Pattern Recognition

Parameter Estimation of Probability Density Functions

Classification Problem (Review)

- The classification problem is to assign an arbitrary feature vector $x \in F$ to one of c classes.
- The classifier is a function from the feature space onto the set of classes,
  $\alpha : F \rightarrow \{ \omega_1 , \ldots , \omega_c \}$. ($\alpha(x)$ is the classifier)
- The parts of the feature space corresponding to classes $\omega_1 , \ldots , \omega_c$ are denoted by $R_1 , \ldots , R_c$

Classification Problem

- Feature vectors x that we aim to classify belong to the feature space *F*.
- The task is to assign an arbitrary feature vector $x \in F$ to one of the *c* classes
- We know the
  - 1. prior probabilities $P(\omega_1), \ldots, P(\omega_c)$ of the classes and
  - 2. the class conditional probability density functions $p(x \mid \omega_1), \ldots, p(x \mid \omega_c)$.

The Bayes classification rule (for two classes M=2)

- Given $\underline{x}$ classify it according to the rule

$$If\ P(\omega_1|\underline{x}) > P(\omega_2|\underline{x})\ \ \underline{x} \rightarrow \omega_1$$
$$If\ P(\omega_2|\underline{x}) > P(\omega_1|\underline{x})\ \ \underline{x} \rightarrow \omega_2$$

- Equivalently: classify $\underline{x}$ according to the rule

$$p(\underline{x}|\omega_1)P(\omega_1)(><)p(\underline{x}|\omega_2)P(\omega_2)$$

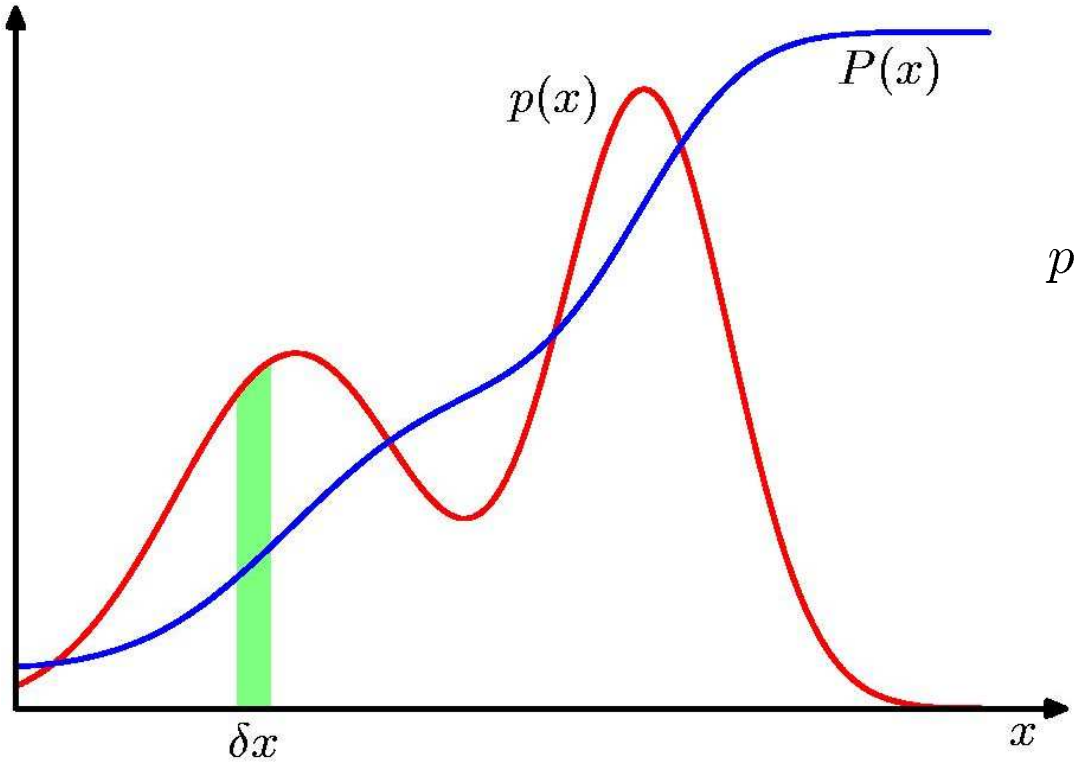
- For equiprobable classes the test becomes

$$p(\underline{x}|\omega_1)(><)P(\underline{x}|\omega_2)$$

Probability Density Function

- A **probability density function** (**pdf**), or **density** of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value.

Probability Densities



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

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

$$p(x) \geqslant 0$$

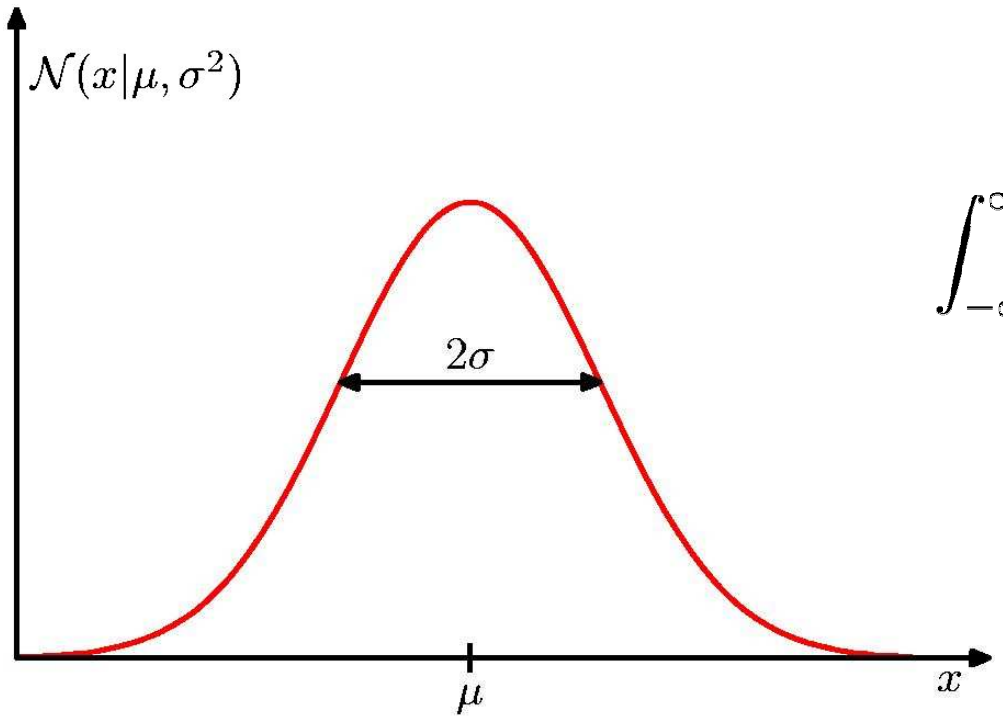
$$\int_{-\infty}^{\infty} p(x)\,\mathrm{d}x = 1$$

Cumulative Distribution Function

- **Cumulative distribution function** (**CDF**), or just **distribution function**, describes the probability that a real-valued random variable *X* with a given probability distribution will be found at a value less than or equal to *x*.

The Gaussian Distribution

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



$$\mathcal{N}(x|\mu,\sigma^2) > 0$$

$$\int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right)\,\mathrm{d}x = 1$$

Gaussian Mean and Variance

$$\mathbb{E}[x] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x \, \mathrm{d}x = \mu$$

$$\mathbb{E}[x^2] = \int_{-\infty}^{\infty} \mathcal{N}\left(x|\mu,\sigma^2\right) x^2 \, \mathrm{d}x = \mu^2 + \sigma^2$$

$$\mathrm{var}[x] = \mathbb{E}[x^2] - \mathbb{E}[x]^2 = \sigma^2$$

Estimating Class Conditional PDFs and a-Priori Probabilities

- In practice, PDFs and a-priori probabilities are not known.
- PDFs and a-priori probabilities are estimated from training samples. (Supervised learning)
- We assume that the training samples are occurrences of the independent random variables. (That is: they were measured from different objects.)
- These random variables are assumed to be distributed according to $p(x|\omega_i)$. (independent and identically distributed (i.i.d.)).

Training Data Types

- Mixture Sampling: A set of objects are randomly selected, their feature vectors are computed and then the objects are hand-classified to the most appropriate classes.

- Separate Sampling: The training data for each class is collected separately.

Estimating a-Priori Probabilities

- For the classifier training, the difference of the two sampling techniques is:

  based on the mixed sampling we can deduce the a-priori probabilities P (ω_1), . . . , P (ω_c) as:

$$P(\omega_i) = \frac{n_i}{\sum_{j=1}^{c} n_j}.$$

Parametric Estimation of PDFs

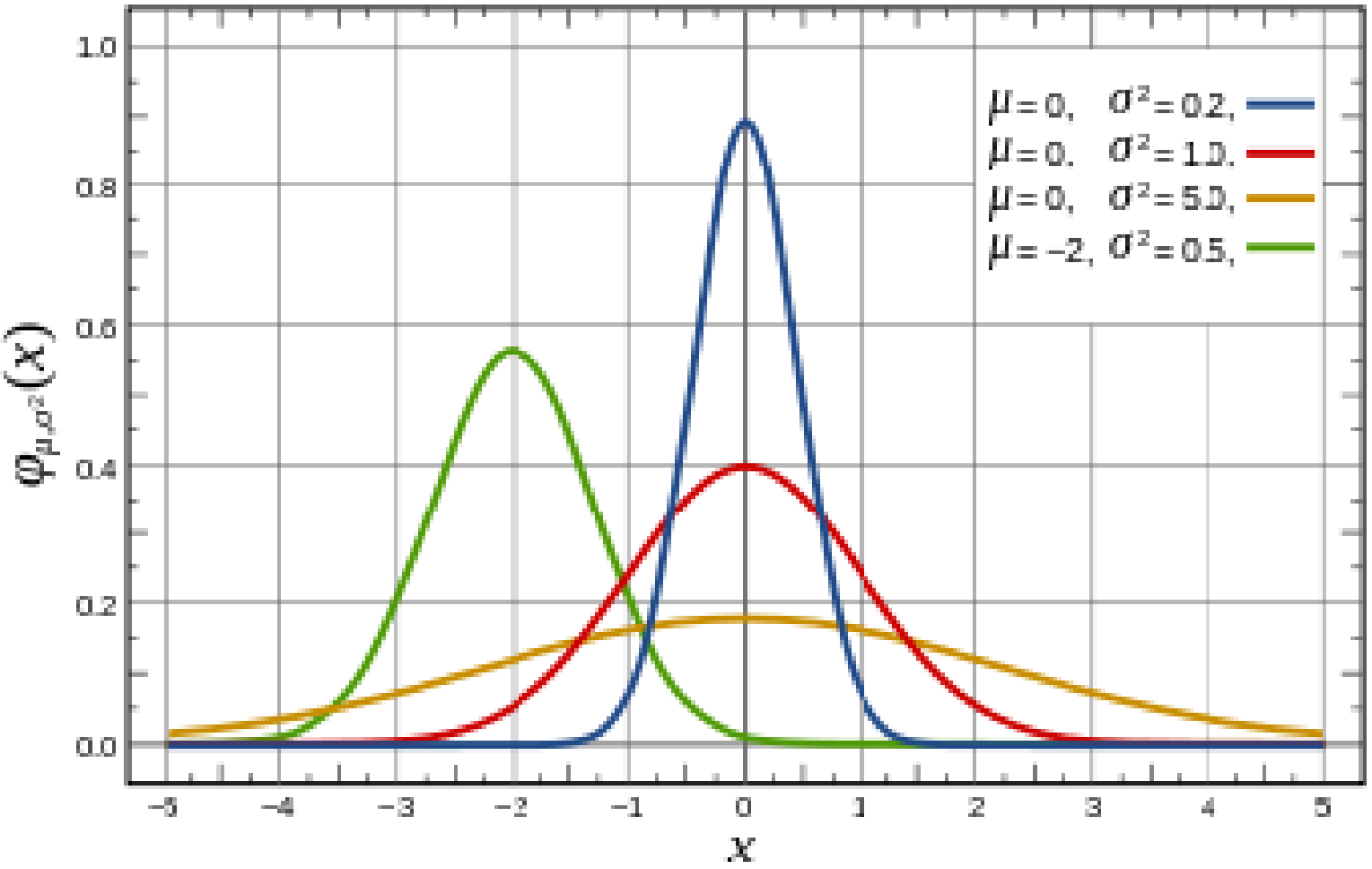
- If we assume that $p(x|\omega_i)$ belongs to some family of parametric distributions, the class conditional pdfs $p(x|\omega_i)$ is reduced to the estimation of the parameter vector θ_i
- For example, we can assume that $p(x|\omega_i)$ is a normal density with unknown parameters $\theta_i = (\mu_i , \Sigma_i)$.

Most Common PDF

- **Normal distribution:** Related to real-valued quantities that grow linearly (e.g. errors, offsets)

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

Normal Distribution

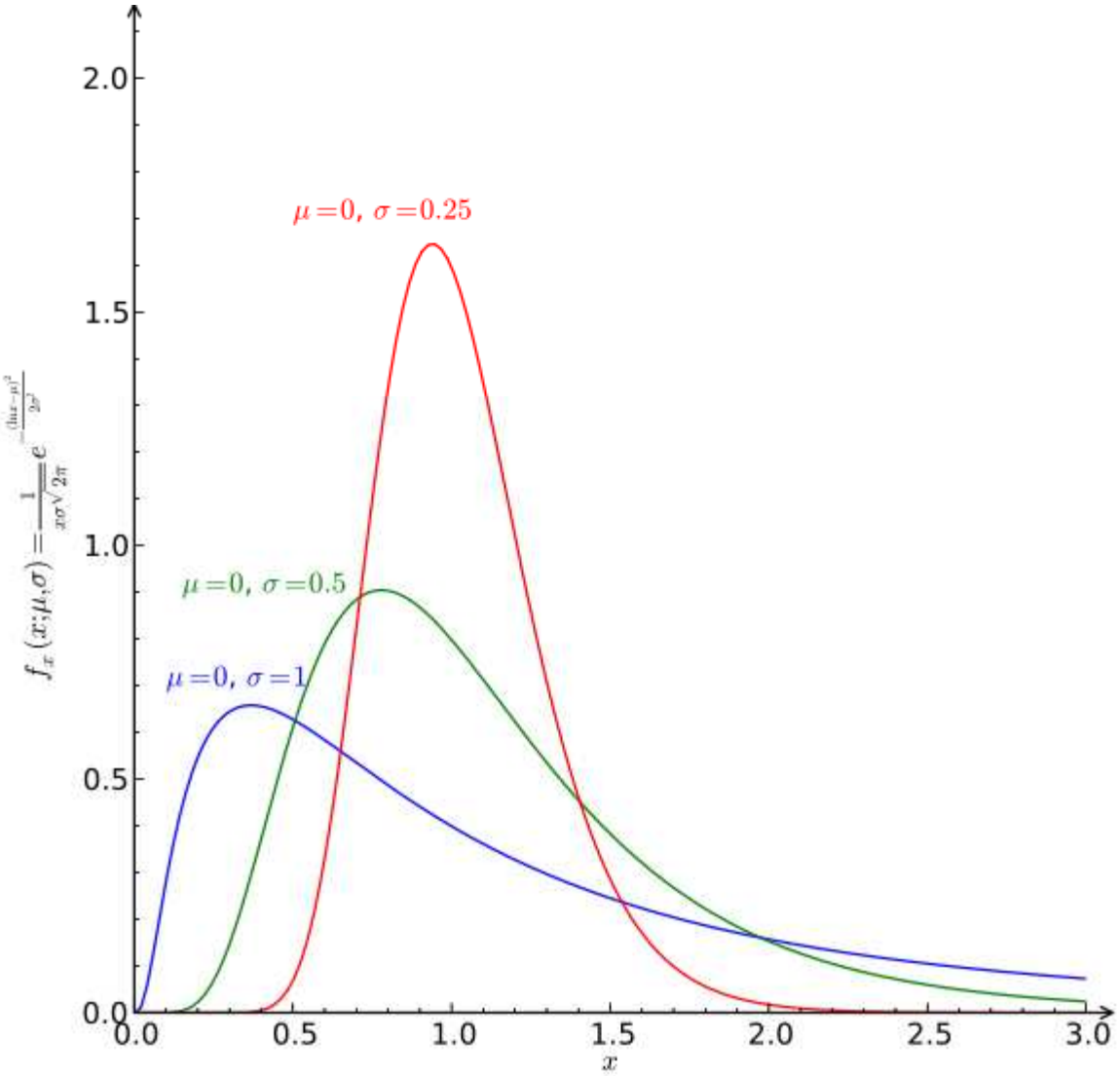


Log-normal distribution

- Related to positive real-valued quantities that grow exponentially

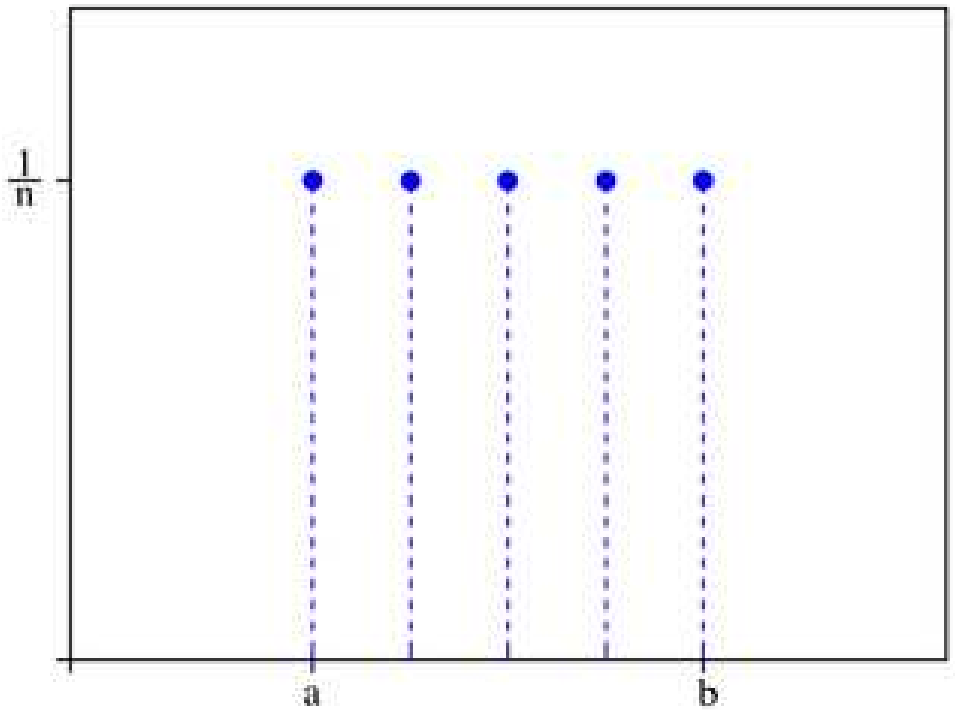
$$f_X(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}}\, e^{-\frac{(\ln x-\mu)^2}{2\sigma^2}}, \quad x > 0$$

Log-Normal Distribution



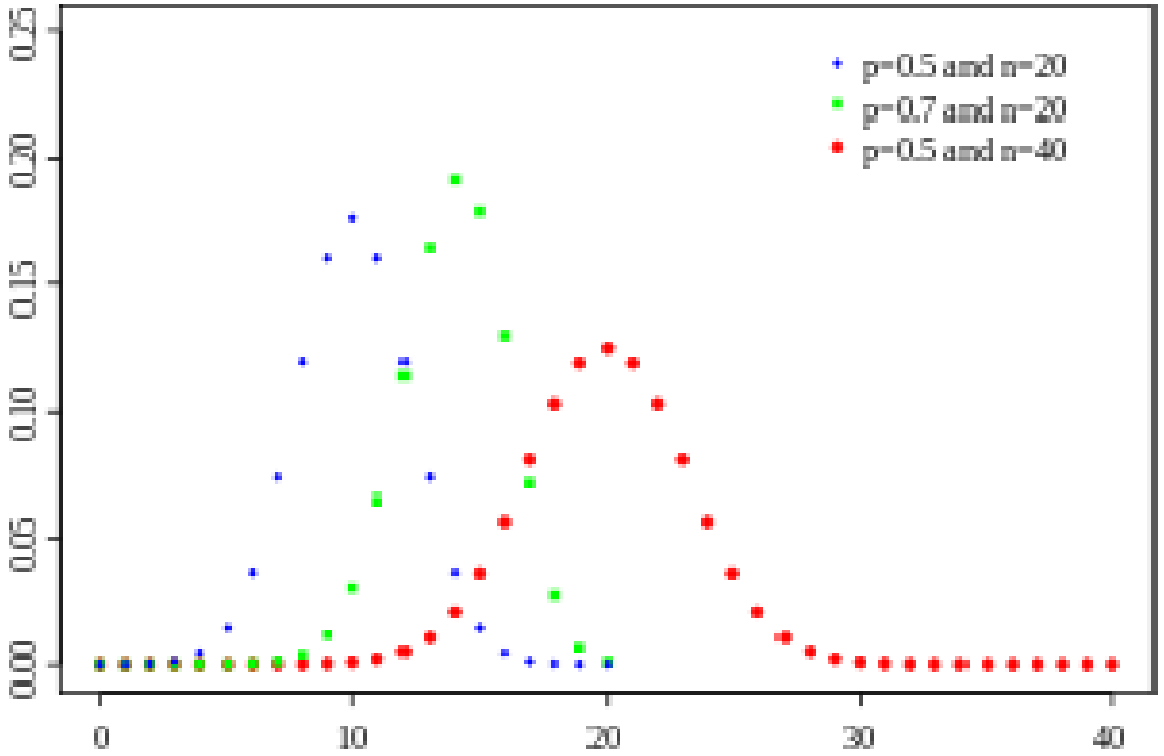
Uniform distribution (discrete)

- Related to real-valued quantities that are assumed to be uniformly distributed over a (possibly unknown) region



Binomial Distribution

- The binomial distribution is the discrete probability distribution of the number of successes in a sequence of n independent yes/no experiments, each of which yields success with probability p



Maximum Likelihood Estimation

- The aim is to estimate the value of the parameter vector θ based on the training samples $D = \{ x_1 , \ldots , x_n \}$.
- We assume that the training samples are occurrences of i.i.d. random variables distributed according to the density $p(x|\theta)$.

Maximum Likelihood Estimation

- The maximum likelihood estimate or the ML estimate θ' maximizes the probability of the training data with respect to θ. Due to the i.i.d. assumption, the probability of D is

$$p(\mathcal{D}|\theta) = \prod_{i=1}^{n} p(\mathbf{x}_i|\theta).$$

Finding ML-Estimate

- Derive the gradient of the likelihood function or the log-likelihood function
- Solve the zeros of the gradient and search also all the other critical points of the likelihood function. (e.g. the points where at least one of the partial derivatives of the function are not defined are critical in addition to the points where gradient is zero.)

Finding ML-Estimate

- Evaluate the likelihood function at the critical points and select the critical point with the highest likelihood value as the estimate.
- warning: the ML-estimate does not necessarily exist for all distributions, for example the likelihood function could grow without limit.

ML-estimates for the Normal Density

- ML-estimates for the normal density

$$p(\mathbf{x}|\theta) = \frac{1}{(2\pi)^{d/2}\sqrt{\det(\Sigma)}} \exp[-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)].$$

- Assuming covariance is fixed and known we have:

$$\sum_{i=1}^{n} \ln p(\mathbf{x}_i|\mu) = -\sum_{i=1}^{n} 0.5(\ln[(2\pi)^{d/2}\det(\Sigma)] + (\mathbf{x}_i-\mu)^T\Sigma^{-1}(\mathbf{x}_i-\mu)).$$

ML-estimates for the Normal Density

- The gradient is:

$$\nabla l(\mu) = -\sum_{i=1}^{n} \Sigma^{-1}(\mathbf{x}_i - \mu).$$

- Setting gradient equal to zero we have:

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} \mathbf{x}_i.$$

ML-estimates for the Normal Density

- ML estimation for covariance is given by:

$$\hat{\Sigma} = \frac{1}{n} \sum_{i=1}^{n} (\mathbf{x}_i - \hat{\mu})(\mathbf{x}_i - \hat{\mu})^T,$$

Numeric Example

x_1	x_2	x_3	x_4	x_5
13	-115	119	33	-19
29	119	-4	17	73

$$\mu = \begin{bmatrix} 6.2000 \\ 46.8000 \end{bmatrix}$$

$$\Sigma = \begin{bmatrix} 5762.6 & -3212.2 \\ -3212.2 & 1937.0 \end{bmatrix}$$

Example

- Assume we have three classes of objects,
- Feature vector has 4 elements <f1,f2,f3,f4>,
- A-priori probabilities are equal = 1/3
- Pdfs have Normal distribution

Example

- Sample data for the first class:

5.1000	3.5000	1.4000	0.2000
4.9000	3.0000	1.4000	0.2000
4.7000	3.2000	1.3000	0.2000
4.6000	3.1000	1.5000	0.2000
5.0000	3.6000	1.4000	0.2000
5.4000	3.9000	1.7000	0.4000
4.6000	3.4000	1.4000	0.3000
5.0000	3.4000	1.5000	0.2000
4.4000	2.9000	1.4000	0.2000
4.9000	3.1000	1.5000	0.1000

Example

- $\mu = < 4.8600 , \ 3.3100 , 1.4500 , \ 0.2200 >$

- $\sum = \begin{bmatrix} 0.0764 & 0.0634 & 0.0170 & 0.0078 \\ 0.0634 & 0.0849 & 0.0155 & 0.0148 \\ 0.0170 & 0.0155 & 0.0105 & 0.0040 \\ 0.0078 & 0.0148 & 0.0040 & 0.0056 \end{bmatrix}$

Example

- For class two :
- μ=<6.1000 , 2.8700, 4.3700 , 1.3800>
- $\Sigma = \begin{bmatrix} 0.4760 & 0.1740 & 0.2910 & 0.0720 \\ 0.1740 & 0.1041 & 0.1191 & 0.0384 \\ 0.2910 & 0.1191 & 0.2141 & 0.0584 \\ 0.0720 & 0.0384 & 0.0584 & 0.0256 \end{bmatrix}$

Example

- For class three:
- μ=< 6.5700 , 2.9400 , 5.7700 , 2.0400 >
- ∑=[$\begin{matrix} 0.5821 & 0.1252 & 0.4141 & 0.0782 \\ 0.1252 & 0.1024 & 0.1052 & 0.0794 \\ 0.4141 & 0.1052 & 0.3241 & 0.0752 \\ 0.0782 & 0.0794 & 0.0752 & 0.0764 \end{matrix}$]

Classifying a Sample Input

- Input: x = [5.9 4.2 3.0 1.5]^T
- Finding the value of $g_i(\mathbf{x}) = p_{normal}(\mathbf{x}|\mu_i, \Sigma_i)P(\omega_i).$

- The input is classified as a member of class 2

- Repeat this example with some more sample inputs

Questions?