Test Exam Solution: Pattern Recognition

Prof. Dr. Hauke Schramm

Exercise 1. (Questions on the lecture)

a. Which of the presented decision rules is equivalent (i.e. same classification result) to the shown MAP-decision rule? It is possible to mark more than one entry.

$$\hat{k} = \arg\max_k p(k \mid x)$$

Answer	Mark your choice here
$\hat{k} = \arg\max_k p(k)$	
$\hat{k} = \arg\max_k p(x \mid k) \cdot p(k)$	X
$\hat{k} = \arg\max_k p(x \mid k)$	
$\hat{k} = \arg\max_k p(x, k)$	X
$\hat{k} = \arg\max_k p(x \mid k) \cdot p(k) / p(x)$	X
$\hat{k} = \arg\max_k \{p(x \mid k) \cdot p(k) + C\}$ C: constant	X

(1.5 points for each correct cross, minus 1.5 points for each mistake)

b. Explain, how Bayes rule (Attention: ***not*** Bayes decision rule!) is used in classification and why it is important.

Answer:

Bayes rule:

$$P(x \mid y) = \frac{P(y|x) \cdot P(x)}{P(y)}$$

Bayes rule allows to invert statistical connections between probability functions.

Usage and importance: In many situations we may easily compute p(x|w), P(w), p(x) but need the inverted function P(w|x) for Bayes decision rule.

c. Give one example for a parametric and for a non-parametric classification technique and discuss the differences between them. What are the advantages / disadvantages of non-parametric techniques?

Answer:

Parametric: Maximum-likelihood parameter estimation of a Gaussian distribution

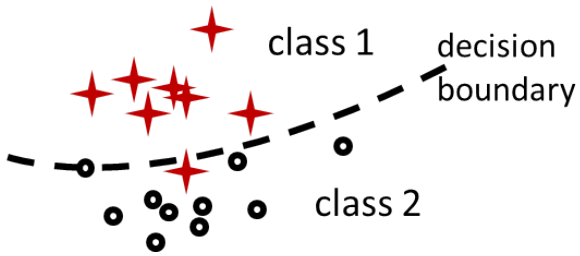
Non-parametric:

Example: Parzen windows, nearest neighbor classifier

Advantage: Exact estimation of posterior probability possible if enough training data available

Disadvantage: Requires much training data and computational resources.

d. Could it happen, using the ***Parzen Window*** approach, that single ***training data*** samples of a class lie on the wrong side of the decision boundary (see illustration)? Explain in detail! (No points without correct and exact justification.)



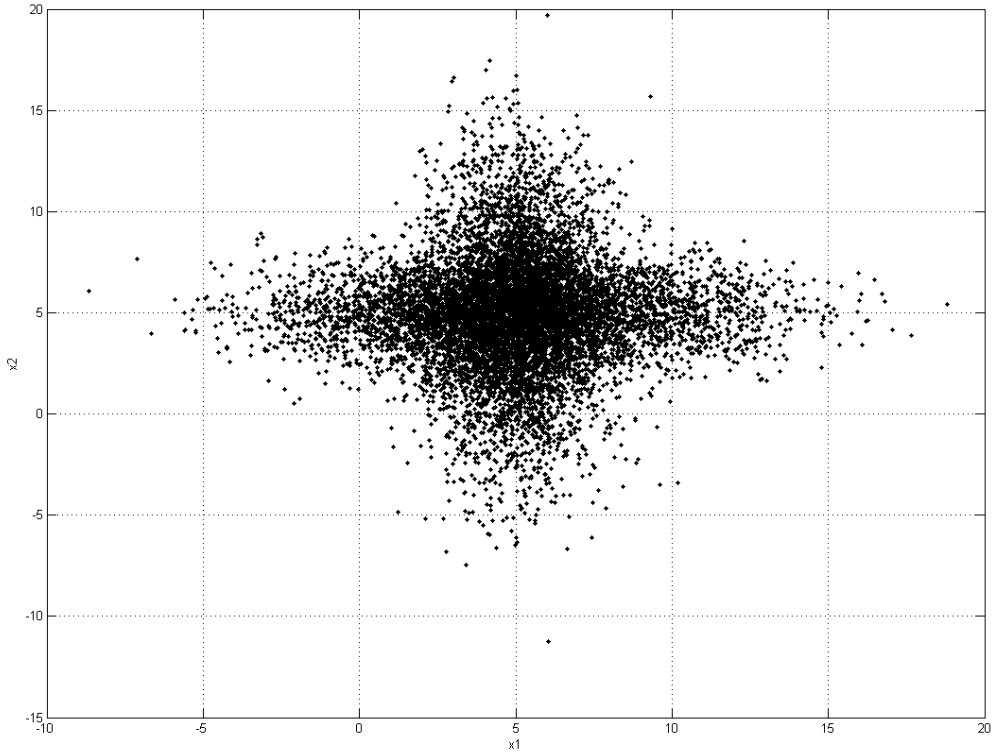
Answer:

Yes, that is possible since the final likelihood is the result of summing all samples' Parzen Windows (cf. equation). If there are many samples of the other class around (see example) the class-conditional probability may be larger even at the sample's position.

$$\hat{p}(x \mid \omega) = \frac{1}{n} \sum_{i=1}^{n} \varphi(x - x_i)$$

Exercise 3. (Joint distribution)

A measurement of 10000 data samples of two sensor signals x1 and x2 shows the distribution of the figure below:



a. Are x1 and x2 statistically independent? Explain!

b. Estimate the form of the marginal distributions from the given figure and illustrate them. Hint: The maximum probability of the marginals amounts to about 0.05.

c. How would you ***calculate*** the marginals from the given data points? Write a short Matlab script that performs this task. Take into consideration that the marginals are probability distributions! Could you reconstruct the joint probability from the marginals? If yes: How? If no: Why not?

d. Assume, you want to model the point cloud with a

   i. multivariate Gaussian distribution with ***diagonal*** covariance matrix
   ii. multivariate Gaussian distribution with ***full*** covariance matrix

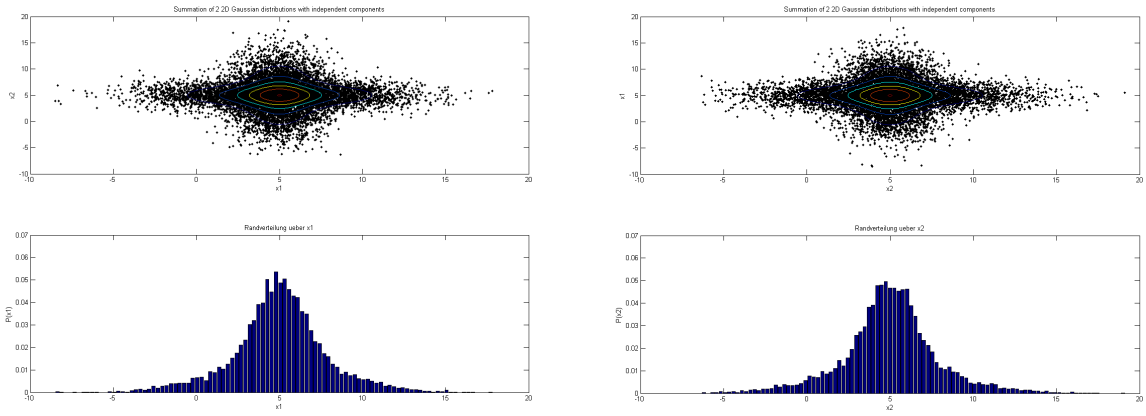
   Describe how you would do that and estimate and illustrate the resulting joint probability for both cases (i. and ii.). Use contour lines for illustration. Is one of the two models better suited to represent the data? Explain!

e. Do you achieve a good representation of the data using the models applied in d.? If not: How could you improve the modeling? Describe your approach in detail and illustrate the expected probability contour image of the improved model.

Answer a.:

The variables are statistically independent since different values of x1 do not influence the distribution over x2 and vice versa.

Answer b.:



Answer c.:

Generation of marginals:

- Calculate histograms over x1 and x2
- Divide the counts by the total number of points → probability distribution

Matlab script:

```
[N, X] = hist(x(:,2),100);

bar(X, N/(size(x,1)));
```

The joint probability can be constructed from the marginals since the random variables are statistically independent:

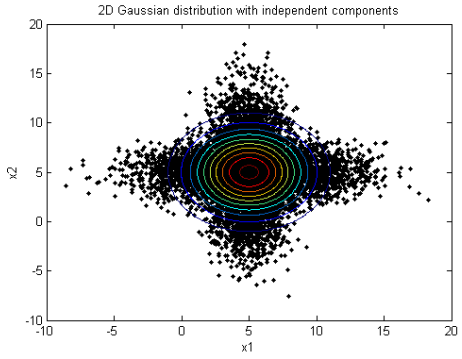
1. Organize the distribution P(x1) as column vector
2. Organize P(x2) as row vector
3. Multiply both → Matrix with probabilities for each point (x1, x2)

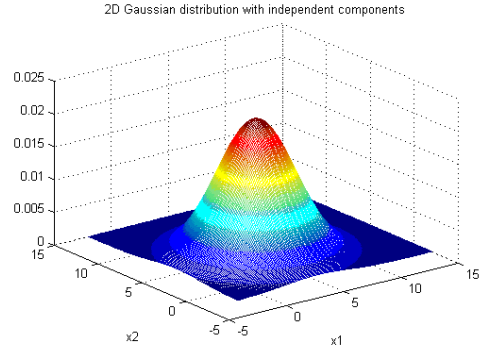
Answer d.:

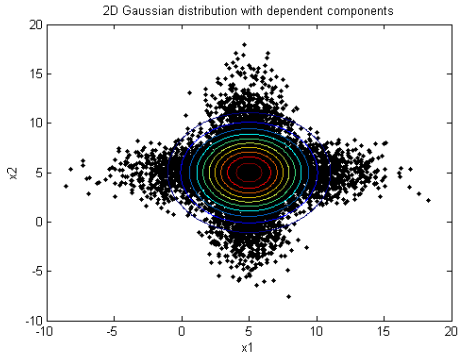
i. Approach:
Estimate variance of x1 and x2 and insert it into the diagonal covariance matrix. Estimate mean vector and insert parameters into formula of the Gaussian distribution. Alternative: Multiply the marginals.

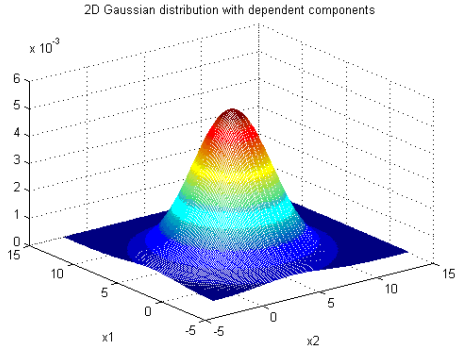
ii. Same approach but estimation of a full covariance matrix (Matlab: cov())

Illustration of the distributions:









Both distributions are badly suited to model the data. The full covariance approach would be better in case of statistically dependent data. This is, however, not the case here. Thus the "full" covariance matrix has only a diagonal structure.

Answer e.:

Approach: Estimate parameters of two Gaussian distributions and combine both distributions but summation and division by 2.

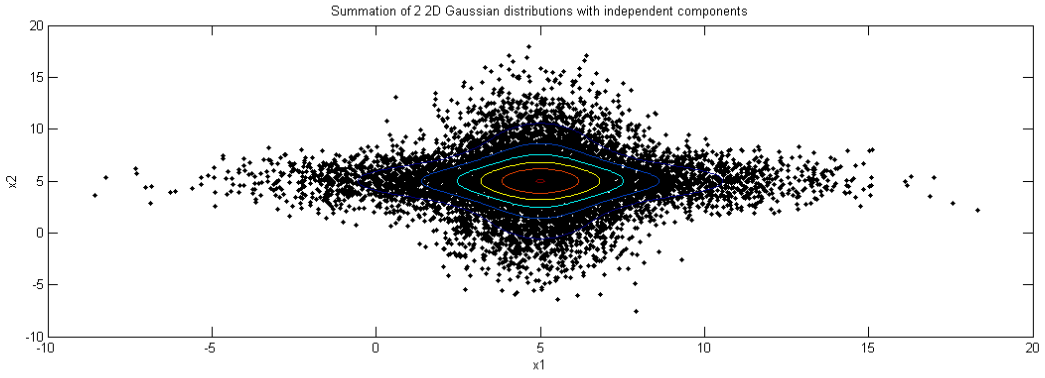
Matlab code for this example: see next page

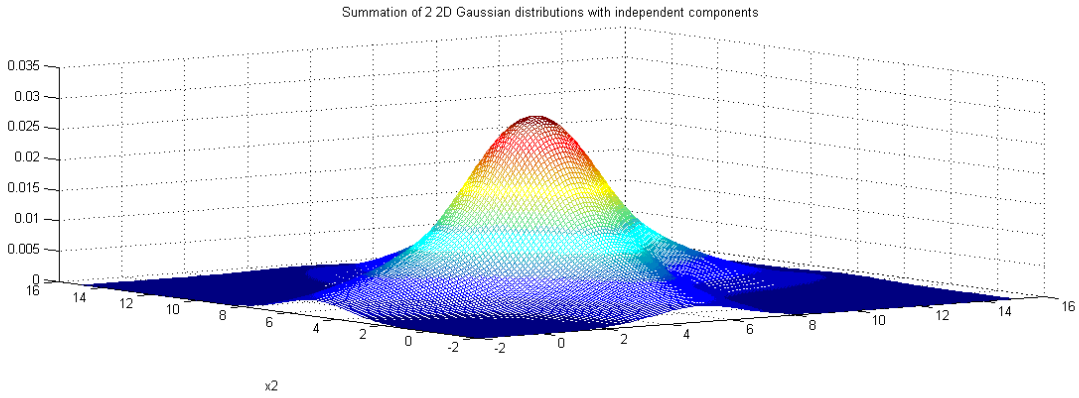
Answer e. (continuation):

Matlab Code:

```
pts = -2:0.1:15;


% first gaussian
m1 = 5;      % mean 1st variable
s1 = sqrt(2);    % std. deviation 1st variable
m2 = 5;      % mean 2nd variable
s2 = sqrt(14);  % std. deviation 2nd variable
px1 = exp(-0.5*((pts-m1)./s1).^2)./(sqrt(2*pi)*s1);
px2 = exp(-0.5*((pts-m2)./s2).^2)./(sqrt(2*pi)*s2);
p_x1x2_1 = px2'*px1;  % Spaltenvektor * Zeilenvektor = Matrix


% second gaussian
m1 = 5;      % mean 1st variable
s1 = sqrt(14);  % std. deviation 1st variable
m2 = 5;      % mean 2nd variable
s2 = sqrt(2);    % std. deviation 2nd variable
px1 = exp(-0.5*((pts-m1)./s1).^2)./(sqrt(2*pi)*s1);
px2 = exp(-0.5*((pts-m2)./s2).^2)./(sqrt(2*pi)*s2);
p_x1x2_2 = px2'*px1;


% adding the two gaussians
p_x1x2 = (p_x1x2_1 + p_x1x2_2)./2;
```

Resulting distribution:





Exercise 3. (Maximum-likelihood estimation)

a. Explain the maximum-likelihood parameter estimation. What is the purpose of this technique and how is it applied? Also write down the central equation(s) and explain the symbols.

Answer a.:

The maximum-likelihood parameter estimation is a technique that can be used for the optimization of the unknown parameters of a given likelihood function:

$$\hat{\theta} = \arg\max_{\theta} \sum_{k} \log p(x_k|\theta)$$

Take each single training sample x_k and compute the likelihood for it with a given parameter theta. Sum over all sample likelihoods → overall likelihood.
Maximize the overall likelihood over possible parameters.
theta_head = optimal parameters

For the following two exercise parts you are given

- a simple ***log-likelihood function*** with unknown parameter a:

  $\log p(x \mid a) = -(a-x)^2 + 1$

  (You may ignore the fact that this function is not normalized.)

- ***training samples***:

  $x_1 = 2,\ x_2 = 1,\ x_3 = 4,\ x_4 = 2$

b. Write a ***Matlab script*** that computes the optimization of parameter `a` by brute-force trial and error strategy. You may assume that `a` lies in the range between -10 and 10.

c. ***Mathematically*** perform the maximum-likelihood optimization of parameter `a` for the given likelihood function and training samples. What is the optimal value of a?

Answer b.:

```
% training samples
x = [2, 1, 4, 2];

max = -inf;
for a = -10:0.01:10
    llik = sum( -(a-x).^2 + 1 );

    fprintf('Current a value: %.3f\n', a);
    fprintf('Likelihood: %.3f\n', llik);
    if llik > max
        best_a = a;
        max = llik;
    end
end
fprintf('Best a value: %.3f\n', best_a);
fprintf('Maximum likelihood: %.3f', max);
```

Answer c.:

Mathematical maximum-likelihood parameter estimate for a single parameter a:

$$\hat{a} = \arg\max_a \sum_{k=1}^{n} \log p(x_k \mid a)$$

$$= \arg\max_a \sum_{k=1}^{4} (-(a - x_k)^2 + 1)$$

$$= \arg\max_a \left[(a - x_1)^2 + (a - x_2)^2 + (a - x_3)^2 + (a - x_4)^2 + 4\right]$$

$$\frac{d}{da}\left[(a - x_1)^2 + (a - x_2)^2 + (a - x_3)^2 + (a - x_4)^2 + 4\right] = 0$$

$$\left[2 \cdot (a - x_1) + 2 \cdot (a - x_2) + 2 \cdot (a - x_3) + 2 \cdot (a - x_4)\right] = 0$$

$$2 \cdot a + 2 \cdot a + 2 \cdot a + 2 \cdot a - 2 \cdot x_1 - 2 \cdot x_2 - 2 \cdot x_3 - 2 \cdot x_4 = 0$$

$$2 \cdot a + 2 \cdot a + 2 \cdot a + 2 \cdot a - 2 \cdot 2 - 2 \cdot 1 - 2 \cdot 4 - 2 \cdot 2 = 0$$

$$8 \cdot a - 18 = 0$$

$$a = \frac{18}{8} = 2.25$$