University of Tehran

School of Electrical and Computer Engineering

Pattern Recognition

Assignment 3

Due Date: 5th Azar

Corresponding TAs:

Hadi Mehdizavareh – mhzavareh90@yahoo.com

Naser Hakimi – na3erhakim93@yahoo.com

Aban 96

Problem 1

“Naive Bayes” classifiers are a family of quite simple classifiers which are based on the Bayes theorem and independence assumption of the features.

Because of the independence assumption, the conditional probability of any feature vector $x = [x_1 \ldots x_L]$ given the class ω_k (the likelihood function) is written as:

$$P(x|\omega_k) = \prod_{i=1}^{L} P(x_i|\omega_k)$$

Therefore, the following decision rule corresponds to a Naïve Bayes classifier:

$$\omega^* = \operatorname{argmax}_{\omega} \left\{ \prod_{i=1}^{L} P(x_i|\omega_k)\ P(\omega) \right\}$$

Assume a Classification Problem where:

- We have two classes with prior probabilities $P(\omega_1)$ and $P(\omega_2)$
- $P(x_i|\omega_k)$ is a probability density function from Exponential Family of distributions.

1.1. Find the decision boundary equation of Naive Bayes classifier.

1.2. Can we consider Naive Bayes as a linear classifier in this particular problem (for exponential distribution)? How about general case (other distributions)? Give your reasons. (If Linearity for Naive Bayes isn’t general, you may give a counterexample for that)

Problem 2

2.1. For the set of samples $X = \{-7, -5, -4, -3, -2, 0, 2, 3, 4, 5, 7\}$ in a one-dimensional problem, find the Parzen window estimate $P_j(x)$ for a rectangular window. Use $h_j = \frac{1}{\sqrt{j}}$. Sketch the results as a function of x for $j = 1,4,11$.

2.2. Suppose $h_j = \frac{h}{\sqrt{j}}$ where h is a constant. Comment on the shape of $P_j(x)$ for various choices of h.

Problem 3

3.1. 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.

3.2. 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

5.1. Repeat Problem 3-1 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.

5.2. Repeat problem 3-1 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.

5.3. 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)

Problem 6

Design a linear classifier. Consider three design methods based on $L-1$, L (one versus all), and $\frac{L(L-1)}{2}$ (one versus one) hyper planes (L is number of classes). Report the correct classification rate in this three cases. You are free to choose the optimization algorithm used to tune the parameters of the classifier.

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)
- One-Vs-One Classifier(OneVsOneClassifier)

7.2. <u>Compare</u> the classifiers of problems 3, 5, 6, and 7.1 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.

Important Notes:

1. In problems 3 to 7, you have to use the small version of uploaded dataset, named "**<u>TinyMNIST</u>**". The dataset is available on your course page in "Dataset" section. <u>Read dataset description before using it.</u>
2. A code file "FeatureSelection.py" is attached, which load the dataset from the .csv files and apply feature selection to them (Number of features are reduced from 196 to 62). In this assignment, you should use the output of this code for training and testing of your algorithms, which are four NumPy arrays "train_data", "train_labels", "test_data", and "test_labels".
3. In <u>problems 3 to 6</u>, you are **<u>not</u>** allowed to use pre-defined functions. Although you can use available online codes as a guide, it is recommended to implement the algorithms from scratch.
4. This assignment is worth more marks(**<u>Two times</u>** of any other assignments)

P.S.: Don't hesitate to ask your question about Python in course forum. The TAs and other students can help you there.