University of Pittsburgh
CS 1675 Introductio to Machine Learning
Professor Milos Hauskrecht

Handout 3
September 14, 2017

Problem assignment 2

Due: Thursday, September 21, 2017

Problem 1. Mean estimates and the effect of the sample size

In this problem we study the influence of the sample size on the estimate of the mean. The data for this experiment are in file mean-study-data.txt in the homework assignment folder. The data were generated from the normal distribution with mean=15 and standard deviation=5.

- (Part 1) Load the data in the mean-study-data.txt Calculate and report the mean and standard deviation of the data.
- (Part 2) Write (and submit) a function [newdata] = subsample(data, k) that randomly selects k instances from the data in the mean-study-data.txt
- (Part 3) Use the above function to randomly generate 1000 subsamples of the data of size 25. For each subsample calculate its mean and save the results in the vector of 1000 means. Plot a histogram of the mean values using 20 bins.
- (Part 4) Include the histogram in your report. Analyze the mean that was calculated in step 1 on all examples in the dataset and the means calculated on 1000 subsamples of size 25. Report your observations.
- (Part 5) Repeat step (part) 3 but now generate 1000 subsamples of size 40. Include the histogram in the report and compare it to the histogram generated in part4 for subsamples of size 25, and to the mean of the original data. What are the differences? What conclusions can you make by comparing means for subsamples of size 25 and 40.

Problem 2. Train-test splitting using k-fold crossvalidation

When testing the performance of a learning algorithm using a simple holdout method the results may be biased by the training/testing data split. To alleviate the problem various random resampling schemes, such as k-fold cross-validation, random subsampling or bootstrap (see lecture notes for Class 4) can be applied to estimate the statistics of interest by averaging the results across multiple train/test splits. Please do the following tasks:

- (Part 1) Please write and submit the function: $[train\ test] = kfold\_crossvalidation(data, k, m)$ that takes the data, k (the number of folds) and m (the target fold) as inputs, and returns the training and testing data sets, such that the testing set corresponds to m-th fold under the k-th fold cross-validation scheme. To implement the procedure please place the folds over indexes of the data, by assuring that each fold has equal number of entries that do not overlap. If this is not possible, the fold sizes (number of instances in each fold) should differ by at most one. The file should be named $kfold\_crossvalidation.m$.
- (Part 2). Run/test your function on data in the file *resampling-data.txt*. More specifically, run your $kfold\_crossvalidation$ function on all data in the file by setting k (number of folds) to 10 and by varying the test fold index (parameter m) from 1, to 10. For each test data (generated for the different value of m) that were returned by your function calculate the mean and std and report them.

Problem 3. Function derivatives

Machine learning as a field builds upon knowledge of math, statistics, control and decision theories. In many cases, the learning process is formulated as an optimization problem with some objective function, say, $\min_\theta f(\theta)$, where θ are parameters we want to optimize. If this function is differentiable, the optimum (either local or global) must satisfy:

$$\frac{d}{d\theta} f(\theta^*) = 0$$

. In this problem we practice the calculation of function derivatives. Please derive:

- (a)
$$\frac{d}{dx}(2x)$$
- (b)
$$\frac{d}{dx}\left(x + 2x^3\right)$$
- (c)
$$\frac{d}{dx}(e^x)$$
- (d)
$$\frac{d}{dx}\left(\sin(x^2)\right)$$
- (e)
$$\frac{d}{dx}\left(\frac{1}{x}\right)$$

- (f)
$$\frac{d}{dx}\left(\frac{1}{x+x^2}\right)$$
- (g)
$$\frac{d}{dx}\left(\ln x^5\right)$$
- (h)
$$\frac{d}{dx}\left(5\right)$$
- (j)
$$\frac{d}{dx}\left(\ln \prod_{i=1}^{n} x^i\right)$$

Problem 4. Probabilities

Part a. Assume you have 2 fair dice. What are the probabilities associated with the different outcomes that are obtained by summing together the numbers on the two dice?

Part b. Calculate the expected value of the outcome for the 2 fair dice roll experiment.

Part c. Assume you play the two dice game from part a. 5 times. What is the probability, we never see the outcome of 4? What is the probability we see even (even number after the sum) outcome in all 5 trials.

Part d. Gaussian distribution is used commonly to model random processes with real-valued outcomes. Use matlab to generate and plot in one figure the probability density functions for:

- Gaussian with the mean 0 and standard deviation 1.
- Guassian with the mean 0.5 and standard deviation 1.
- Gaussian with the mean 0 and standard deviation 2.

Use observed probability density function to explain how the different parameters influence the position and the shape of the function.