Assignment #1 (Due date: 1403/04/20)

Part I: Analytical problems

1.

In a two-class one-dimensional problem, the pdfs are the Gaussians $\mathcal{N}(0, \sigma^2)$ and $\mathcal{N}(1, \sigma^2)$ for the two classes, respectively. Show that the threshold x_0 minimizing the average risk is equal to

$$x_0 = 1/2 - \sigma^2 \ln \frac{\lambda_{21} P(\omega_2)}{\lambda_{12} P(\omega_1)}$$

where $\lambda_{11} = \lambda_{22} = 0$ has been assumed.

2.

Prove that the covariance estimate

$$\hat{\Sigma} = \frac{1}{N-1} \sum_{k=1}^{N} (\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})(\boldsymbol{x}_k - \hat{\boldsymbol{\mu}})^T$$

is an unbiased one, where

$$\hat{\boldsymbol{\mu}} = \frac{1}{N} \sum_{k=1}^{N} \boldsymbol{x}_k$$

3.

Show that for the lognormal distribution

$$p(x) = \frac{1}{\sigma x \sqrt{2\pi}} \exp\left(-\frac{(\ln x - \theta)^2}{2\sigma^2} \right), \quad x > 0$$

the ML estimate is given by

$$\hat{\theta}_{ML} = \frac{1}{N} \sum_{k=1}^{N} \ln x_k$$

4.

Consider Cauchy distributions in a two-class one-dimensional classification problem

$$f(x \mid \omega_i) = \frac{1}{\pi b} \cdot \frac{1}{1+(\frac{x-a_i}{b})^2} \qquad i = 1,2 \qquad a_2 > a_1$$

(a) By explicit integration, check that the distributions are indeed normalized.

(b) Assuming $P(\omega_1) = P(\omega_2)$, show that $P(\omega_1 \mid x) = P(\omega_2 \mid x)$ if $x = (a_1+a_2)/2$. Plot $P(\omega_1 \mid x)$ for the case $a_1 = 3$, $a_2 = 5$ and $b = 1$. How do $P(\omega_1 \mid x)$ and $P(\omega_2 \mid x)$ behave as $x \to -\infty$? $x \to +\infty$? Explain.

(c) Show that the minimum probability of error is given by

$$P(\text{error}) = \frac{1}{2} - \frac{1}{\pi} \tan^{-1} \left| \frac{a_2 - a_1}{2b} \right|$$

Plot this as a function of $|a_2 - a_1|/b$.

(d) What is the maximum value of P(error) and under which conditions can this occur? Explain.

(e) Design the Bayes minimum error classifier in terms of a_i and b if $P(\omega_1) = P(\omega_2)$. Show the decision boundaries in this case. What is the probability of error?

(f) Design the Bayes minimum risk classifier with the following error weights

$$\begin{pmatrix} \lambda_{11} & \lambda_{12} \\ \lambda_{21} & \lambda_{22} \end{pmatrix} = \begin{pmatrix} 0 & 1 \\ 2 & 0 \end{pmatrix}$$

Show the decision boundaries in this case. What is the probability of error? Compare the results in (e) and (f).

5. We have a SVM to be trained on a set of inputs X_n, where $n = 1, \ldots, N$, together with a corresponding set of target values ($t_n = +1$ *or* -1). The goal is to maximize the margin and at the same time allows some small misclassifications. An example of classify as follows:

$$y(\mathbf{x}_n) = \mathbf{w}^T \phi(\mathbf{x}_n) + b; \qquad \hat{t}_n = \begin{cases} +1, & \text{if } y(\mathbf{x}_n) \geq 0 \\ -1, & \text{otherwise} \end{cases}$$

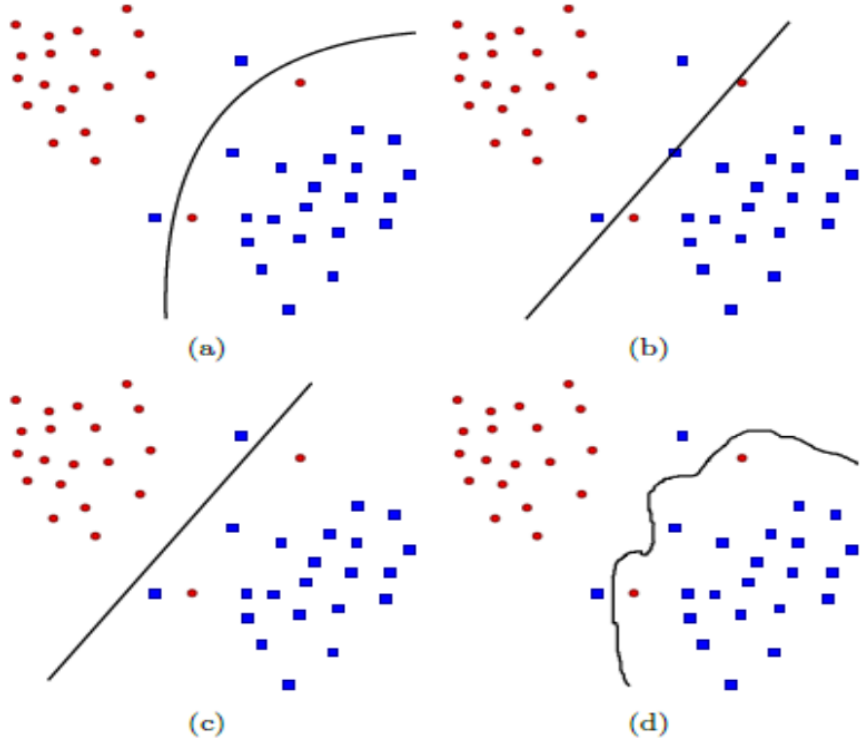
We therefore minimize:

$$\begin{aligned} & C \sum_{n=1}^{N} \xi_n + \frac{1}{2} \|\mathbf{w}\|^2 \\ \text{s.t.} \quad & t_n y(\mathbf{x}_n) \geq 1 - \xi_n, && n = 1, \ldots, N \\ & \xi_n \geq 0, && n = 1, \ldots, N \end{aligned}$$

Where $C > 0$ is a parameter that controls the trade-off between the stack variable penalty and the margin. Figure below illustrates the decision boundaries for four different SVMs

using different values of C and different kernels. For each, specify which setting was used to get the result.

(Note: You should explain why you have chosen each specific choice for the corresponding figures)



Decision boundaries corresponding to the four different SVMs.

1. $C = 1$ and no kernel is used,
2. $C = 0.1$ and no kernel is used,
3. $C = 0.1$ and the kernel is $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \exp -\frac{\|\mathbf{x}_i - \mathbf{x}_j\|^2}{2}$.
4. $C = 0.1$ and the kernel is $\mathbf{K}(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T \mathbf{x}_j + (\mathbf{x}_i^T \mathbf{x}_j)^2$

Part II: Programming assignments

5.

Compute the value of a Gaussian pdf, $\mathcal{N}(m, S)$, at $x_1 = [0.2, 1.3]^T$ and $x_2 = [2.2, -1.3]^T$, where

$$m = [0, 1]^T, \quad S = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

6.

Generate $N = 500$ 2-dimensional data points that are distributed according to the Gaussian distribution $\mathcal{N}(m, S)$, with mean $m = [0, 0]^T$ and covariance matrix $S = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{12} & \sigma_2^2 \end{bmatrix}$, for the following cases:

$\sigma_1^2 = \sigma_2^2 = 1, \sigma_{12} = 0$

$\sigma_1^2 = \sigma_2^2 = 0.2, \sigma_{12} = 0$

$\sigma_1^2 = \sigma_2^2 = 2, \sigma_{12} = 0$

$\sigma_1^2 = 0.2, \sigma_2^2 = 2, \sigma_{12} = 0$

$\sigma_1^2 = 2, \sigma_2^2 = 0.2, \sigma_{12} = 0$

$\sigma_1^2 = \sigma_2^2 = 1, \sigma_{12} = 0.5$

$\sigma_1^2 = 0.3, \sigma_2^2 = 2, \sigma_{12} = 0.5$

$\sigma_1^2 = 0.3, \sigma_2^2 = 2, \sigma_{12} = -0.5$

Plot each data set and comment on the shape of the clusters formed by the data points.

7.

Consider a 2-class classification task in the 2-dimensional space, where the data in both classes, ω_1, ω_2, are distributed according to the Gaussian distributions $\mathcal{N}(m_1, S_1)$ and $\mathcal{N}(m_2, S_2)$, respectively. Let

$$m_1 = [1, 1]^T, \quad m_2 = [3, 3]^T, \quad S_1 = S_2 = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Assuming that $P(\omega_1) = P(\omega_2) = 1/2$, classify $x = [1.8, 1.8]^T$ into ω_1 or ω_2.

8.

The Poisson distribution for the discrete variable $x = 0, 1, 2, \ldots$ and real parameter λ is

$$f_x(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!}$$

a) The *mode* of a distribution is the value of x that has the maximum probability. Prove that the mode of a Poisson distribution is the greatest integer that does not exceed λ, i.e. the mode is $[\lambda]$ (if λ is an integer, then both λ and $\lambda - 1$ are modes).

b) Consider two equally probable categories having Poisson distributions but with differing parameters; assume for definiteness $\lambda_1 > \lambda_2$. What is the Bayes classification decision?

9.

Consider a 2-class classification task in the 3-dimensional space, where the two classes, ω_1 and ω_2, are modeled by Gaussian distributions with means $m_1 = [0, 0, 0]^T$ and $m_2 = [0.5, 0.5, 0.5]^T$, respectively. Assume the two classes to be equiprobable. The covariance matrix for both distributions is

$$S = \begin{bmatrix} 0.8 & 0.01 & 0.01 \\ 0.01 & 0.2 & 0.01 \\ 0.01 & 0.01 & 0.2 \end{bmatrix}$$

Given the point $x = [0.1, 0.5, 0.1]^T$, classify x (1) according to the Euclidean distance classifier and (2) according to the Mahalanobis distance classifier. Comment on the results.

10.

Maximum Likelihood (ML)

Generate 50 2-dimensional feature vectors from a Gaussian distribution, $\mathcal{N}(m, S)$, where

$$m = [2, -2]^T, \; S = \begin{bmatrix} 0.9 & 0.2 \\ 0.2 & 0.3 \end{bmatrix}$$

Let X be the resulting matrix, having the feature vectors as columns. Compute the ML estimate of the mean value, m, and the covariance matrix, S, of $\mathcal{N}(m, S)$ and comment on the resulting estimates.

11.

Generate two data sets, X (training set) and X_1 (test set), each consisting of $N = 1000$ 3-dimensional vectors that stem from three *equiprobable* classes, ω_1, ω_2, and ω_3. The classes are modeled by Gaussian distributions with means $m_1 = [0, 0, 0]^T$, $m_2 = [1, 2, 2]^T$, and $m_3 = [3, 3, 4]^T$, respectively; their covariance matrices are

$$S_1 = S_2 = S_3 = \begin{bmatrix} 0.8 & 0 & 0 \\ 0 & 0.8 & 0 \\ 0 & 0 & 0.8 \end{bmatrix} = \sigma^2 I$$

1. Using X, compute the maximum likelihood estimates of the mean values and the covariance matrices of the distributions of the three classes. Since the covariance matrices are known to be the same, estimate them for each class and compute their average. Use the latter as the estimate of the (common) covariance matrix.

2. Use the Euclidean distance classifier to classify the points of X_1 based on the ML estimates computed before.
3. Use the Mahalanobis distance classifier to classify the points of X_1 based on the ML estimates computed before.
4. Use the Bayesian classifier to classify the points of X_1 based on the ML estimates computed before.
5. For each case, compute the error probability and compare the results (all classifiers should result in almost the same performance. Why?).

12. Expectation Maximization (EM) Algorithm

Generate a set X of $N = 500$ 2-dimensional points that stem from the following pdf:

$$p(x) = \sum_{j=1}^{J=3} P_j p(x|j)$$

where the $p(x|j)$'s, $j = 1,2,3$ are (2-dimensional) normal distributions with mean values $m_1 = [1,\ 1]^T$, $m_2 = [3,\ 3]^T$, $m_3 = [2,\ 6]^T$ and covariance matrices $S_1 = 0.1I$, $S_2 = 0.2I$, $S_3 = 0.3I$, respectively (I is the 2×2 identity matrix). In addition, $P_1 = 0.4$, $P_2 = 0.4$, and $P_3 = 0.2$.

The idea is to use the previously generated data and pretend that we do not know how they were generated. We assume that the pdf $p(x)$ underlying X is a weighted sum of J normal distributions with covariance matrices of the form $S_i = \sigma_i^2 I$, and we employ the EM algorithm to estimate the unknown parameters in the adopted model of $p(x)$. The goal is to demonstrate the dependence of the EM algorithm on the initial conditions and the parameter J. To this end, we use the following sets of initial parameter estimates:

- $J = 3$, $m_{1,ini} = [0,\ 2]^T$, $m_{2,ini} = [5,\ 2]^T$, $m_{3,ini} = [5,\ 5]^T$, $S_{1,ini} = 0.15I$, $S_{2,ini} = 0.27I$, $S_{3,ini} = 0.4I$ and $P_{1,ini} = P_{2,ini} = P_{3,ini} = 1/3$
- $J = 3$, $m_{1,ini} = [1.6,\ 1.4]^T$, $m_{2,ini} = [1.4,\ 1.6]^T$, $m_{3,ini} = [1.3,\ 1.5]^T$, $S_{1,ini} = 0.2I$, $S_{2,ini} = 0.4I$, $S_{3,ini} = 0.3I$ and $P_{1,ini} = 0.2$, $P_{2,ini} = 0.4$, $P_{3,ini} = 0.4$
- $J = 2$, $m_{1,ini} = [1.6,\ 1.4]^T$, $m_{2,ini} = [1.4,\ 1.6]^T$, $S_{1,ini} = 0.2I$, $S_{2,ini} = 0.4I$ and $P_{1,ini} = P_{2,ini} = 1/2$

Comment on the results.

13. Two coins are used for a coin-tossing experiment, that is, coin A and coin B. The probability that coin A returns heads is 0.6, and the respective probability for coin B is 0.4. An individual standing behind a curtain decides which coin to toss as follows: the first coin to be tossed is always coin A, the probability that coin A is re-tossed is 0.4, and similarly, the probability that coin B is re-tossed is 0.6. An observer can only have access to the outcome of the experiment, that is, the sequence of heads and tails that is produced. (a) Model the experiment by means of a HMM (i.e., define the vector of the initial state probabilities, the transition matrix and the matrix of the emission probabilities) and (b) use the *Baum Welch* algorithm to compute the HMM score for the sequence of observations $\{H, H, T, H, T, T\}$ where H stands for heads and T stands for tails.

Good Luck!