Classification without measurements

We have to classify objects not knowing any objects' features.
The only premise, which allows us to make classification decision, is the probability of occurrence of objects of different classes.
In case of two classes c_1 and c_2 we make a decision, that object belongs to class: c_1 if $P(c_1) > P(c_2)$
c_2 otherwise.
$P(c_1)$ and $P(c_2)$ are called *a priori* probabilities.
Classifier makes always the same decision, dependent only on *a priori* probability.
Classification error: $P(e) = min[P(c_1), P(c_2)\,]$

In case of gender classification of Pattern Recognition lecture participants, current semester's classifier is quite good:
$P(e) = 7/45$ (<20% errors!)

Probability density function

When we classify using one continuous feature, we can treat it as the continuous random variable X (it could be height in our gender classification example).
Distribution of x values is described by probability density function $p(x)$ (pdf).
For each interval $\langle x_1, x_2 \rangle$ we compute the probability of X belonging to this interval from equation:

$$P(c \in C : x_1 \leq X \leq x_2) = \int_{x_1}^{x_2} p(x)dx$$

Distribution function:

$$F(x) = P(c \in C : X < x) = \int_{-\infty}^{x} p(x)dx$$

Of course: $\int_{-\infty}^{+\infty} p(x)dx = 1$

Classification using pdf

For classification purposes more interesting than general pdf $p(x)$ for all objects is the pdf of objects belonging to given class. It is class conditional probability density function $p(x|c_i)$.
We seek probability $P(c_i|x)$ that object belongs to c_i, under condition of measuring feature of value x:

$$P(c_i|x) \sim p(x|c_i)P(c_i)$$

Note, that $P(c_i|x)$ must sum to 1 over all classes.
We introduce normalizing factor: $p(x) = \sum_i p(x|c_i)P(c_i)$

Finally, *a posteriori* probability sought after:

$$P(c_i|x) = \frac{p(x|c_i)P(c_i)}{p(x)} = \frac{p(x|c_i)P(c_i)}{\sum_i p(x|c_i)P(c_i)}$$

(see Bayes theorem)

Less formally: $posterior = \frac{likelihood \times prior}{evidence}$

Bayes Rule(1)

Having *a posteriori* probability we decide that object belongs to:

c_1 if $P(c_1|x) > P(c_2|x)$
c_2 otherwise.

$p(x)$ in preceding equations is only scaling factor, so we can rewrite Bayes rule as:

object belongs to

c_1 if $p(x|c_1)P(c_1) > p(x|c_2)P(c_2)$
c_2 otherwise.

Of course we can develop many other transformations which still will produce optimal Bayes decision…

Loss and Risk

Classification decisions can be associated with different costs (or losses) in classification system.
For each classification decision $\alpha_i, i = 1 \ldots C$ we define loss function $\lambda(\alpha_i|c_j)$. It is the loss connected with taking decision α_i, when the object is really of class c_j. We usually represent this function as a square matrix of decision vs *state of nature*.
The expected loss of decision α_i is:

$$R(\alpha_i|x) = \sum_{j=1}^{C} \lambda(\alpha_i|c_j) P(c_j|x)$$

is called conditional risk.

Bayes Rule (2)

Optimal classifier minimizes the overall risk. We can achieve this taking for every x value decision minimizing conditional risk $R(\alpha(x)|x)$.

Bayes decision rule:
Compute $R(\alpha_i|x)$ for $i = 1 \ldots C$ and select decision α_i, for which $R(\alpha_i|x)$ takes minimum value.

For given loss function, Bayes decision rule achieves the **best** classification results (hence Bayes optimal classifier).

Why not to classify everything optimally?!

Minimum-error-rate classifier

Often used loss function is so called *zero-one* loss function. It takes only two values: for correct decision (0) and error decision (1).

Formally: $\lambda(\alpha_i|c_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$

Conditional risk equals in this case:

$$R(\alpha_i|x) = \sum_{j=1}^{C} \lambda(\alpha_i|c_j) P(c_j|x) = \sum_{i \neq j} P(c_j|x) = 1 - P(c_i|x)$$

Note, that decision condition is in this situation reduced to selecting largest *a posteriori* probability, i.e. we select c_i when $P(c_i|x) > P(c_j|x), i \neq j$ which is exactly Bayes rule (1)

Discriminant Functions

For each class we define function $g_i(x), i = 1 \ldots C$. We decide that object belongs to class c_i if $g_i(x) > g_j(x), i \neq j$

In this representation classifier consists of C modules computing discriminant functions for all classes and maximum selector making final class membership decision.

Discriminant functions divide feature space into *decision regions* $\Re_1 \ldots \Re_C$ of individual classes, separated by *decision boundaries*. Regions $\Re_i$ and $\Re_j$ (for c_i and c_j classes respectively) are separated by boundary given by: $g_i(x) - g_j(x) = 0$.

Discriminant Functions

Bayes classifier: $g_i(x) = -R(\alpha_i|x)$

Minimum-error rate classifier: $g_i(x) = P(c_i|x)$

Note, that the above function assignment is not unique. Every new function $\dot{g}_i(x) = ag_i(x) + b, a > 0$ $\dot{g}_i(x) = ag_i(x) + b, a > 0$ will give the same classification results.

Generally, if we have monotonically increasing function f, than discriminant function $\dot{g}_i(x) = f\big(g_i(x)\big)$ does not change classification.

For example:

$$g_i(x) = P(c_i|x) \equiv p(x|c_i)P(c_i) \equiv ln\big(p(x|c_i)\big) + ln\big(P(c_i)\big)$$

Properties of random variable

Expected value:
For continuous r.v. X with pdf $p(x)$:

$$EX = \int_{-\infty}^{\infty} xp(x)dx$$

For discrete r.v. X with value set W and probabilities $p_i = P(X = x_i)$:

$$EX = \sum_{x_{i \in W}} x_i p_i$$

Variance: $\sigma^2 = E(X - EX)^2$

Moment of *r*-th order relatively to constant *c* is the number $E(X - c)^r$
$c = 0$ normal moments,
$c = EX$ central moments

Univariate normal distribution

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

Example:

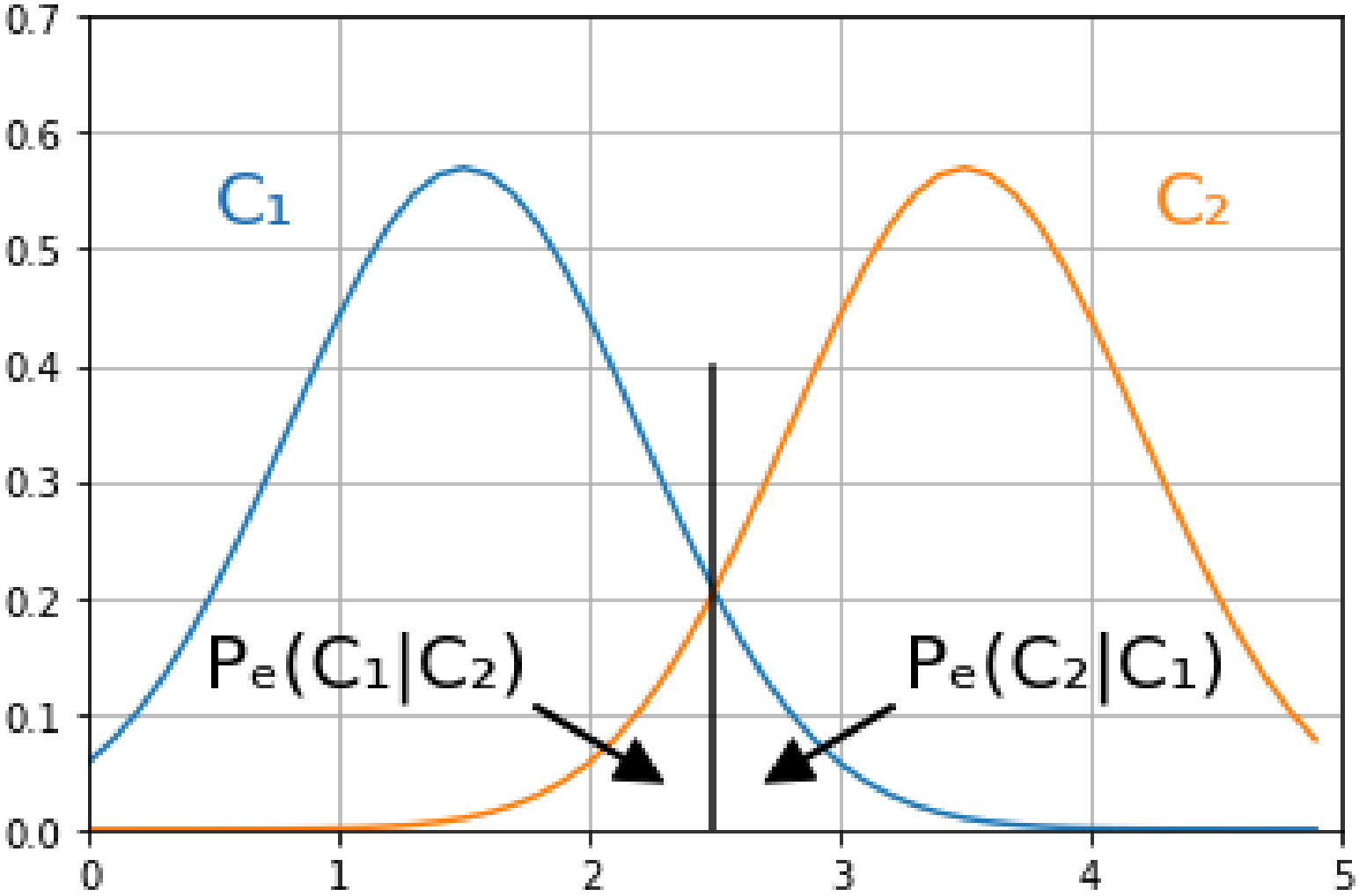
μ_1= 1.5 σ_1= 0.7

μ_2= 3.5 σ_2= 0.7

for $P(c_1) = P(c_2) = 0.5$ $\quad x_{opt} = \frac{\mu_1+\mu_2}{2} = 2.5$

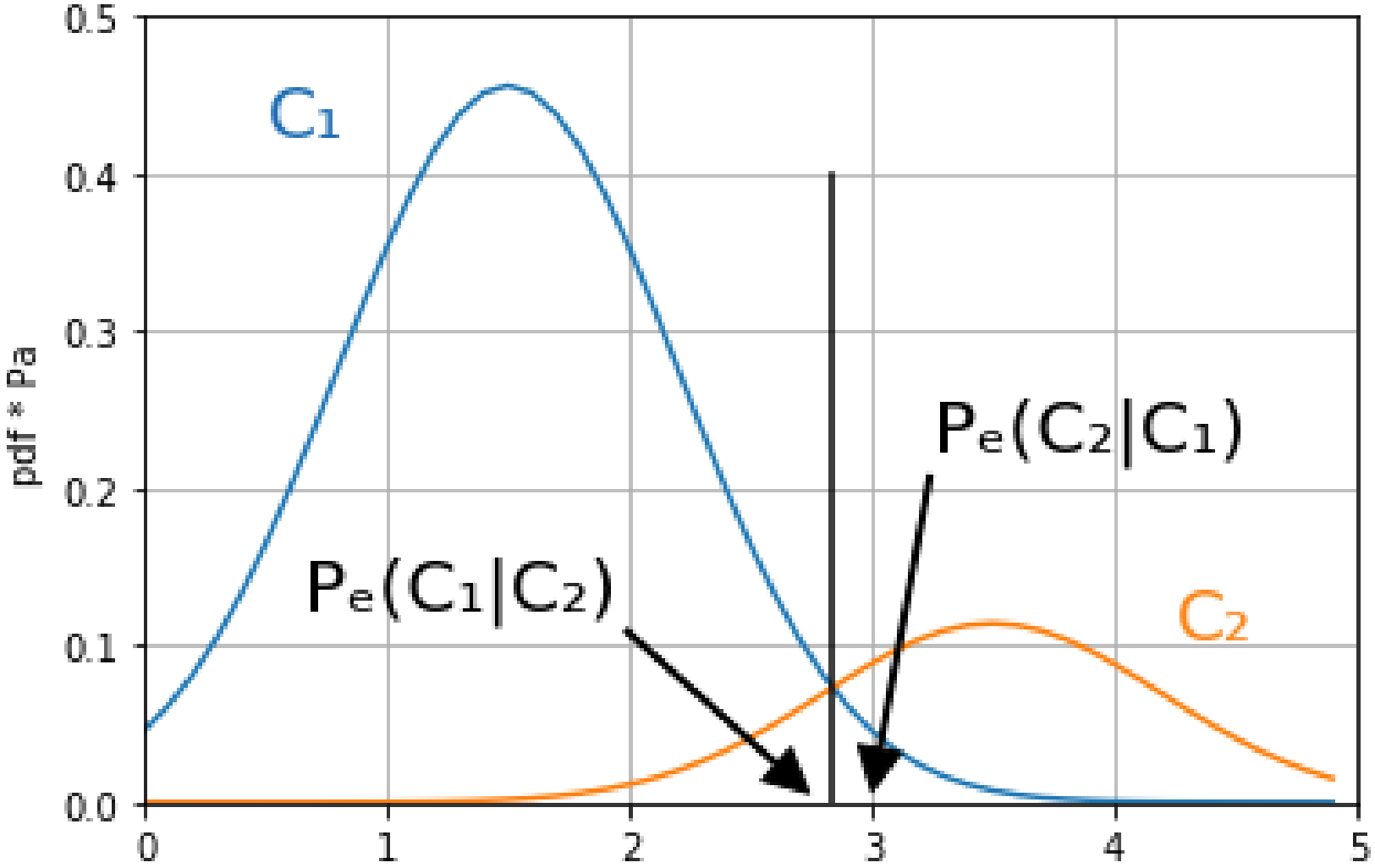
for $P(c_1) = 0.8$ $P(c_2) = 0.2$

$$x_{opt} = \frac{\mu_1 + \mu_2}{2} - \frac{\sigma^2 ln\left({P_2}/{P_1}\right)}{\mu_2 - \mu_1} = 2.5 + 0.34 = 2.84$$



$$P_e(c_1|c_2) = F_{c2}(x_{opt}) = 0.0766 \quad P_e(c_2|c_1) = 1 - F_{c1}(x_{opt}) = 0.0766$$

Classifier error probability: $0.5 * P_e(c_1|c_2) + 0.5 * P_e(c_2|c_1) = 0.0766$



$P_e(c_1|c_2) = F_{c2}(x_{opt}) = 0.1729 \quad P_e(c_2|c_1) = 1 - F_{c1}(x_{opt}) = 0.0278$

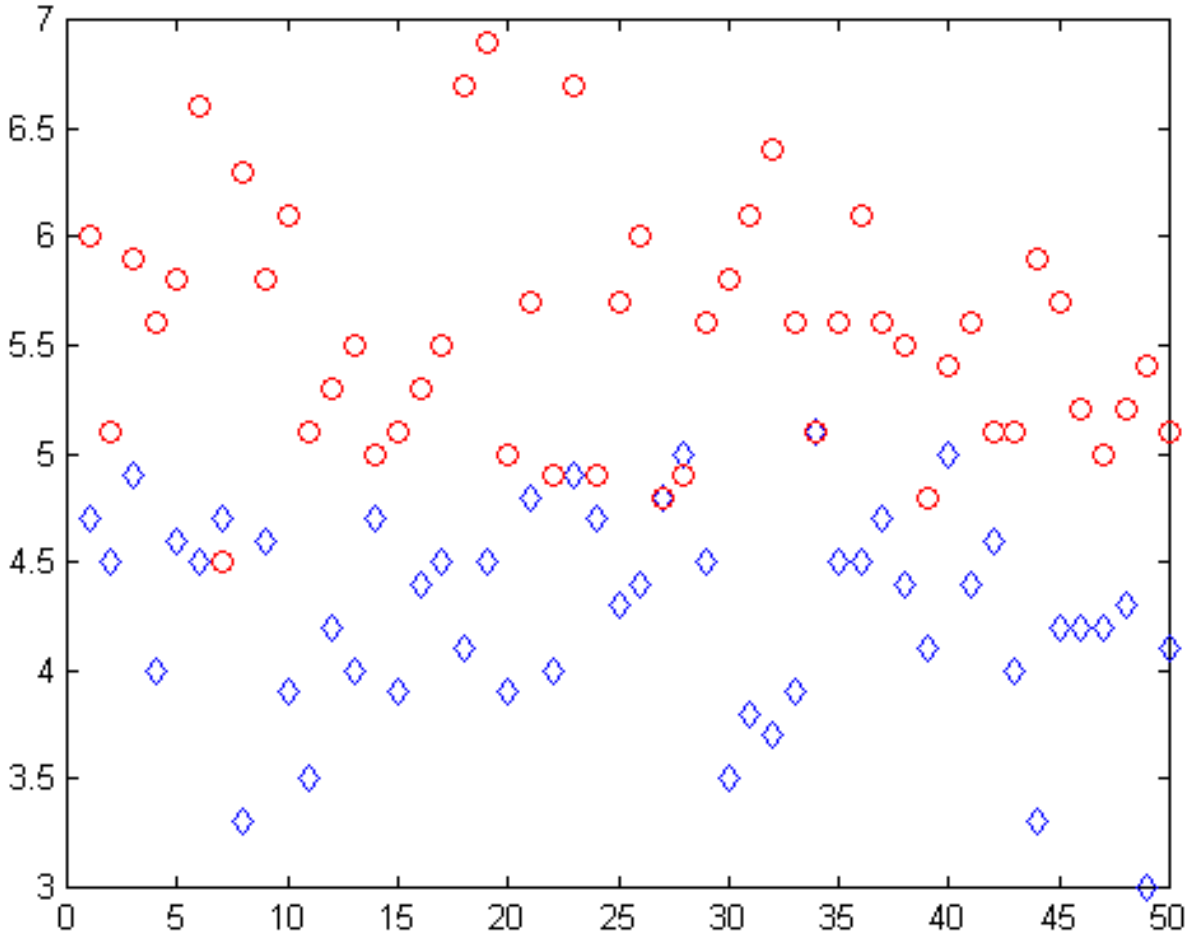
Classifier error probability: $0.2 * P_e(c_1|c_2) + 0.8 * P_e(c_2|c_1) = 0.0568$

Flower classification

Iris virginica

Iris versicolor

Feature: petal length



Values:

$\mu_1 = 5.55$

$\sigma_1 = 0.47$

$\mu_2 = 4.29$

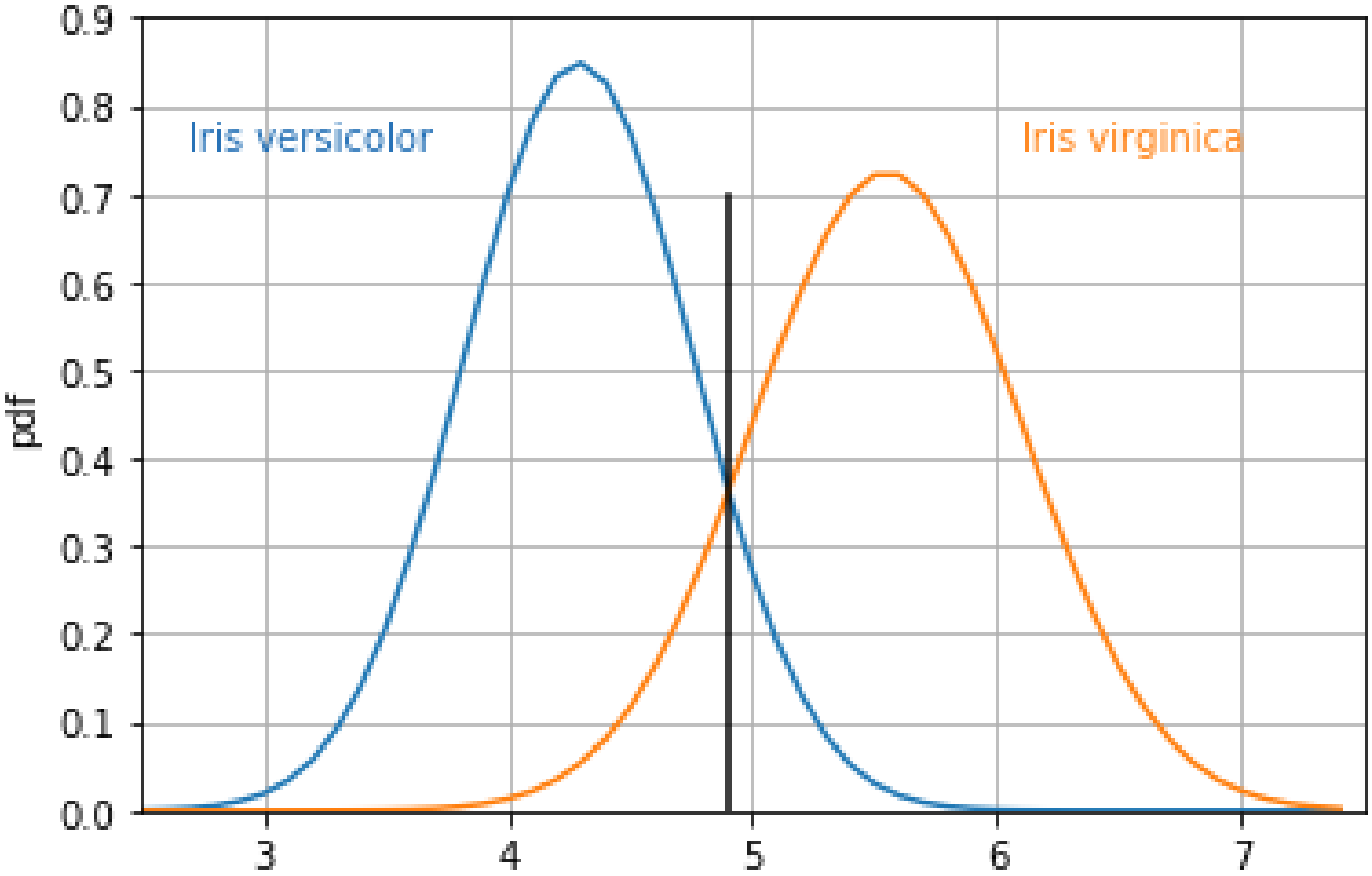
$\sigma_2 = 0.55$

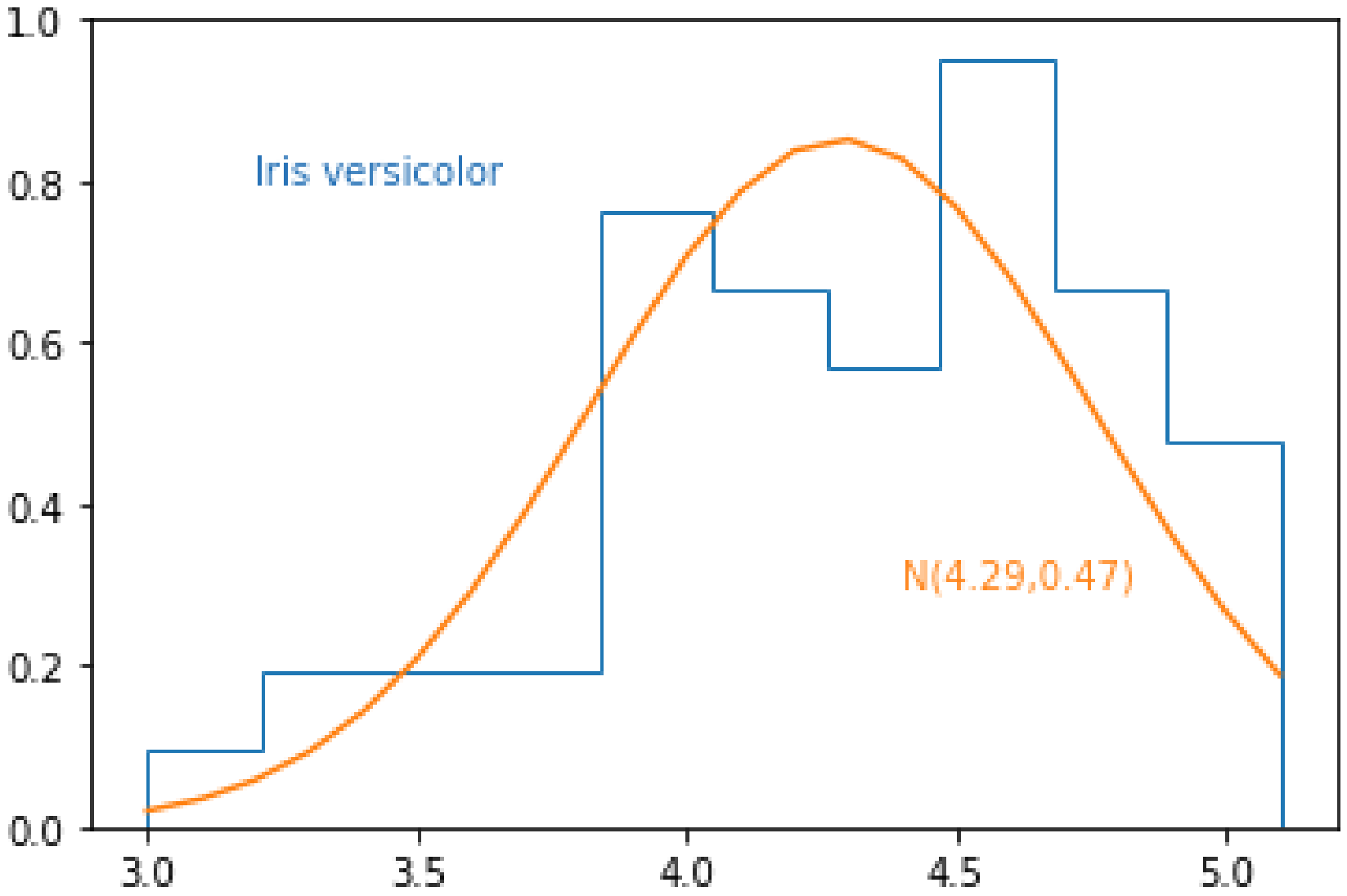
$x_{opt} \approx 4.91$

$P_{err} \approx 0.11$

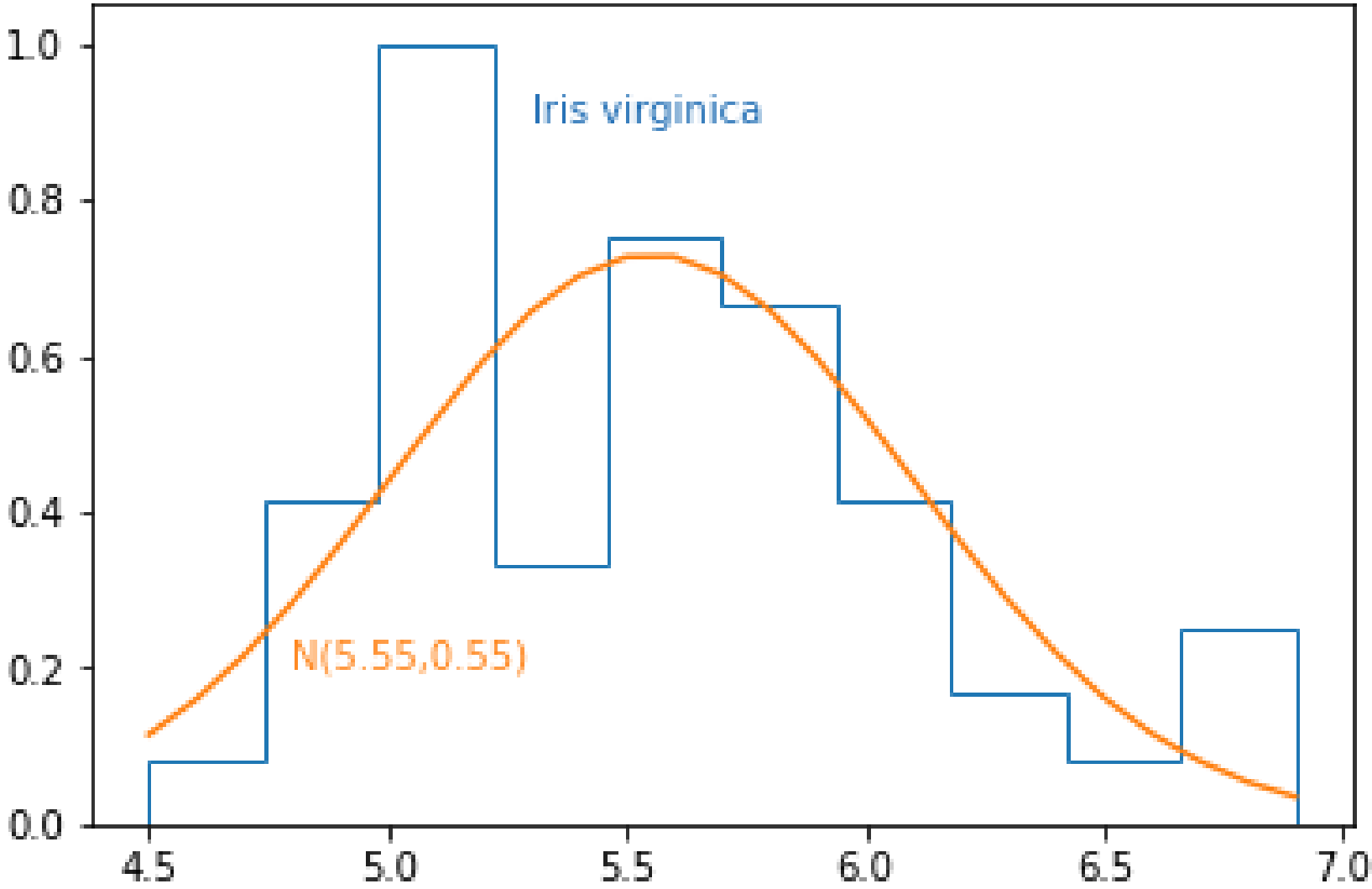
Measured error:

$P_{err} \approx 0.09$





P_{err}=0.5*0.0961=0.048 Measured error = 0.03



P_{err}=0.5*0.1197=0.0599 Measured error = 0.06

Multivariate normal distribution

$$p(x) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left[-\frac{(x-\mu)^T \Sigma^{-1} (x-\mu)}{2}\right]$$

Feature vector: $\mathbf{x}^T = (x_1, x_2, \cdots, x_d)$

Expected value: $E\mathbf{X} = \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{x}_i$

Variance $\rightarrow$ covariance matrix

$$\boldsymbol{\Sigma} = E[(\mathbf{x}-\boldsymbol{\mu})(\mathbf{x}-\boldsymbol{\mu})^T] = \{\sigma_{ij}\}$$

$$\sigma_{ij} = E\left[(x_i - \mu_i)(x_j - \mu_j)^T\right]$$

$\sigma_{ii} = \sigma_i^2$ variance of feature *i*

σ_{ij} covariance between features *i* and *j* ($\sigma_{ij} = \sigma_{ji}$)

Covariance properties

Features are uncorrelated when

$$E\left[x_i\, x_j\right] = E\left[x_i\right] E\left[x_j\right]$$

Features are orthogonal when

$$E\left[x_i\, x_j\right] = 0$$

Features are independent when

$$p(x_i, x_j) = p(x_i)\, p(x_j)$$

Population - sample

Expected value estimate: $\bar{x} = \frac{1}{N} \sum_{i=1}^{N} x_i$

Sample variance: $\sigma^2 = \frac{1}{N-1} \sum_{i=1}^{N} (x_i - \bar{x})^2$

Sample covariance matrix:

$$\sigma_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} (x_{ik} - \bar{x}_i)(x_{jk} - \bar{x}_j)$$

$$\sigma_{ij} = \frac{1}{N-1} \sum_{k=1}^{N} d_{ik} d_{jk}, \quad d_{ik} = x_{ik} - \bar{x}_i$$

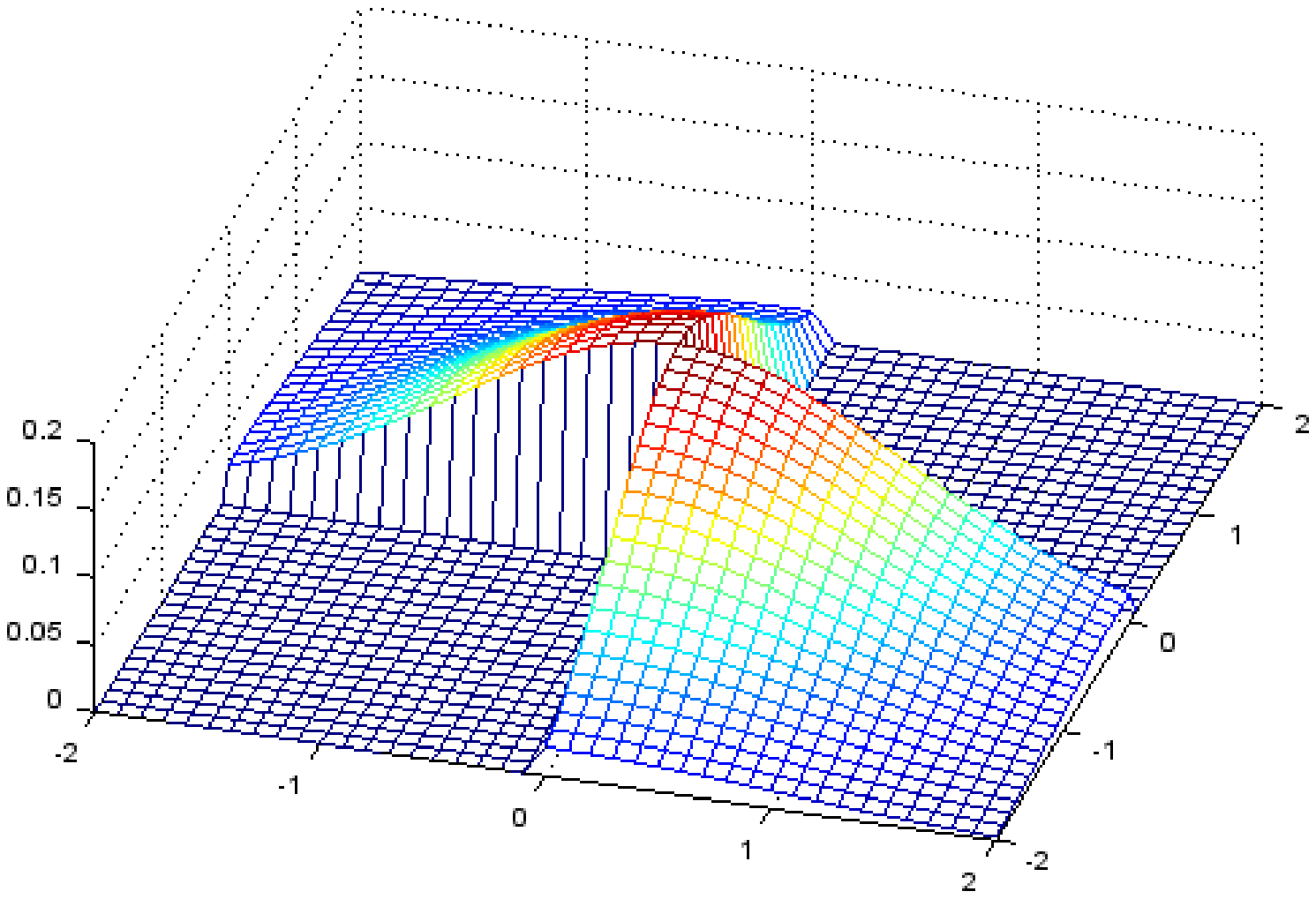
Multivariate normal distributions

Marginal distribution:
If X is multivariate normal distribution, then each marginal distribution is also normal distribution.

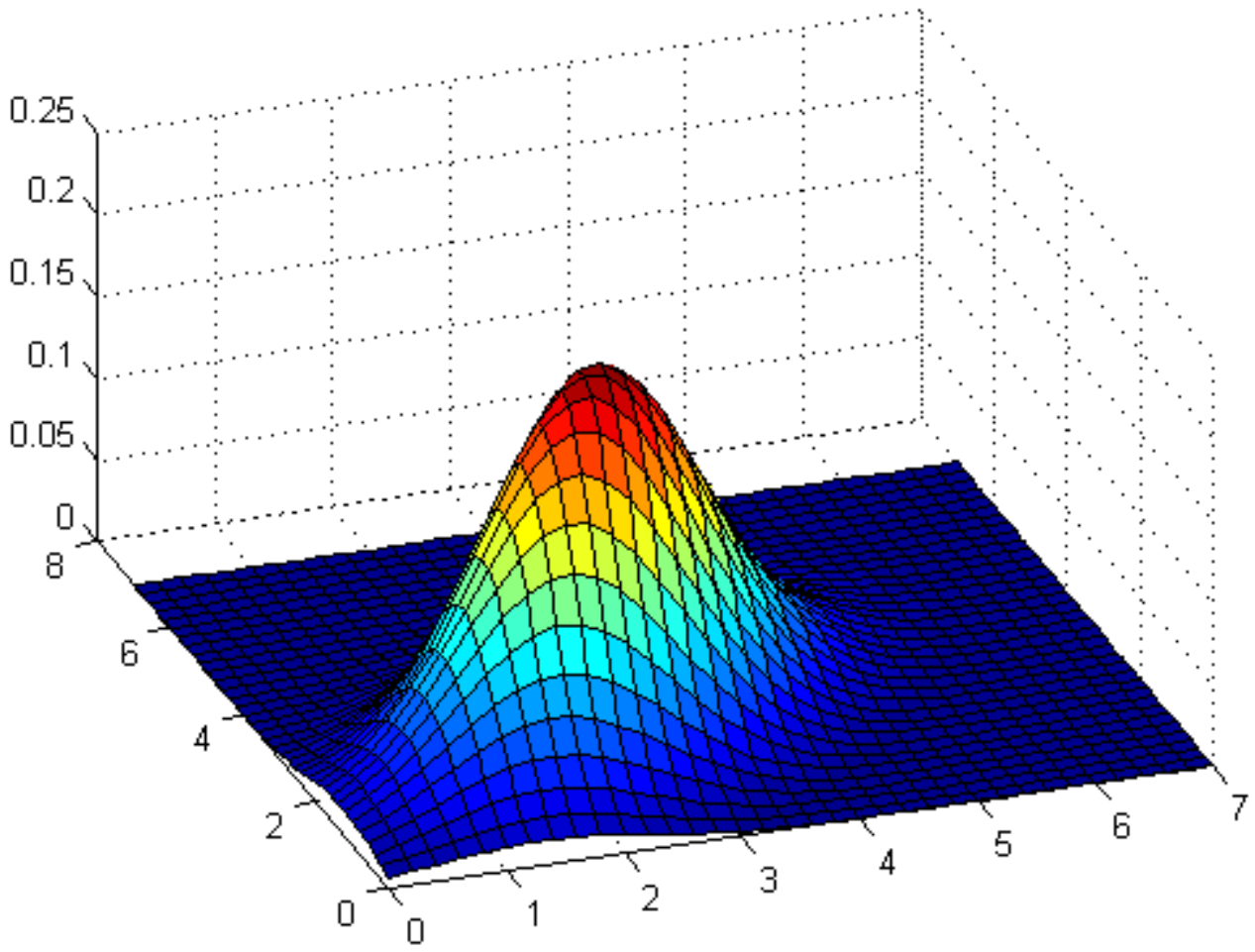
Conditional distribution:
If X is multivariate normal distribution, then each conditional distribution is also normal distribution.

Combination distribution:
If X is multivariate normal distribution, then each combination of coordinates distribution is also normal distribution.

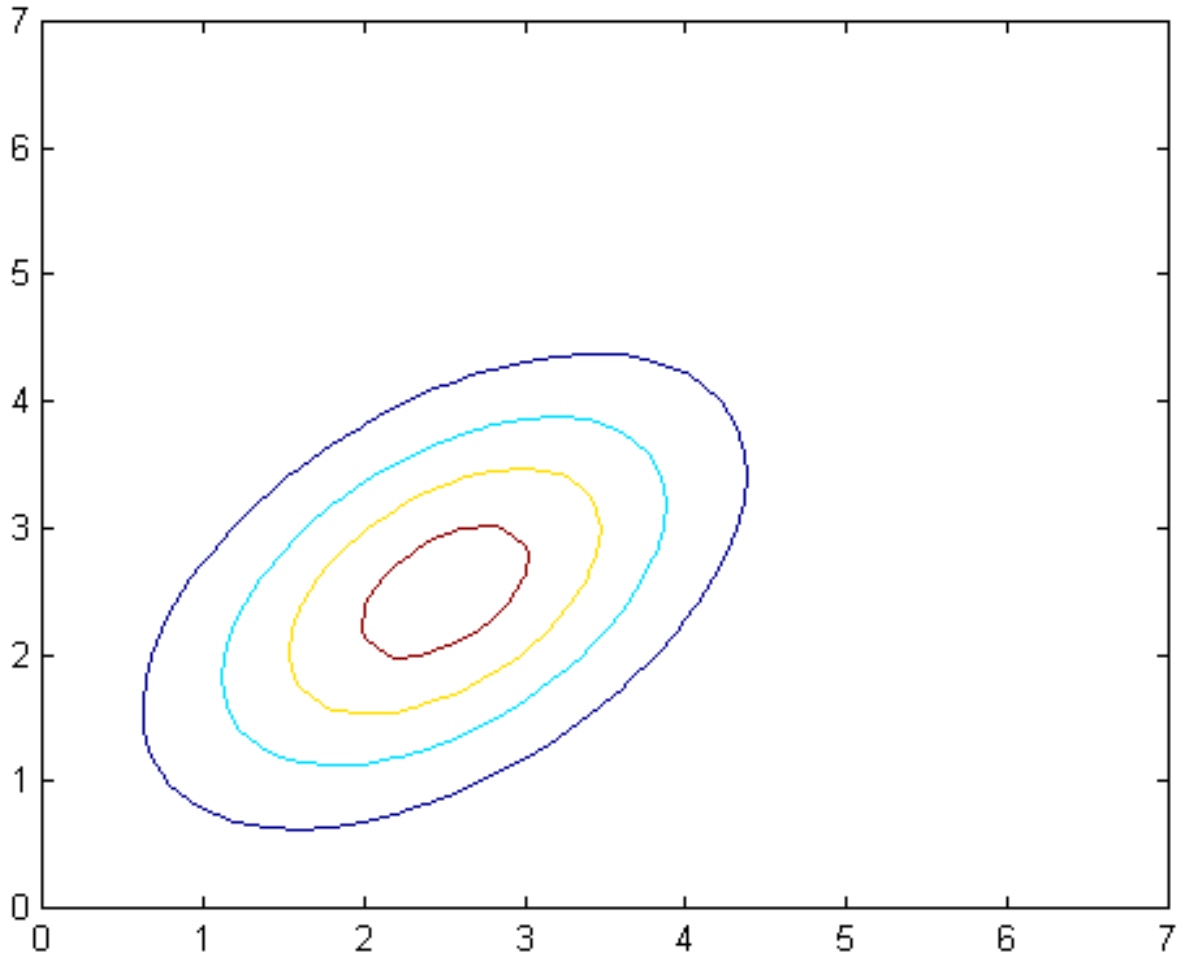
The bad news is that it does not work the other way: i.e. having normal distributions of individual features does not mean that the multivariate distribution is normal ☹



Marginal distributions are normal in this case, two-dimensional distribution evidently not!

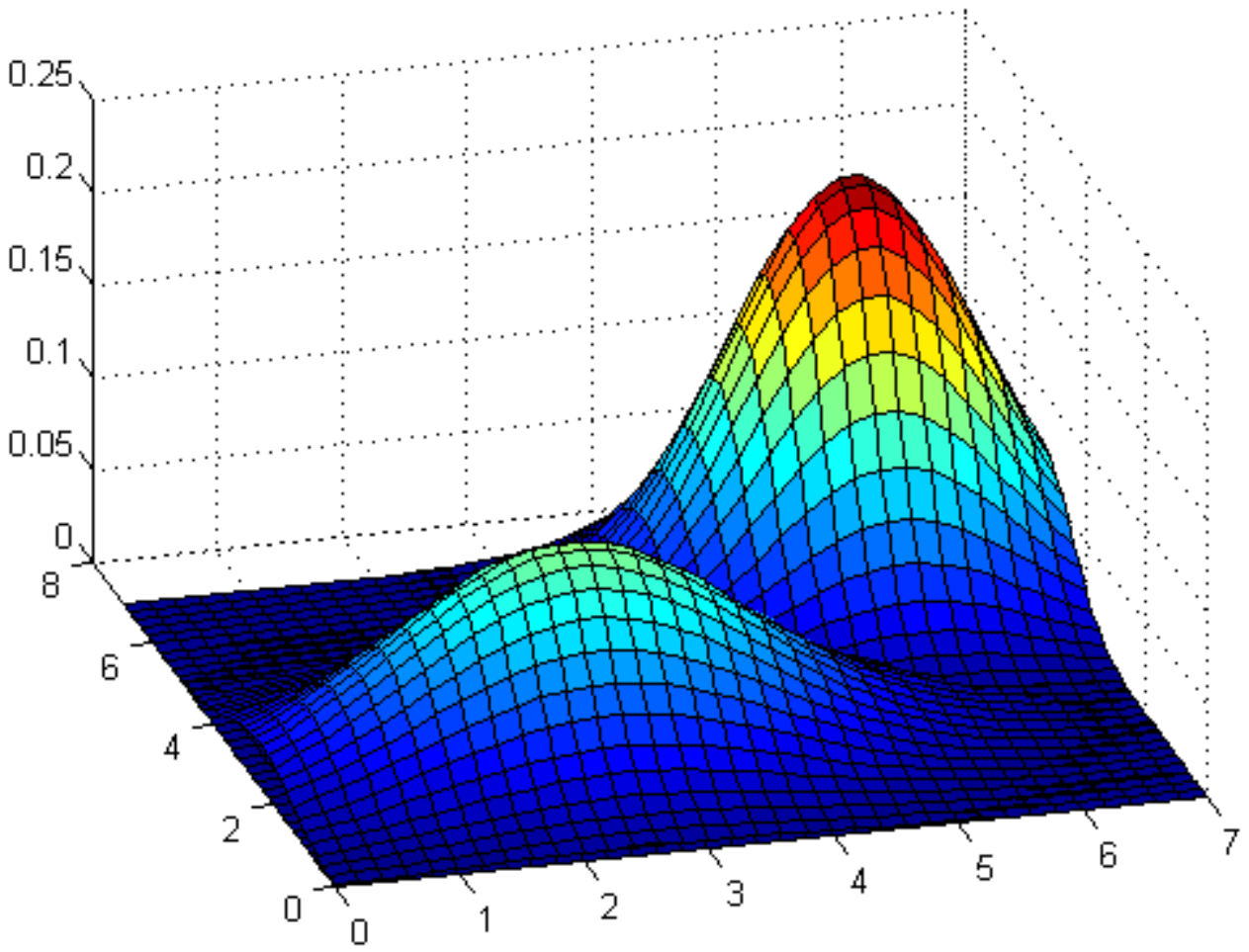


Bivariate normal distribution

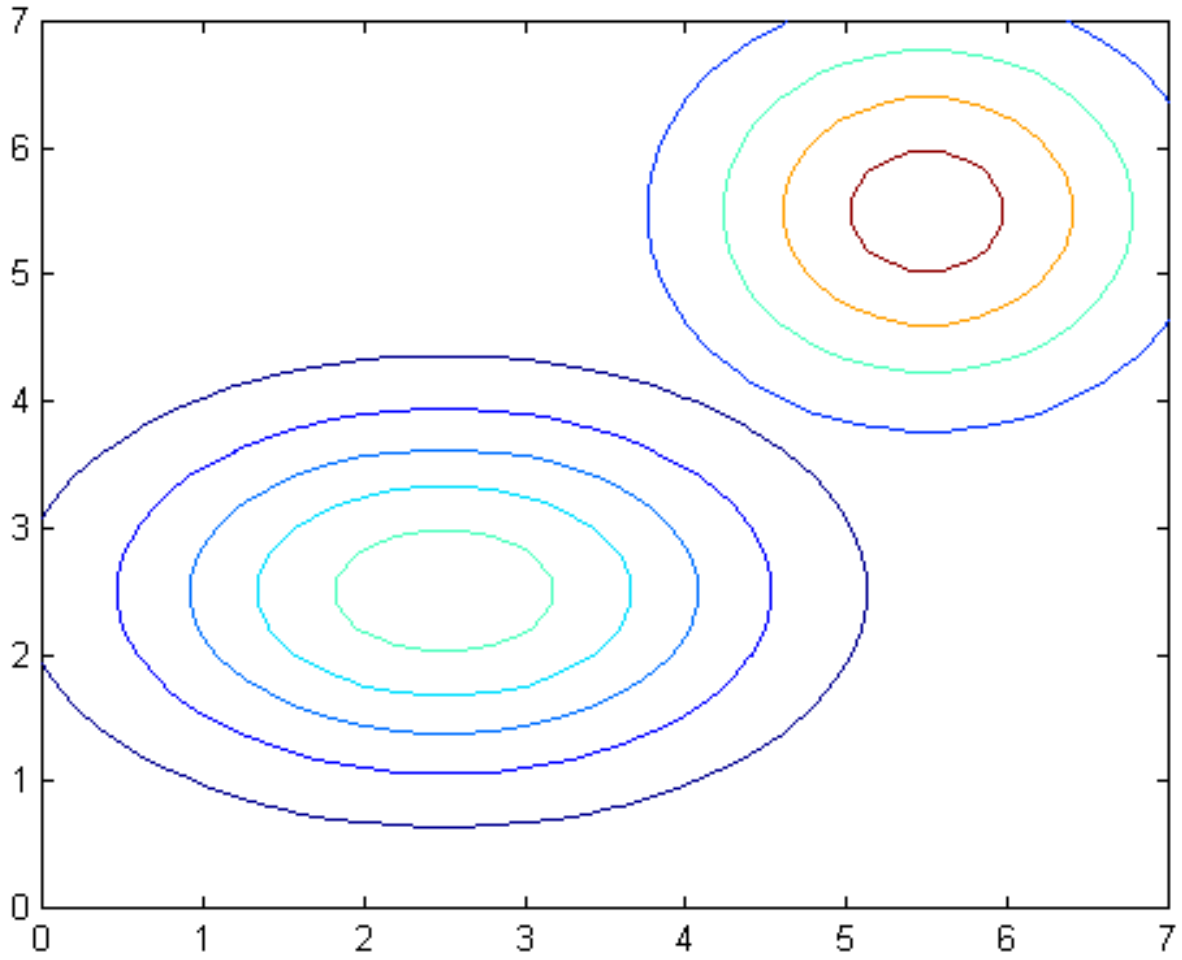


... and its contour plot

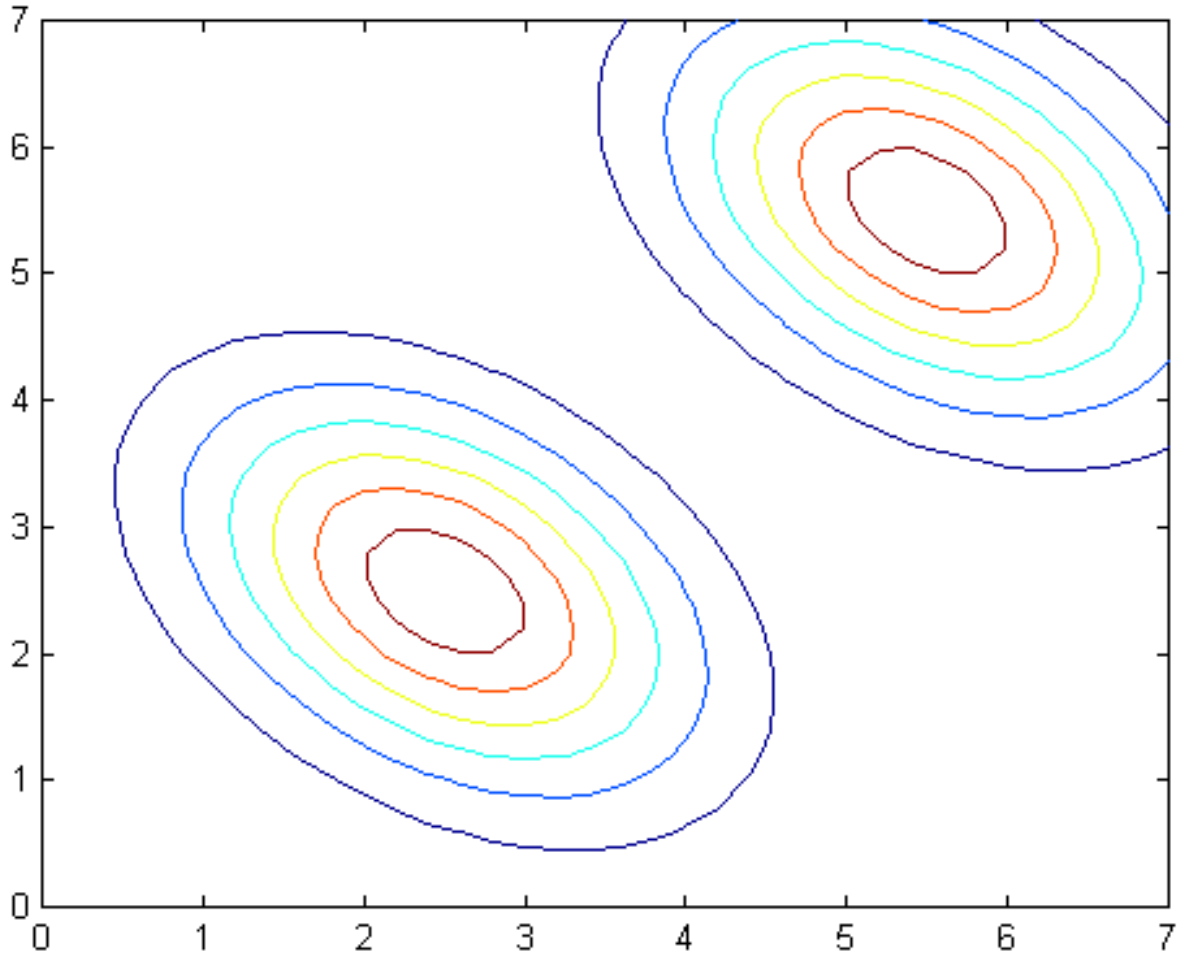
Mahalanobis distance r: $r^2 = (\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$



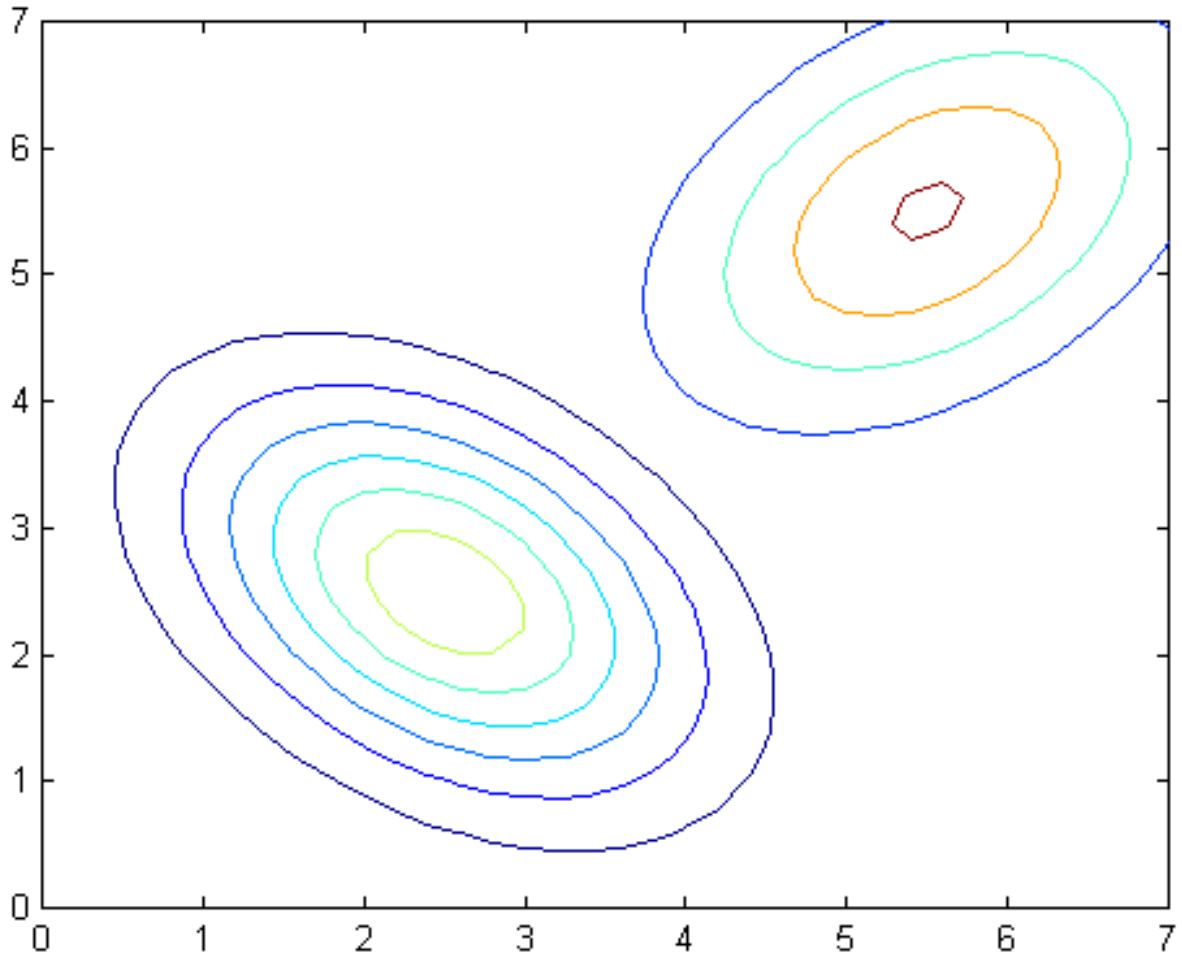
Two classes with normal distributions



... and the same on contour plot.



Equal covariance matrices



Unequal covariance matrices

Discriminant functions for normal distribution

We consider minimum error rate classifier (i.e. zero-one loss function).

Discriminant functions: $g_i(\mathbf{x}) = \ln p(\mathbf{x} \mid \omega_i) + \ln P(\omega_i)$

We substitute: $p(\mathbf{x} \mid \omega_i) \sim N(\boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$

$$g_i(\mathbf{x}) = -\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu}_i)^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu}_i) - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\boldsymbol{\Sigma}| + \ln P(\omega_i)$$

$\boldsymbol{\Sigma}_i = \sigma^2 \mathbf{I}$ **Linear discriminant function**

$\boldsymbol{\Sigma}_i = \boldsymbol{\Sigma}$ **Linear discriminant function**

$\boldsymbol{\Sigma}_i$ Second order surface

Goodness-of-fit test χ^2

Let's assume that we want to check, if the data at our disposal, is consistent with normal distribution with given parameters. Of course, parameters of this distribution: mean and standard deviation we compute from our data.

The basic procedure for this test consists in separating the data into *k* bins. For each bin we'll have a N_i number representing the number of samples in the i^{th} bin. Knowing the theoretical distribution and its parameters we can compute n_i numbers, representing number of samples from the theoretical distribution belonging to the i^{th} bin.

We can now compute χ^2 statistics:

$$\chi^2 = \sum_{i=1}^{k} \frac{\left(N_i - n_i\right)^2}{n_i}$$

Goodness-of-fit test χ^2

The value of statistics is used to disprove the null hypothesis, that our data is consistent with given theoretical distribution. Generally, the greater χ^2 statistics value, the more probable that null hypothesis is false.
Critical values are usually read from tables, but important is the number of **degrees of freedom** in the test. Initially, the number of degrees of freedom is the same as the count of bins, but we have subtract the number of distribution parameters computed from the data. In our example there are two parameters: mean and standard deviation. Additionally, theoretical probabilities have to be multiplied by the number of samples in our data set, which counts for next parameter.

χ^2 test for iris data

4 bins located around the mean value will be used:

1: $x<\mu-\sigma/2$ 2: $\mu-\sigma/2 \geq x < \mu$ 3: $\mu \geq x < \mu+\sigma/2$ 4: $x \geq \mu+\sigma/2$

Selection of ranges is not insignificant. It is recommended that for each bin the condition $n_i \geq 5$ should be fulfilled.

We have following numbers in the bins:

Bin	N_i	p_i	n_i	χ^2_i
1	15	0.3048	15.2	0.0026
2	7	0.1952	9.8	0.8000
3	13	0.1952	9.8	1.0449
4	15	0.3048	15.2	0.0026

Total χ^2=1.8502
Number of degrees of freedom: 4 (bins) – 3 (parameters) = 1
From χ^2 distribution tables we get critical value for given significance level (say α=0.05): 3.841 (>1.85).
It cannot be disproved (with significance 0.05) that the data set comes from a normal distribution.

Problems with Bayes rule

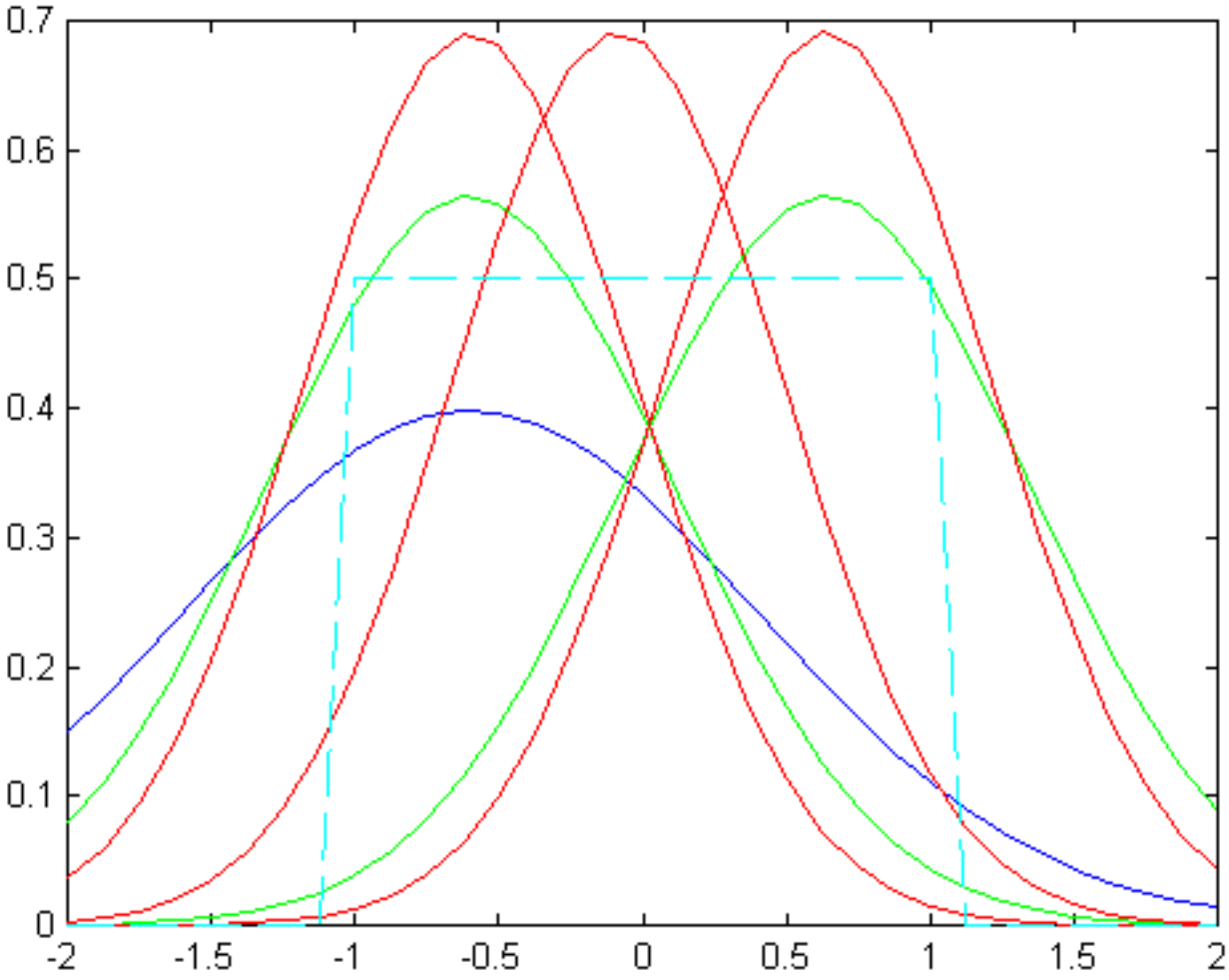
- Decision criterion can be analytically complicated even for normal distribution.
- Even worse situation is when distribution is not normal.
- It can be hard to establish feature distribution.
- Classes with small *a priori* probability have weak influence on decision criterion.
- Where to find *a priori* probability?
- Completeness ?! - reject decision.

Estimating pdf with Parzen window

We can estimate unknown probability density with so called Parzen window. The general concept of this method consists in “building” of the unknown density from the partial densities introduced by training samples. Partial densities are delivered by a function, called window function $\varphi(u)$. There are no special constraints on the window function, but it should be valid density function (i.e. non-negative and integral over the whole domain should be equal to 1).
We can use for example Gaussian-like window function:

$\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-u^2/2}$ Additionally we’ll introduce h_1 parameter, called window’s width, scaled by the number of samples: $h_n = \frac{h_1}{\sqrt{n}}$

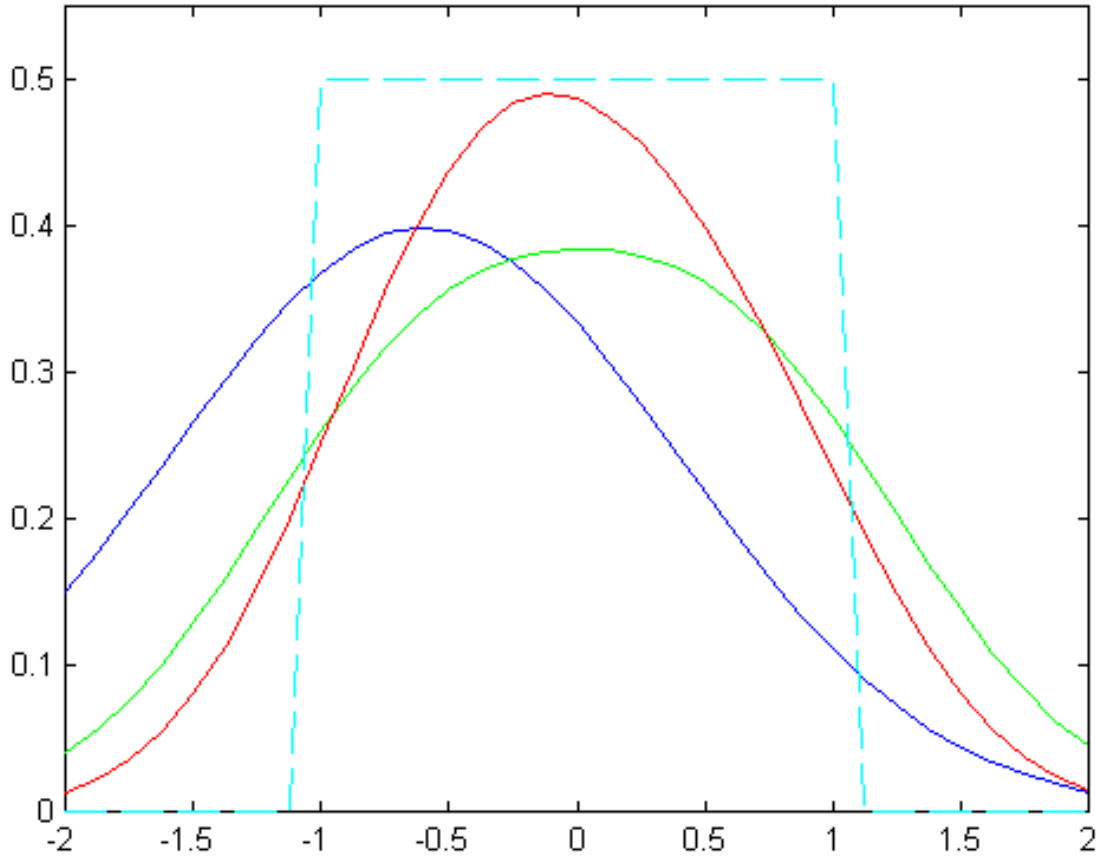
Estimating pdf with Parzen window



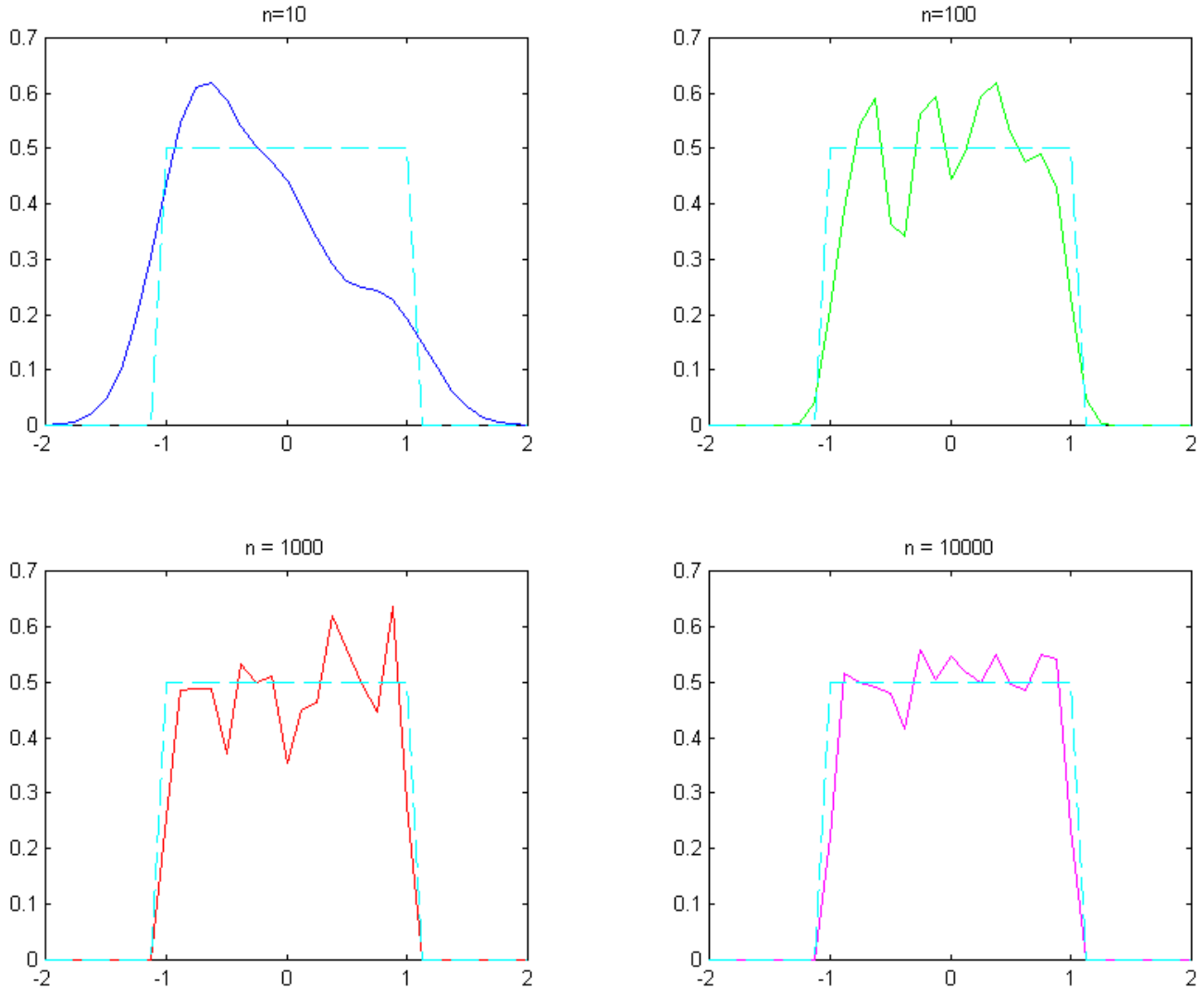
Window function for 1, 2 and 3 samples of the training set.

Estimating pdf with Parzen window

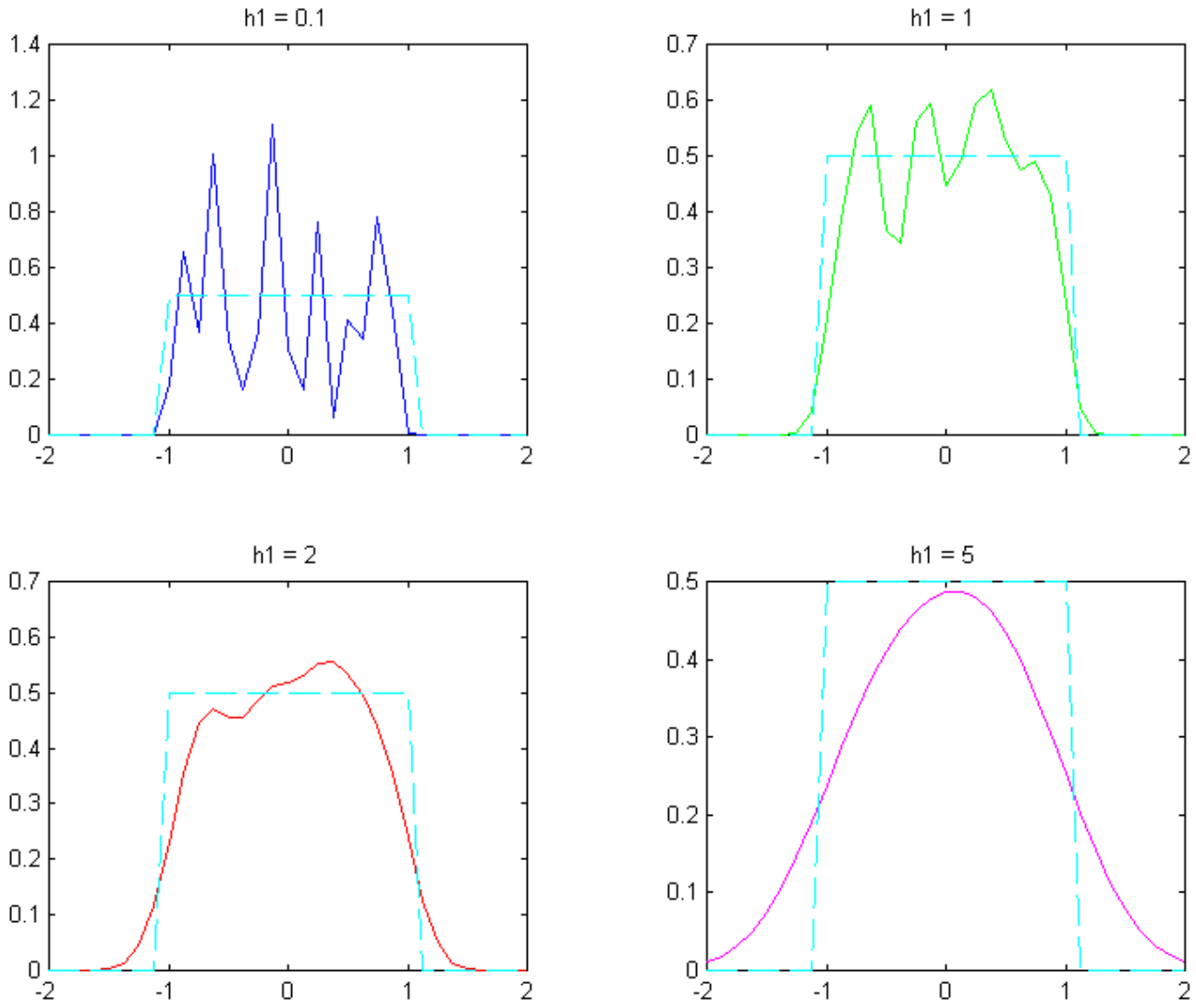
Finally, pdf estimate is: $p_n(x) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{h_n}\varphi\left(\frac{x - x_i}{h_n}\right)$



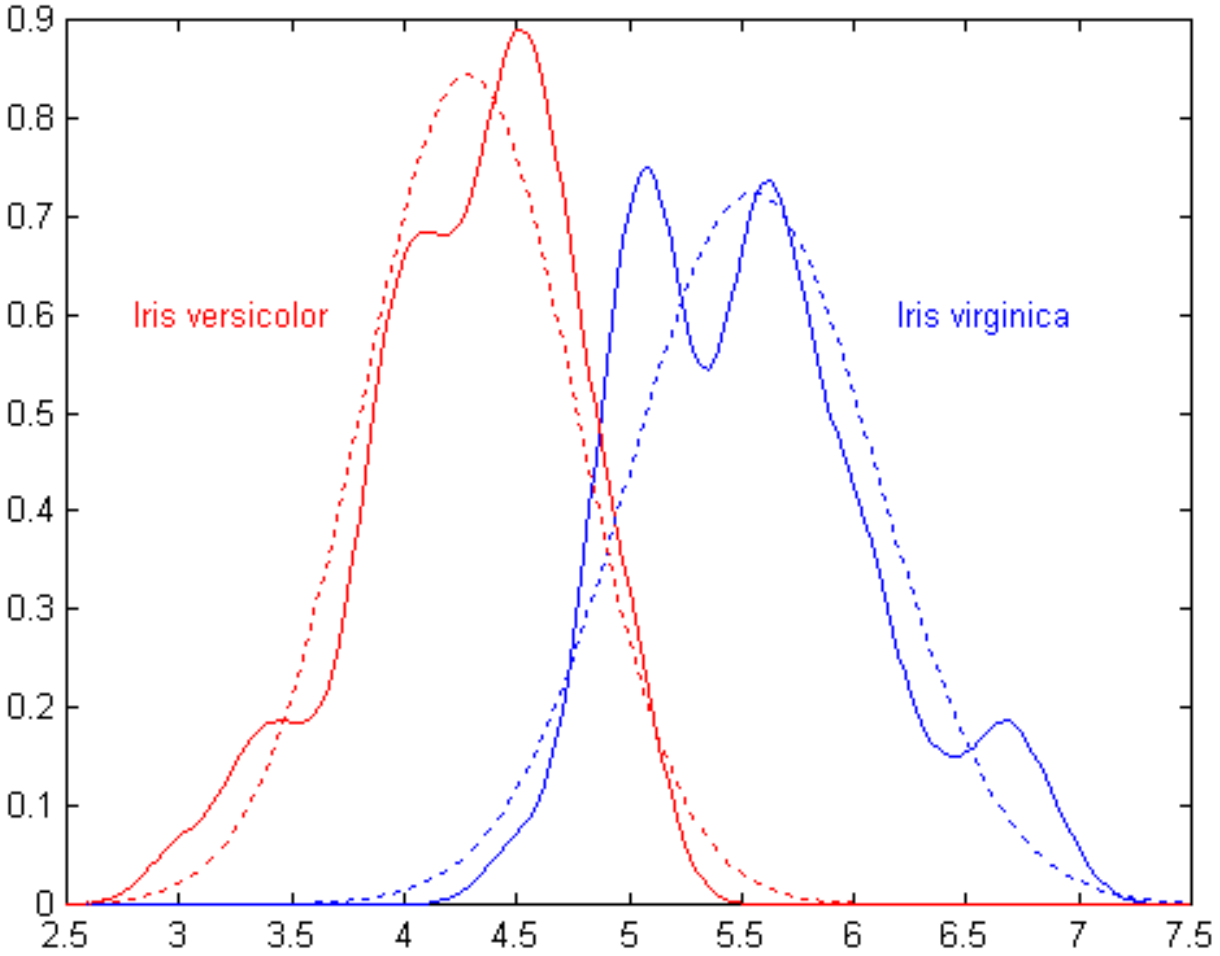
Parzen window with different number of samples



Different Parzen window *widths*



Parzen window on Iris data



Note a very small difference of the decision boundary in both cases.