University of Tehran

School of Electrical and Computer Engineering

Pattern Recognition

Assignment 3

Due Date: 14th of Ordibehesht

Corresponding TAs:

Saba Tabatabaee – sabatabatabaee@ut.ac.ir

Sima Hooshangi –sima.hooshangi@gmail.com

Farvardin 98

PROBLEM 1

The Naive Bayes classifier selects the most likely classification V_{nb} given the attribute values $a_1, a_2, \ldots a_n$. This results in:

$$V_{nb} = argmax_{v_j \epsilon V} P(v_j) \prod P(a_i|v_j)$$

We generally estimate $P(a_i|v_j)$ using m-estimates:

$$P(a_i|v_j) = \frac{n_c + mp}{n + m}$$

where:

$$n = the\ number\ of\ training\ examples\ for\ which\ v = v_j$$
$$n_c = number\ of\ examples\ for\ which\ v = v_j\ and\ a = a_j$$
$$p = a\ priori\ estimate\ for\ P(a_i|v_j)$$
$$m = the\ equivallent\ sample\ size$$

Attributes are Color, Type, Origin, and the subject, stolen can be either yes or no.

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Classify a Red Domestic SUV. Note there is no example of a Red Domestic SUV in our data set.

PROBLEM 2

Consider a normal data distribution $p(x) \sim N(\mu, \sigma^2)$ and a normal Parzen window function $\phi(x) \sim N(0,1)$. Show that the Parzen window estimate

$$p_n(x) = \frac{1}{nh_n} \sum_{i=1}^{n} \varphi\left(\frac{x - x_i}{h_n}\right)$$

has the following properties:

1. $\bar{p}_n(x) \sim N(\mu, \sigma^2 + h_n^2)$
2. $var[p_n(x)] \simeq \frac{1}{2nh_n\sqrt{\pi}} p(x)$
3. $p(x) - \bar{p}_n(x) \simeq \frac{1}{2}\left(\frac{h_n}{\sigma}\right)^2 \left[1 - \left(\frac{x-\mu}{\sigma}\right)^2\right] p(x)$

For small h_n .(note that if $h_n = h_1/\sqrt{n}$, this result implies that the error due to bias goes to zero as $1/n$, whereas the standard deviation of the noise only goes to zero as $\sqrt[4]{n}$.)

PROBLEM 3

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 4

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 $(C+1)^{th}$ action, rejection, and λ_s is the loss incurred for making a substitution error. Here, we assume the following values for the losses:

$$\lambda_r = 0.8, \lambda_s = 1$$

Modify the classifier that you designed in problem 3 to add the option of rejection.

PROBLEM 5

I. Repeat Problem 3 (part I) with Parzen non-parametric estimate of pdfs. Study the effect of window size carefully, report the probability of classification error and correct classification rate. Consider two different windows: a. Rectangular and b. Gaussian. Compare the results for the windows.

II. Repeat problem 3 (part I) 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.

Important Note: In this problem, if the classifiers take a lot of time to run, you may examine your classifiers on a portion of test dataset. (e.g. first 500 sample). 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 500 test samples)

IV. Design a minimum mean distance classifier. Report the correct classification rate.

PROBLEM 6

6.1. 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)

6.2. Compare the classifiers of problems 3, 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.

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 CamsScanner 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.zip

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