"Machine Learning and Computational Statistics"

7^{th} Homework

Exercise 1:

Consider the following data stemming from a two-dimensional normal distribution $p(x_1,x_2)$~$N(\mu,\Sigma)$

X={(3.2, 2.9), (2.4, 6.0), (0.70, 4.3), (1.9, 3.5), (2.2, 4.8), (1.2, 2.1), (1.5, 2.1), (2.6, 4.8), (4.2, 7.5), (-1.5, 3.5)}

(a) Estimate the mean and the covariance matrix of $p(x_1,x_2)$ (use the ML method).

(b) Consider the one-dimensional normal pdfs $p_1(x_1)$, $p_2(x_2)$, modeling the features x_1 and x_2, respectively. Do you believe that the independence assumption (i.e. $p(x_1,x_2)=p_1(x_1)$, $p_2(x_2)$) is valid in this case? Explain **very briefly**.

Exercise 2 (python code + text):

Consider a two-class, two-dimensional classification problem for which you can find attached two sets: one for training and one for testing (file *HW6.mat*). Each of these sets consists of pairs of the form $(y_i,\mathbf{x}_i)$, where y_i is the class label for vector $\mathbf{x}_i$. Let N_{train} and N_{test} denote the number of training and test sets, respectively. The data are given via the following arrays/matrices:

- ***train_x*** (a N_{train} x2 **matrix** that contains in its rows the training vectors $\mathbf{x}_i$)
- ***train_y*** (a N_{train}–dim. column **vector** containing the **class labels** (1 or 2) of the corresponding training vectors $\mathbf{x}_i$ included in ***train_x***).
- ***test_x*** (a N_{test}x2 **matrix** that contains in its rows the test vectors $\mathbf{x}_i$)
- ***test_y*** (a N_{test}–dim. column **vector** containing the **class labels** (1 or 2) of the corresponding test vectors $\mathbf{x}_i$ included in ***test_x***).

Assume that the two classes, ω_1 and ω_2 are modeled by normal distributions.

(a) Adopt the **Bayes classifier**.

i. Use the training set to **estimate** $P(\omega_1)$, $P(\omega_2)$, $p(\mathbf{x}|\omega_1)$, $p(\mathbf{x}|\omega_2)$ (Since $p(\mathbf{x}|\omega_j)$ is modeled a normal distribution, it is completely identified by $\boldsymbol{\mu}_j$ and Σ_j. Use the **ML estimates** for them as given in the lecture slides).

ii. **Classify** the points $\mathbf{x}_i$ of the test set, using the **Bayes classifier** (for each point apply the Bayes classification rule and keep the class labels, to an a N_{test}–dim. column **vector** , called ***Btest_y*** containing the **estimated class labels** (1 or 2) of the corresponding test vectors $\mathbf{x}_i$ included in ***test_x.***).

iii. Estimate the error classification probability ((1) **compare** ***test_y*** and ***Btest_y*** , (2) **count** the positions where both of them have the same class label and (3) **divide** with the total number of test vectors).

(b) Adopt the **naïve Bayes classifier**.

i. Use the training set to estimate $P(\omega_1)$, $P(\omega_2)$, $p(x_1|\omega_1)$, $p(x_2|\omega_1)$, $p(x_1|\omega_2)$, $p(x_2|\omega_2)$ (Each $p(\mathbf{x}|\omega_j)$ is written as $p(\mathbf{x}|\omega_j)= p(x_1|\omega_j)* p(x_2|\omega_j)$. Use the **ML estimates** of the mean and variance for each one of the 1-dim. pdfs).

ii. Classify the points $\mathbf{x}_i =[x_{i1}, x_{i2}]^T$ of the test set, using the naïve Bayes classifier (Estimate $p(\mathbf{x}|\omega_j)$ with $p(x_{i1}|\omega_j)* p(x_{i2}|\omega_j)$ and then apply the Bayes rule. Keep the class labels, to an a N_{test}–dim. column **vector** , called ***NBtest_y*** containing the **estimated class labels** (1 or 2) of the corresponding test vectors $\mathbf{x}_i$ included in ***test_x***)

iii. Estimate the error classification probability (work as in the previous case).

(c) Adopt the **minimum Euclidean distance classifier**.

i. Estimate the means of the classes.

ii. Classify the points $\mathbf{x}_i =[x_{i1}, x_{i2}]^T$ of the test set, using the minimum Euclidean distance classifier (Compute the Euclidean distances $||\mathbf{x}-\mu_1||^2$ and $||\mathbf{x}-\mu_2||^2$ and assign $\mathbf{x}$ to the class corresponding to the minimum distance. Keep the class labels, to an a N_{test}–dim. column **vector** , called ***NBtest_y*** containing the **estimated class labels** (1 or 2) of the corresponding test vectors $\mathbf{x}_i$ included in ***test_x***)

iii. Estimate the error classification probability (work as in case (a)).

(d) For each classifier, depict graphically the training set, using different colors for points from different classes.

(e) Report the classification results obtained by the three classifiers and comment on them. Under what conditions, the classifiers would exhibit the same performance?

Hint: After downloading the attached MATLAB file, use the following python code to retrieve the above mentioned matrices and vectors:

```
import scipy.io as sio

Dataset = sio.loadmat('HW6.mat')

train_x = Dataset['train_x']
train_y = Dataset['train_y']
test_x = Dataset['test_x']
test_y = Dataset['test_y']
```