COMPUTER ENGINEERING && IT DEPARTMENT

AMIRKABIR UNIVERSITY OF TECHNOLOGY

Statistical Pattern Recognition

Submitted To:
Mohammad Rahmati
Assoc. Professor
Computer Engineering
Department

Submitted By :
Ahmad Asadi
94131091
Group-G1
Fall-95

Contents

1 Problem 1

Let $\{x_k\}, k = 1, 2, ..., n$ denote independent training data from one of the following densities. Obtain the Maximum Likelihood estimate of θ in each case.

- $f(x_k, \theta) = \frac{x_k}{\theta^2} e^{-\frac{x_k^2}{2\theta^2}}, x_k \geq 0, \theta > 0$

  We will first form likelihood function and then find the optimum value of θ maximizing it.

$$l(\theta) = P(D|\theta) = \Pi_{i=1}^{n} P(X_i|\theta) = \Pi_{i=1}^{n} \frac{x_i^2}{\theta^2} e^{-\frac{x_i^2}{2\theta^2}} = \frac{1}{\theta^{2n}} \Pi_{i=1}^{n} (x_i^2 e^{-\frac{x_i^2}{2\theta^2}})$$
$$\rightarrow l(\theta) = \frac{1}{\theta^{2n}} (\Pi_{i=1}^{n} x_i^2) e^{\Sigma_{i=1}^{n} - \frac{x_i^2}{2\theta^2}} \rightarrow ln(l(\theta)) = ln \frac{\Pi_{i=1}^{n} x_i^2}{\theta^{2n}} - \Sigma_{i=1}^{n} \frac{x_i^2}{2\theta^2}$$
$$\nabla_\theta ln(l(\theta)) = \frac{-2n\theta^{-2n-1} \Pi_{i=1}^{n} X_i}{\theta^{-2n} \Pi_{i=1}^{n} X_i} - 4\theta^{-3} \Sigma_{i=1}^{n} x_i^2 = \frac{-2n}{\theta} - 4\theta^{-3} \Sigma_{i=1}^{n} X_i^2 = 0$$
$$\rightarrow \theta^* = \sqrt{\frac{2}{n} \Sigma_{i=1}^{n} X_i^2}$$

- $f(x_k, \theta) = \sqrt{\theta} x_k^{\sqrt{\theta}-1}, 0 \leq x_k \leq 1, \theta > 0$

$$l(\theta) = P(D|\theta) = \Pi_{i=1}^{n} \sqrt{\theta} X_i^{\sqrt{\theta}-1} = \theta^{\frac{n}{2}} \Pi_{i=1}^{n} X_i^{\sqrt{\theta}-1}$$
$$\rightarrow ln(l(\theta)) = \frac{n}{2} ln(\theta) + (\sqrt{\theta} - 1) ln(\Pi_{i=1}^{n} X_i)$$
$$\rightarrow \nabla_\theta ln(l(\theta)) = \frac{n}{2\theta} + \frac{1}{2} \theta^{-\frac{1}{2}} ln(\Pi_{i=1}^{n} X_i) = 0 \rightarrow \theta^* = (\frac{-n}{\Sigma_{i=1}^{n} ln X_i})^2$$

---

2 Problem 2

Let x have uniform density:

$$f_x(x, \theta) \propto U(0, \theta) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq 1 \\ 0 & otw \end{cases} \quad (1)$$

1. Suppose that n samples $D = \{x_1, x_2, \cdots, x_n\}$ are drawn independently according to $f_x(x|\theta)$. Show that the maximum likelihood estimate for θ is $max[D]$, i.e., the value of the maximum element in D.

   Since all samples are drawn independently, we have:

   $$P(D|\theta) = \Pi_{i=1}^{n} P(X_i|\theta) \tag{2}$$

   According to proposed uniform density function, if there exists $X_i \in D$ such that $X_i < \theta$ then $P(X_i|\theta) = 0$, therefore according to (2), $P(D|\theta) = 0$. So the likelihood of such values of θ will be equal to zero, hence we can conclude that:

   $$\forall\theta, \exists X_i \in D; X_i < \theta \Leftrightarrow l(\theta) = 0. \tag{3}$$

   According to (3), it is meaningful to assume $\theta \geq max[D]$ to get regions of nonzero likelihood. Having uniform density in mind and considering $\theta \geq max[D]$ as the result of (3):

   $$P(D|\theta) = \Pi_{i=1}^{n} P(X_i|\theta) = \frac{1}{\theta^n} \tag{4}$$

   As it is obviously clear, (4) does not have optimal value globally but taking (3) constraint into consideration yields $\theta^* = max[D]$, since the gradient of likelihood function is negative and so the likelihood function is monotonically decreasing in the range of $[max[D], +\inf)$.
   Figure 1 represents the proposed likelihood function in (4).

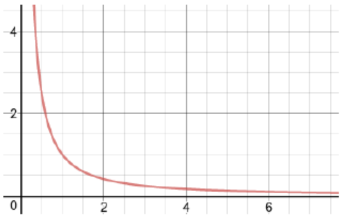


Figure 1: The proposed likelihood function in 4

---

2. Suppose that $n = 5$ point are drawn from the distribution and the maximum value of which happens to be $maxx_k = 0.6$. Plot the likelihood function $f_x(D|\theta)$ in the range $0 \leq \theta \leq 1$. Explain in words why you do not need to know the values of other four points.
   As discussed in the first part, the likelihood function $f_x(D|\theta)$ is zero when

$\theta < max[D]$, because the conditional probability of each sample is represented as $P(X_i|\theta) \propto U(0, \theta)$.
giving more details, whenever $\theta < max[D]$, there exists at least one sample $X_i \in D$ such that $\theta < X_i$. Therefore, taking the fact that all samples are drawn from a uniform distribution, $P(X_i|\theta) = 0$ which implies that $P(D|\theta) = 0$. Keeping this fact in mind, knowing the values of points rather than the maximum one, is absolutely useless.
On the other cases in which $\theta \leq max[D]$, we have $P(D|\theta) = \frac{1}{\theta^n}$. As here $n = 5$, the likelihood function will be in the form of $P(D|\theta) = \frac{1}{\theta^5}$.
Figure 2 displays the likelihood function of this problem.

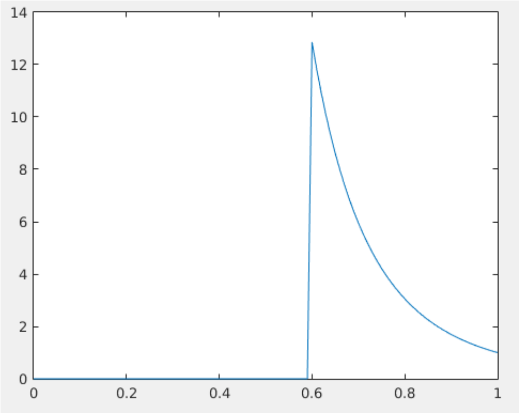


Figure 2: The likelihood functon of the problem 2_2.

3 Problem 3

Consider the standard two class SVM with the hinge loss. Argue that under a given value of regularization parameter:

$$\text{Leave-one-out-Error} < \frac{\#SVs}{l}$$

Where l is the size of training data and $\#SVs$ is the number of support vectors obtained by training SVM on the entire set of training data.
To calculate leave-one-out-error, we first train SVM on $n - 1$ training samples and test the other remaining one. If the remaining point is a support vector for the full n sample case, then there will be an error. Note that if we can find a transformation

$\phi(.)$ that well separates the data, then Leave-one-out-Error $< \frac{\#SVs}{l}$ shows that the expected error rate will be lower.
This bound is also independent of the dimensionality of the space of transformed vectors determined by $\phi(.)$.

4 Problem 4

Consider the 2-dimensional points and their classification (+ or -) below:

x	y	class
0	4	+
8	3	+
6	-2	-
4	0	-
2	1	-

- The points with classification + and - corresponds to the point sets M_+ and M_- , respectively. Draw the points and determine first whether or not the sets M_+ and M_- are linearly separable. And then whether or not the two sets are linearly separable by a 2- dimensional perceptron.
  To determine that if a data set containing samples of two different classes is linearly separable, it is sufficient to check whether there exists a separating line such that $a^T X + w_0.\underset{<_{M_-}}{\overset{M_+}{\geq}}0$.

  Figure 4 displays the given data points and a proposed separating line $0.5x_1 + x_2 - 3 = 0$. As it is illustrated in this figure, the given data set is linearly separable.

  Since, none of the separating lines will not cross the origin point, it is necessary to have bias terms augmented in perceptron model. Therefore, a 2-dimensional perceptron, without augmented bias terms, which is just able to draw lines crossing zero point, is not able to classify given dataset.

  Another way rather than drawing line graphically is to first negate all samples from one of the classes, for example instances from class M_-, and then find any vector a such that $a^T y > 0$ for all $y \in D$, denoting dataset by D.

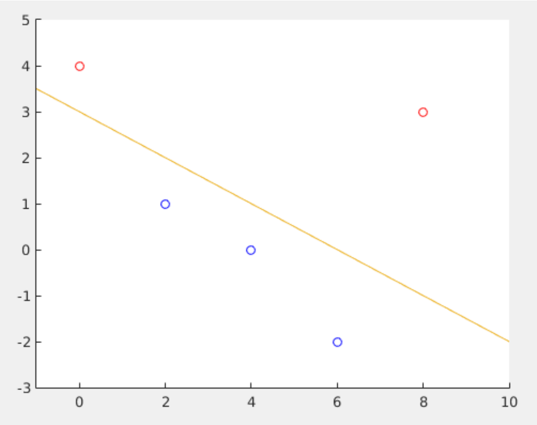


Figure 3: Given dataset and separating line in problem 4. The red points are points from M_+ and the blue points are points from M_-.

Negating all samples from second class and considering 3-dimensional data samples, in bias augmented model, yields following vector:

$$a = [-310 - 4]^T$$

Calculation of this vector is illustrated in table 1 and figure **??** represents the underlying line by it.

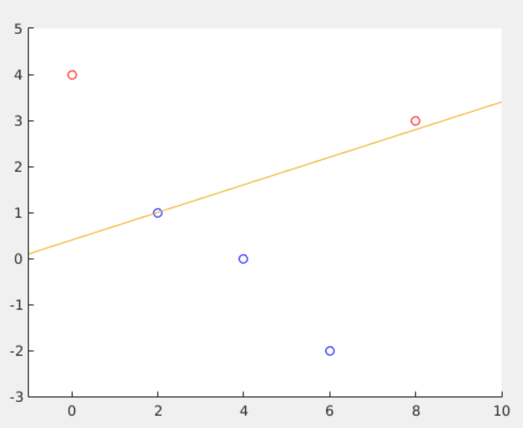


Figure 4: Given dataset and separating line by a 3-dimensional perceptron in problem 4. The red points are points from M_+ and the blue points are points from M_-.

- Manually execute the perceptron learning algorithm on this dataset. Based on your answer from part (a) decide whether or not you need a bias. Use a vector of all ones as the initial weight vector. Write all the intermediate results of your perceptron computation in a table.

We first negate all the samples of class M_- and then augment a bias component to perceptron weights and an augmented 1 to data instances. It is obviously clear that augmented 1s in samples from M_- will be negated too.

Table 1 illustrates all steps of perceptron learning algorithm. In all cases that $a^T y < 0$ we will consider a misclassification and we will consider all $y \in D$ in which $a^T y = 0$, $y \in< M_-$.

- Give the linear function that has been learned by this perceptron.
  Linear function representing by this perceptron is presented in equation (5)

$$f(X) = -3X_1 + 10X_2 - 4. \gtrless^{M_+}_{\leq_{M_-}} 0 \tag{5}$$

- Classify point (5,2) base on the trained perceptron.
  Computing the value of $f([5,2])$:

$$f([5,2]) = -15 + 20 - 4 = 1 > 0$$

  So the point (5,2) is in M_+.

---

5 Computer Exercises

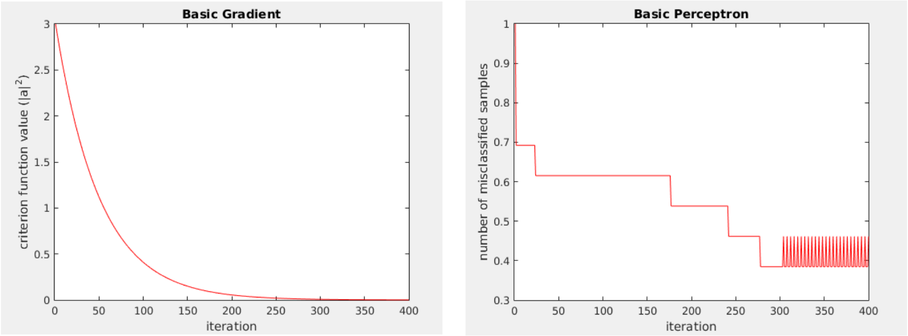
All the scripts are properly named and attached to the submitted file.

5.1 Exercise 1

1. The criterion function used in gradient descent method is the $l2 - norm$ of the transformation matrix a and for preceptron is the number of misclassified samples. Figure 5 displays criterion function changes over iterations for both algorithms.

Table 1: Perceptron learning algorithm steps in case of considering bias components.

y_1	y_2	y_3	a_1	a_2	a_3	$a^T y$
0	4	1	1	1	1	>0
8	3	1	1	1	1	>0
-6	2	-1	1	1	1	<0
-4	0	-1	-5	3	0	>0
-2	-1	-1	-5	3	0	>0
0	4	1	-5	3	0	>0
8	3	1	-5	3	0	<0
-6	2	-1	3	6	1	<0
-4	0	-1	-3	8	0	>0
-2	-1	-1	-5	7	-1	<0
0	4	1	-5	7	-1	>0
8	3	1	-5	7	-1	<0
-6	2	-1	3	10	0	>0
-4	0	-1	3	10	0	<0
-2	-1	-1	-1	10	-1	<0
0	4	1	-3	9	-2	>0
8	3	1	-3	9	-2	>0
-6	2	-1	-3	9	-2	>0
-6	2	-1	-3	9	-2	>0
-4	0	-1	-3	9	-2	>0
-2	-1	-1	-3	9	-2	<0
0	4	1	-5	8	-3	>0
8	3	1	-5	8	-3	<0
-6	2	-1	3	11	-2	>0
-4	0	-1	3	11	-2	<0
-2	-1	-1	-1	11	-3	<0
0	4	1	-3	10	-4	>0
8	3	1	-3	10	-4	>0
-6	2	-1	-3	10	-4	>0
-4	0	-1	-3	10	-4	>0
-2	-1	-1	-3	10	-4	=0



(a) Basic gradient algorithm

(b) Basic perceptron algorithm

Figure 5: Criterion function changes over iterations in both perceptron and gradient algorithms.

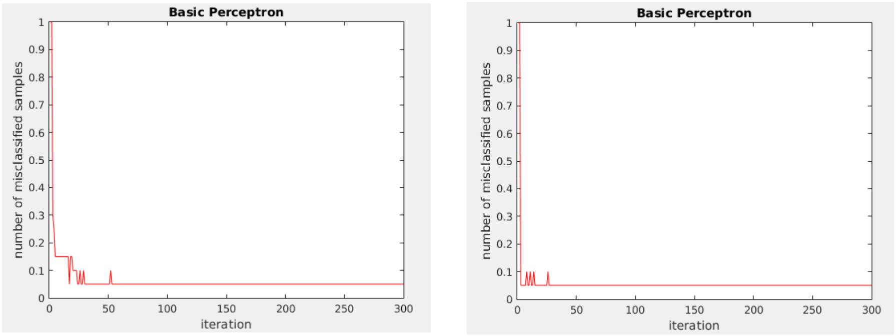
2. In basic gradient algorithm in each iteration there exists just 2 mathematical operations. Therefore after 400 iterations the estimated total number of mathematical operations would be 800. On the other hands, in basic perceptron algorithm there exists $1 + |Y|$ mathematical operations in each iteration, note that $|Y|$ is 1 if the current sample is misclassified and 0 otherwise. So the estimated total mathematical operations should be $\#iterations + \#misclassifiedSamples$. The estimated number is $300 + 2700 = 3000$.

5.2 Exercise 2

As it is clearly obvious from distribution of samples from 3 classes, in cases that separabality is higher, the perceptron algorithm can reach the optimum value faster and the convergence speed would be higher. Figure 6 illustrates classification error during epochs of perceptron in two cases. The right side figure is error in classifying W_1 and W_2 and the left side figure illustrates that of W_2 and W_3.

5.3 Exercise 9

The script is implemented in MATLAB and its source file is attached to submitted zip file.



(a) classification of W_1 and W_2

(b) classification of W_2 and W_3

Figure 6: Classification error in cases of different levels of separabalities in data.