University of Tehran

School of Electrical and Computer Engineering

Pattern Recognition

Assignment 3

Due Date:

Azar 16th

Corresponding TAs:

Maryam kazemi (Q1, Q4, Q7): m.kazemi7535@ut.ac.ir

Mahyar Khordehforoosh (Q2, Q6, Q7): m.khordehforoosh@gmail.com

Hamidreza Aftabi (Q3, Q5, Q7): hamid.aftabi@gmail.com

Aban 98

PROBLEM 1

In this exercise, we will take a look at one estimation technique of the probability density functions $P(x|\omega_i), i = 1, 2, \dots, L$, based on the Bayes classification rule, and the available training set, X. Naive Bayes classifiers are a family of simple classifiers, based on the Bayes theorem.

1.1 Write the conditional probability of any feature vector X=[$x_1 \dots x_L$] given the class ω_k (the likelihood function) and explain the assumption under which you wrote the probability. Then, write the decision rule corresponding to a Naive Bayes classifier.

1.2 Imagine we have two classes with prior probabilities (ω_1) and (ω_2); And ($x_i|\omega_k$) is a probability density function from Exponential Family of distributions. Find the decision boundary equation of Naive Bayes classifier.

1.3 Now, we apply Naive Bayes classifier to compute the probability of poisonous under some circumstances.

In the following table, you can find the dataset of playing tennis in 8 example in a row, containing some conditions, leading to a conclusion of it is poisonous or not .

Using a Naive Bayes classifier rule to classify example 8.

example	color	toughness	fungus	appearance	poisonous
1	Green	Hard	No	Smooth	No
2	Green	Hard	Yes	Smooth	No
3	Brown	Soft	No	Wrinkled	No
4	Orange	Hard	No	Wrinkled	Yes
5	Green	Soft	Yes	Smooth	Yes
6	Green	Hard	Yes	Wrinkled	Yes
7	Orange	Hard	No	Wrinkled	Yes
8	Green	Soft	Yes	Wrinkled	?

Hint: More details about Naive Bayes classifiers, may be found in “Pattern Recognition; Sergios Theodoridis; Chapter 2.5.7”.

PROBLEM 2

I. Consider a sample of n observations from a pdf p which is shown in figure 1. An estimate of the pdf $\hat{p}$ is calculated using a kernel (Parzen) density estimate with Gaussian kernels for various bandwidths h. How would you expect the number of relative maxima of $\hat{p}$ to vary as h increases?

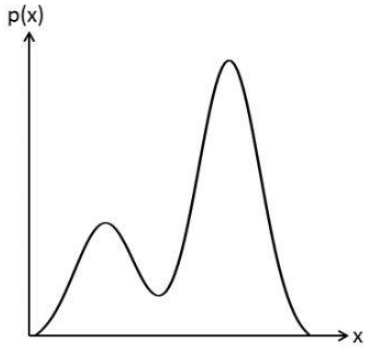

Figure 1: The Probability Density Function for problem 2 part I

II. The natural way for choosing h is to plot out several curves and choose the estimate that best matches one's prior (subjective) ideas. However, this method is not practical in pattern recognition since we typically have high-dimensional data. A rule of thumb is to reference to a standard distribution and find the value of bandwidth that minimizes the Mean of the Integral of the Square Error (MISE).

$$h_{MISE} = argmin\{E\left[\int \left(\hat{f}(x) - f(x)^2 dx\right]\right\}$$

If we assume that the true distribution is Gaussian and we use a Gaussian kernel, it can be shown that the optimal value of h is:

$$h^* = 1.06\sigma N^{-1/5}$$

Where σ is the sample standard deviation and N is is the number of training examples. Discuss about the practicality of this bandwidth providing examples.

III. Let $p(x) \sim U(0, a)$ be uniform in interval $[0, a]$, and let a Parzen window be defined as bellow:

$$\phi(x) = \begin{cases} e^{-x} & x > 0 \\ 0 & otherwise \end{cases}$$

Show that the mean of such a Parzen-window estimate with n samples is given by:

$$\bar{p}_n(x) = \begin{cases} 0 & x < 0 \\ \frac{1}{a}\left(1 - e^{-\frac{x}{h_n}}\right) & 0 \le x \le a \\ \frac{1}{a}\left(e^{\frac{a}{h_n}} - 1\right)e^{-\frac{x}{h_n}} & a \le x \end{cases}$$

PROBLEM 3

Show that the variance $\sigma_N^2(x)$ of the pdf estimate, given by:

$$\hat{p}(x) = \frac{1}{N}\left(\sum_{i=1}^{N} \varphi\left(\frac{x_i - x}{b}\right)\right)$$

is upper bounded by:

$$\sigma_N^2(x) \leq \frac{\sup(\varphi)\, E[\hat{p}(x)]}{N b^l}$$

where sup(.) refers to the supremum of the associated function.

PROBLEM 4

I. Design and implement a Bayes optimal classifier with Gaussian parametric estimate of pdfs to minimize the probability of classification error. You must state the equations which are used for the parameter estimation, and also explain how you choose the prior probabilities of the classes.

II. When estimating the parameters of a Gaussian distribution, sometimes a singular matrix is obtained as the covariance of the data.

a) Why this situation is problematic?

b) This difficulty arises for the given dataset. By using the following hint, study the proposed methods, and apply one of them to your classifier. Evaluate your classifier by means of correct classification rate and confusion matrix.

Hint: https://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/old_IDAPILecture16.pdf

PROBLEM 5

In many pattern classification problems, one has the option either to assign the pattern to one of the C classes, or to reject it as being unrecognizable. If the cost for rejects is not too high, rejection may be a desirable action. Let:

$$\lambda_{ij} = \begin{cases} 0 & i = j \\ \lambda_r & i = c+1 \\ \lambda_s & otherwise \end{cases}$$

Where λ_r is the loss incurred for choosing the $(\mathrm{C}+1)^{th}$ action, rejection, and λ_s is the loss incurred for making a substitution error.

Modify the classifier that you designed in problem 4-1 to add the option of rejection. By utilizing an exhaustive search, plot a 3D graph using CCR, λ_r, and λ_s values. What is the highest CCR? Report λ_r and λ_s of the classifier which yields the highest CCR.

PROBLEM 6

I. Repeat Problem 4-1 with Parzen non-parametric estimate of pdfs. Study the effect of window size carefully and report the probability of classification error and correct

classification rate. Consider two different windows: Rectangular and Gaussian. Compare the results for the windows.

II. Repeat problem 4-1 with k-nearest neighbor (k-NN) non-parametric estimate of pdfs. Study the effect of number of samples k. Report the probability of classification error and correct classification rate

III. Design and Implement a k-nearest neighbor classifier. Report the correct classification rate for k= 1, 3, 5, 10.

PROBLEM 7

7.1. You have already implemented a number of different classifiers. Using pre-defined functions of Scikit-Learn package try to implement these classifiers:

- KNN Classifier (KNeighborsClassifier)
- Parzen non-parametric estimate of pdfs (RadiusNeighborsClassifier)
- Gaussian Naive Bayes(GaussianNB)

7.2. Compare the classifiers of problems 4, 5 and 6 in terms of:

a) Correct Classification Rate

b) Confusion Matrix

c) Confidence Matrix

d) Required time for Training the algorithm

e) Required time for testing the algorithm

Which classifier is your choice for given dataset? Explain why.

DATASET DESCRIPTION

1. In questions 4,5,6 and 7 you should use pen-based recognition of handwritten digits data set. For further information visit here. In the *data folder*, use *pendigits.tra* and *pendigits.tes* to train and test your classifiers.
2. In these problems, if the classifiers take a lot of time to run, you may examine your classifiers on a portion of test dataset. If you do so, please state in your report how many test samples you used and extrapolate the time needed to run algorithms for all test samples. (Use at least 3000 train and 1000 test samples)
3. Make sure to include train and test data in your submitted homework. Your code should also be runnable on any device without changing the directory path of the dataset.

NOTES

1. Please make sure you reach the deadline because there would be no extra time available.
2. Late policy would be as bellow:
   - Every student has a budget for late submission during the semester. This budget is two weeks for all the assignments.
   - Late submission more than two weeks may cause lost in your scores.
3. Analytical problems can be solved on papers and there is no need to type the answers. The only thing matters is quality of your pictures. Scanning your answer sheets is recommended. If you are using your smartphones you may use scanner apps such as CamsSanner or google drive application.
4. Simulation problems need report as well as source codes and results. This report must be prepared as a standard scientific report.
5. You have to prepare your final report including the analytical problems answer sheets and your simulation report in a single pdf file.
6. Finalized report and your source codes must be uploaded to the course page as a ".zip" file (not ".rar") with the file name format as bellow:

   PR_Assignment #[Assignment Number]_Surname_Name_StudentID.pdf

7. Plagiarisms would be strictly penalized.
8. You may ask your questions from corresponding TAs.