COM2004/3004
Data Driven Computing

Week 3a: Parameter Estimation

Autumn Semester

Overview

Overview

Bayesian Classifiers

- the normal distribution
- parameter estimation
- classification
- Bayes decision theory
- risk
- ROC (receiver operating characteristic)
- Curse of dimensionality and naive Bayes classification

Review

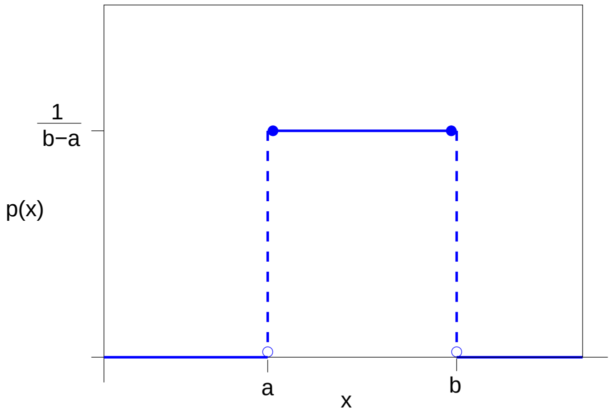
Discrete vs. continuous probabilities

Remember from last Friday,

- Discrete probability distributions, $P(X)$, represent the probability of discrete observations,
  - they can be represented by tables
  - e.g. for a dice roll, P(X=1) = 1/6, P(X=2)=1/6 etc.
  - table entries must sum to 1.
- Continuous probability distributions, $p(x)$, represent probability of continuous observerations
  - they can't be tabulated
  - we define the probability density function, pdf (gradient of the cdf).
  - area under the pdf must be 1
  - the pdf can be represented by a function p(x)

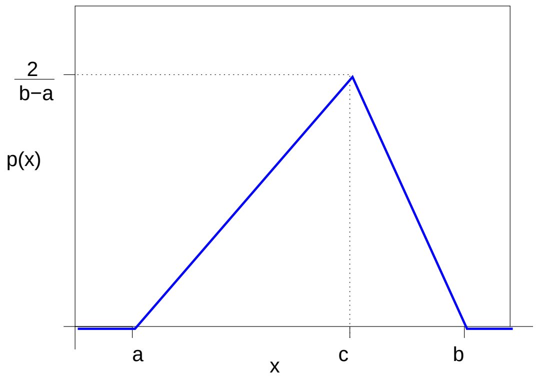
The uniform distribution

$$p(x; a, b) = \mathcal{U}(a, b) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \text{otherwise} \end{cases}$$

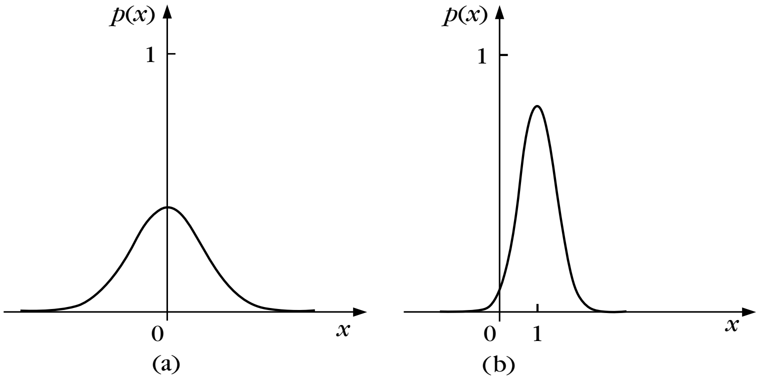


The triangle distribution

$$p(x; a, b, c) = \begin{cases} \frac{2(x-a)}{(b-a)(c-a)} & x \in [a, c] \\ \frac{2(b-x)}{(b-a)(b-c)} & x \in [c, b] \\ 0 & \text{otherwise} \end{cases}$$



Normal Distribution



Normal Distribution

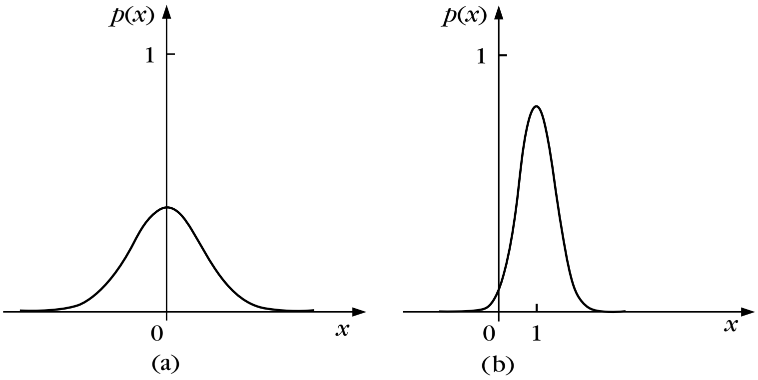
Univariate normal distribution

- pdf:

$$p(x;\mu,\sigma) = \underbrace{\frac{1}{\sqrt{2\pi\sigma^2}}}_{\text{normalisation}} \overbrace{\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)}^{\text{shape}}$$

  - parameter μ controls the position of the peak
  - parameter σ controls the width of the peak
  - defined for $-\infty < x < \infty$
  - also referred to as the Gaussian distribution

Normal Distribution



(a) $\mu = 0$, $\sigma^2 = 1$ (b) $\mu = 1$, $\sigma^2 = 0.2$

Normal Distribution

– mean:

$$
\begin{aligned}
E[X] &= \int_{-\infty}^{\infty} x p(x) dx \\
&= \dots \\
&=
\end{aligned}
$$

– variance:

$$
\begin{aligned}
V[X] &= E[(x-\mu)^2] \\
&= \int_{-\infty}^{\infty} (x-\mu)^2 p(x) dx \\
&= \dots \\
&=
\end{aligned}
$$

Normal Distribution

- mean:

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$
$$= \dots$$
$$= \mu$$

- variance:

$$V[X] = E[(x-\mu)^2]$$
$$= \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx$$
$$= \dots$$
$$=$$

Normal Distribution

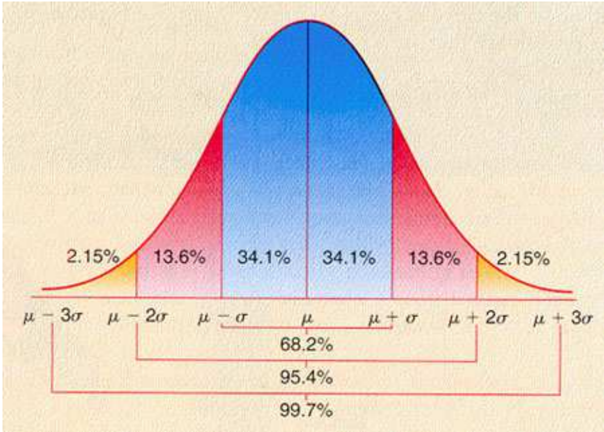
- mean:

$$E[X] = \int_{-\infty}^{\infty} xp(x)dx$$
$$= \dots$$
$$= \mu$$

- variance:

$$V[X] = E[(x-\mu)^2]$$
$$= \int_{-\infty}^{\infty} (x-\mu)^2 p(x)dx$$
$$= \dots$$
$$= \sigma^2$$

Normal Distribution



σ is referred to as the standard deviation

Normal Distribution

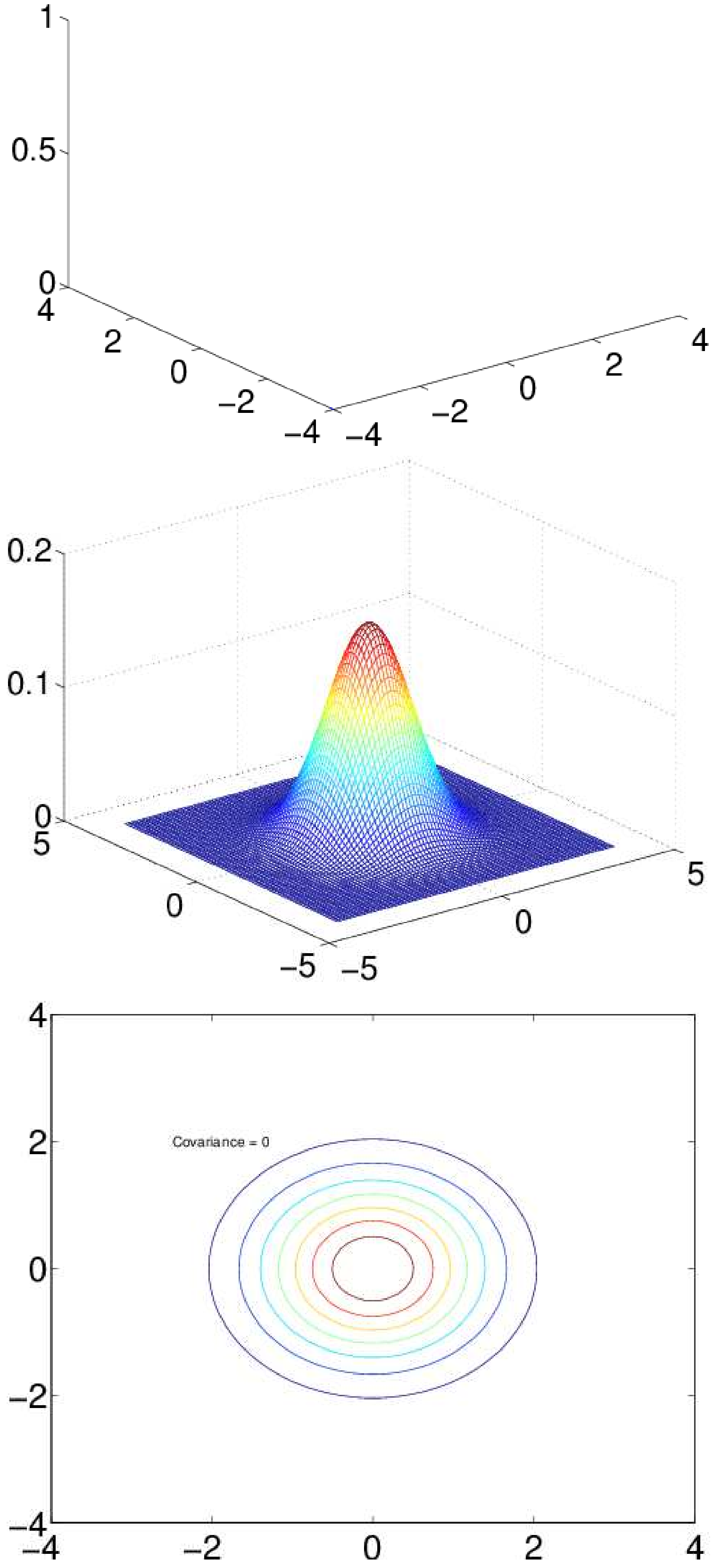
(*e.g.*) Weight of one apple is normally distributed with the mean 300 grams and the standard deviation 50 grams. What is the probability that the weight is between 200 and 400 grams? From the previous figure we see it is roughly 95%.

(*e.g.*) Suppose that the weight of a chicken egg is normally distributed, and the average (μ) is 60 grams and the standard deviation (σ) is 5 grams. Eggs are classified by weight into three categories: $\mu - \sigma$ or less (small), $\mu + \sigma$ or more (large), and the rest (medium). Further, one egg can be sold at 10p (small), 12p (medium), or 16p (large). If selling 1000 eggs how much will the farmer expect to earn?

$$0.10 \times 159 + 0.12 \times 682 + 0.16 \times 159 = 123.18 \quad \text{(pounds)}$$

2D Normal Distribution

$p(x_1, x_2)$



Normal Distribution

Multivariate normal distribution

□ L-dimensional pdf:

$$p(\boldsymbol{x}; \boldsymbol{\mu}, \Sigma) = \frac{1}{\sqrt{(2\pi)^L |\Sigma|}} \exp\left(-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x} - \boldsymbol{\mu})\right)$$

- mean $\boldsymbol{\mu}$ is a column vector with L elements
- Σ is an $L \times L$ covariance matrix
  $|\Sigma|$ is the determinant and Σ^{-1} is the inverse
- denoted as $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$

Normal Distribution

(*e.g.*) $L = 1$:

$$\Rightarrow \ \boldsymbol{x} = (x), \ \boldsymbol{\mu} = (\mu), \ \Sigma = (\sigma_{11}) = (\sigma^2)$$

$$\begin{aligned} p(x) &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2}(x-\mu)^T(\sigma^2)^{-1}(x-\mu)\right) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \end{aligned}$$

Normal Distribution

(*e.g.*) $L = 2$:

$$\Rightarrow \ \boldsymbol{x} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \ \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \ \Sigma = \begin{pmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{21} & \sigma_{22} \end{pmatrix}$$

calculating $\boldsymbol{x} - \boldsymbol{\mu}$, $|\Sigma|$ and Σ^{-1}

$$\begin{aligned} \boldsymbol{x} - \boldsymbol{\mu} &= \begin{pmatrix} x_1 - \mu_1 \\ x_2 - \mu_2 \end{pmatrix} \\ |\Sigma| &= \sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21} \\ \Sigma^{-1} &= \frac{1}{\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21}} \begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix} \end{aligned}$$

(continued)

Normal Distribution

calculating $p(\boldsymbol{x})$

$$p(\boldsymbol{x}) = \frac{1}{\sqrt{(2\pi)^2(\sigma_{11}\sigma_{22} - \sigma_{12}\sigma_{21})}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

where

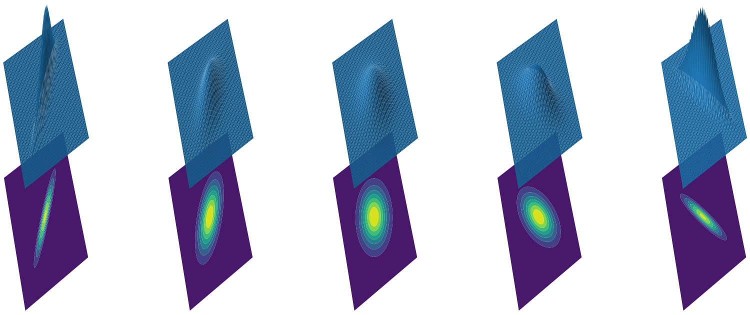
$$\begin{aligned}
&(\boldsymbol{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) \\
&= (x_1-\mu_1, x_2-\mu_2)\frac{1}{\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21}}\begin{pmatrix} \sigma_{22} & -\sigma_{12} \\ -\sigma_{21} & \sigma_{11} \end{pmatrix}\begin{pmatrix} x_1-\mu_1 \\ x_2-\mu_2 \end{pmatrix} \\
&= \frac{\sigma_{22}(x_1-\mu_1)^2 - (\sigma_{12}+\sigma_{21})(x_1-\mu_1)(x_2-\mu_2) + \sigma_{11}(x_2-\mu_2)^2}{\sigma_{11}\sigma_{22}-\sigma_{12}\sigma_{21}}
\end{aligned}$$

18 / 28

Normal Distribution

When we know the parameters, i.e. μ and Σ

$\Rightarrow$ simple to plot the probability distribution



19 / 28

Parameter Estimation

- Typically we have some data but don't know the parameters of the underlying distribution.
- But we need to know the parameters in order to build classifiers etc
- So how do we estimate the distribution parameters given some data samples?

Parameter Estimation

- parameter estimation: data $\Rightarrow$ distribution parameters

- many approaches
  - maximum likelihood (ML) estimation
  - maximum a posteriori probability (MAP) estimation
  - maximum entropy (MaxEnt) estimation
  - non parametric probability density estimation

    ...

    ...

Parameter Estimation

X is a set of random samples drawn from pdf $p(\boldsymbol{x};\boldsymbol{\theta})$:

$$X = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \dots, \boldsymbol{x}_N\}$$

- $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N$ are statistically independent
- $p(\boldsymbol{x};\boldsymbol{\theta})$ is a shorthand for $p(\boldsymbol{x}|\omega_i;\boldsymbol{\theta}_i)$
  - feature vectors $\boldsymbol{x}$ in class ω_i are distributed according to the pdf $p(\boldsymbol{x}|\omega_i;\boldsymbol{\theta}_i)$
  - $\boldsymbol{\theta}_i$ is there to show that pdf is parametrised

Likelihood function of $\boldsymbol{\theta}$ with respect to X:

$$p(X;\boldsymbol{\theta}) \equiv p(\boldsymbol{x}_1, \dots, \boldsymbol{x}_N;\boldsymbol{\theta}) = \prod_{k=1\dots N} p(\boldsymbol{x}_k;\boldsymbol{\theta})$$

Parameter Estimation

Maximum likelihood estimate:

$$\hat{\boldsymbol{\theta}}_{ML} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ p(X;\boldsymbol{\theta})$$

- how to solve this?

  (hint) the gradient of the likelihood must be zero at $\hat{\boldsymbol{\theta}}_{ML}$:

$$\frac{\partial p(X;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

Parameter Estimation

Solution: maximum likelihood estimate

- use the log likelihood function:

$$L(\boldsymbol{\theta}) = \ln p(X; \boldsymbol{\theta}) = \ln \prod_{k=1\ldots N} p(\boldsymbol{x}_k; \boldsymbol{\theta}) = \sum_{k=1\ldots N} \ln p(\boldsymbol{x}_k; \boldsymbol{\theta})$$

hence the gradient:

$$\frac{\partial L(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{k=1\ldots N} \frac{\partial \ln p(\boldsymbol{x}_k; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \sum_{k=1\ldots N} \frac{1}{p(\boldsymbol{x}_k; \boldsymbol{\theta})} \frac{\partial p(\boldsymbol{x}_k; \boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = 0$$

(note) $p(X; \boldsymbol{\theta})$ and $\ln p(X; \boldsymbol{\theta})$ attain the maximum with the same $\boldsymbol{\theta}$ because the logarithmic function is monotone

24 / 28

Parameter Estimation

Example: normal distribution parameters

- Assume that N data points $x_1, \ldots, x_N$ have been generated by 1-dimensional normal pdf with unknown mean μ and unknown variance σ:

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

- the log likelihood:

$$\begin{aligned} L(\mu, \sigma^2) &= \ln p(x_1, \ldots, x_N; \mu, \sigma^2) \\ &= \ln \prod_{k=1\ldots N} p(x_k; \mu, \sigma^2) \\ &= -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{k=1\ldots N} (x_k - \mu)^2 \end{aligned}$$

25 / 28

Parameter Estimation

Example: normal distribution parameters (continued)

- the mean

$$\frac{\partial L(\mu, \sigma^2)}{\partial \mu} = \frac{1}{2\sigma^2} \sum_{k=1\ldots N} (x_k - \mu) = 0 \quad \Rightarrow$$

$$\hat{\mu}_{ML} = \frac{1}{N} \sum_{k=1\ldots N} x_k$$

- the variance

$$\frac{\partial L(\mu, \sigma^2)}{\partial \sigma^2} = -\frac{N}{2\sigma^2} - \frac{1}{2\sigma^4} \sum_{k=1\ldots N} (x_k - \mu)^2 = 0 \quad \Rightarrow$$

$$\hat{\sigma}^2_{ML} = \frac{1}{N} \sum_{k=1\ldots N} (x_k - \mu)^2$$

Summary

Summary

- Introduced some simple continuous distibutions
- Introduced the normal distribution
  - 1st in the 1-D (univariate) case...
  - ... then generalised to N dimension
- Introduced concept of Maximum Likelihood (ML) parameter estimation
- Example: estimating parameters of the univariate normal distriubtion

What next?

- ☐ Tutorial session
  - Problems to aid understanding of continuous probability
  - Lab class briefing
- ☐ Lab Class
  - Parameter estimation and classification
- ☐ Lecture on Friday
  - Bayesian Classification Theory
  - Lab class review