University of Pittsburgh
CS1675 Introduction to Machine Learning
Professor Milos Hauskrecht

Handout 3
February 11, 2021

Problem assignment 3

Due: Thursday, February 18, 2021

Problem 1. Multivariate Gaussian

Assume the pairs of real valued measurements in file 'gaussian.txt'

- (a) Plot the data using the scatter plot matlab function.
- (b) Calculate and report the ML estimate of the mean and the covariance matrix from the data. Please use the unbiased estimate of the variance. Plot and report the resulting Gaussian distribution. (Note: you need to plot this in 3D).
- (c) Now consider each measurement in 'gaussian.txt' separately. Calculate the ML estimate of the mean and variance of these measurements. Plot and report the individual distributions.
- (d) Do you believe the multivariate Gaussian model is better than two separate univariate Gaussian models? Explain why yes or why not? How would you use the data to answer that question?

Problem 2. Poisson distribution

The Poisson distribution is used to model the number of random arrivals to a system over a fixed period of time. Examples of systems in which events are determined by random arrivals are: arrivals of customers requesting the service, occurrence of natural disasters, such as floods, etc. The Poisson distribution is defined as:

$$p(x|\lambda) = \frac{e^{-\lambda}\lambda^x}{x!} \text{ for } x = 0, 1, 2, \cdots$$

where λ is a parameter. The mean of the Poisson distribution is λ.

Answer the following questions:

- (a) Plot and report the probability function for Poisson distributions with parameters $\lambda = 2$ and $\lambda = 6$. Note that the Poisson model is defined over nonenegative integers only.

- (b) Given a set of independent observations $x_1, x_2, \ldots, x_n$ from a Poisson distribution, the ML estimate of the parameter λ is:

$$\lambda_{ML} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Assume the data in 'poisson.txt' file that represent the number of incoming phone calls received over a fixed period of time. Compute and report the ML estimate of the parameter λ. Also plot and report the probability fuction for the ML parameter.

- (c) The conjugate prior for λ defining the Poisson distribution is Gamma distribution. It is defined as:

$$p(\lambda|a,b) = \frac{1}{b^a\Gamma(a)}\lambda^{(a-1)}e^{-\frac{\lambda}{b}}.$$

Plot and report the Gamma distribution for the following set of parameters ($a = 1, b = 2$) and ($a = 3, b = 5$).

- (d) Assuming the prior distribution on λ is $Gamma(\lambda|a,b)$, the posterior distribution for λ after seeing observations $D = \{x_1, x_2, \ldots, x_n\}$ is again gamma distribution:

$$p(\lambda|D) \sim Gamma(\lambda|a + \sum_{i=1}^{n} x_i, \frac{b}{nb+1}).$$

Please use data in 'poisson.txt' to calculate and plot the posterior distributions of λ for both priors in Part c.

Problem 3: Non-parametric density estimation

In this problem we implement two non-parametric density estimation methods: Parzen window, and KNN, and after that apply them to calculate $p(x)$. The (training) data are stored in file 'NDE_data.txt'. The training data consists of 75 one-dimensional instances. The test data to which the methods will be applied to are in file 'NDE_test.txt. The instances are samples from [0,1] interval generated from an 'unknown' density which we will try to estimate.

Part 1. Load the data in 'NDE_data.txt' and plot them using hist function with 20 bins.

Part 2. Write (and submit) a matlab function $[density\_x] = Parzen\_window(x, h, D)$, where x is an instance (a real-number) for which we want to estimate the density, h is the width parameter that defines the Parzen window, and D is the training dataset that is used to calculate the estimate. The function should returns $p(x)$ estimate for x.

- (a) Use your $Parzen\_window$ function with the width parameter $h = 0.025$ and apply it to all instances in the testing data file. Report the results. In addition, use data

instances data instances X=0:0.01:1 to generate a plot of density function modeled by the $Parzen\_window$ methods with $h = 0.025$.

- (b) Repeat the process in Part a, but now use the width parameter $h = 0.1$.
- (c) Compare the results for Parts a and b. Are the estimates for the testing data instances very different? Discuss the shape, properties and diferences between the density function plots.

Part 3. Write (and submit) a matlab function $[density\_x] = kNN(x, k, D)$, where x is an instance (a real-number) for which we want to estimate the density, k is the number of instances used to calculate the estimate, and D is the training dataset that is used to calculate the estimate. The function should returns $p(x)$ estimate for x.

- (a) Use your kNN function with $k = 3$ and apply it to all instances in the testing data file. Report the results. In addition, use data instances X=0:0.01:1 to generate a plot of the density function modeled by the kNN method with $k = 3$.
- (b) Repeat the process in Part a, but now use $k = 5$.
- (c) Compare the results for Parts a and b. Discuss the differences between density values obtained for the testing data, as well as, the shape, properties and differences of your density function plots.