4, \ldots$
- Estimate: $p(x)$

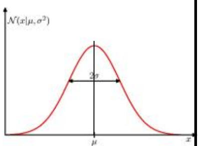


- Methods
  - Parametric representations (today)
  - Non-parametric representations (lecture 3)
  - Mixture models (lecture 4)

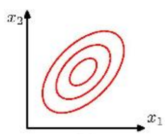
The Gaussian (or Normal) Distribution

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

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



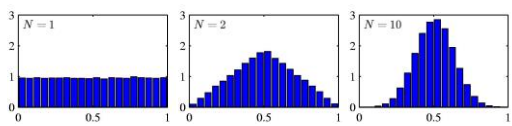
- Multi-dimensional case
  - Mean μ
  - Covariance Σ



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

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.

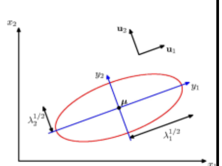


Gaussian Distribution – Properties

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

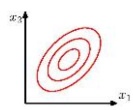
$\Rightarrow$ Constant density on ellipsoids with main directions along the eigenvectors $\mathbf{u}_i$ and scaling factors $\sqrt{\lambda_i}$

Gaussian Distribution – Properties

- Special cases
  - Full covariance matrix

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

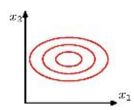
$\Rightarrow$ General ellipsoid shape



  - Diagonal covariance matrix

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

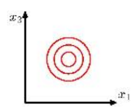
$\Rightarrow$ Axis-aligned ellipsoid



  - Uniform variance

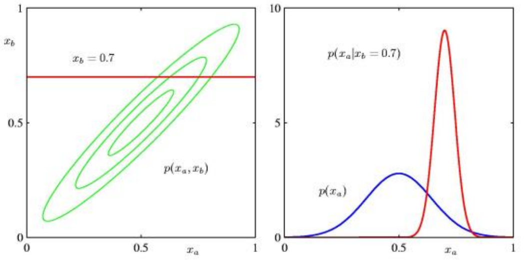
$$\boldsymbol{\Sigma} = \sigma^2\mathbf{I}$$

$\Rightarrow$ Hypersphere



Gaussian Distribution – Properties

- The marginals of a Gaussian are again Gaussians:



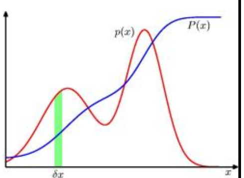
Topics of This Lecture

- 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

Probability Densities

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

$$p(x \in (a,b)) = \int_a^b p(x)\,\mathrm{d}x$$



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

$$P(z) = \int_{-\infty}^{z} p(x)\,\mathrm{d}x$$

Expectations

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

$$\mathbb{E}[f] = \sum_x p(x) f(x) \qquad \mathbb{E}[f] = \int p(x) f(x)\,\mathrm{d}x$$

discrete case — 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

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

Variances and Covariances

- The variance provides a measure how much variability there is in $f(x)$ around its mean value $\mathbb{E}[f(x)]$.

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

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

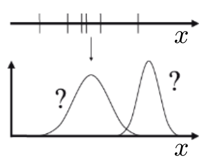
$$\begin{aligned}\mathrm{cov}[x,y] &= \mathbb{E}_{x,y}\left[\{x - \mathbb{E}[x]\}\{y - \mathbb{E}[y]\}\right] \\ &= \mathbb{E}_{x,y}[xy] - \mathbb{E}[x]\mathbb{E}[y]\end{aligned}$$

- If $\mathbf{x}$ and $\mathbf{y}$ are vectors, the result is a covariance matrix

$$\begin{aligned}\mathrm{cov}[\mathbf{x},\mathbf{y}] &= \mathbb{E}_{\mathbf{x},\mathbf{y}}\left[\{\mathbf{x} - \mathbb{E}[\mathbf{x}]\}\{\mathbf{y}^{\mathrm{T}} - \mathbb{E}[\mathbf{y}^{\mathrm{T}}]\}\right] \\ &= \mathbb{E}_{\mathbf{x},\mathbf{y}}[\mathbf{x}\mathbf{y}^{\mathrm{T}}] - \mathbb{E}[\mathbf{x}]\mathbb{E}[\mathbf{y}^{\mathrm{T}}]\end{aligned}$$

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

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

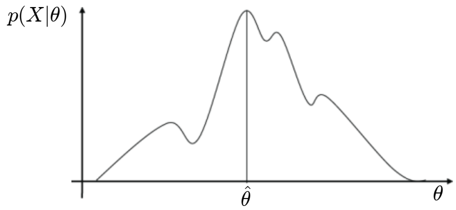
  - Log-likelihood

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

  - 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)

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

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

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

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

Maximum Likelihood Approach

- We thus obtain

$$\hat{\mu} = \frac{1}{N} \sum_{n=1}^{N} x_n \qquad \textit{“sample mean”}$$

- In a similar fashion, we get

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

- $\hat{\theta} = (\hat{\mu}, \hat{\sigma})$ 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

$$\begin{aligned} \mathbb{E}(\mu_{\text{ML}}) &= \mu \\ \mathbb{E}(\sigma^2_{\text{ML}}) &= \left(\frac{N-1}{N}\right)\sigma^2 \end{aligned}$$

⇒ The ML estimate will underestimate the true variance.

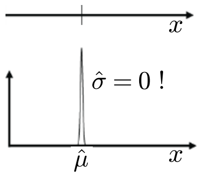
- Corrected estimate:

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

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…

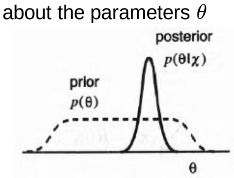
Bayesian vs. Frequentist View

- 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

- Bayesian Learning is an important concept
  - However, it would lead to far here.

  ⇒ 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