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

Handout 2
February 4, 2021

Problem assignment 2

Due: Thursday, February 11, 2021

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. Compare them to the true mean and std above.
- (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 1000 mean values using 20 bins.
- (Part 4) Include the histogram in your report. Analyze the means calculated on 1000 subsamples of size 25 and compare them to the true mean and the mean that was calculated in step 1 on all examples in the dataset. Report your observations.
- (Part 5) Repeat steps from part 3 but now generate 1000 subsamples of size 40. Include the histogram in the report and compare it to the histogram generated in part 4 for subsamples of size 25, and to the mean of the original data. What are the differences? What conclusions can you make by comparing the means for subsamples of size 25 and 40.
- (Part 6) Take first 25 examples from the original data in the *mean_study_data.txt* and calculate their mean. Use the function t-test in Matlab to calculate and report the 0.95 confidence interval for the mean estimate. Does the true mean value fall into the 0.95 confidence interval?

Problem 2. k-fold cross-validation

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) 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. 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 odd-sum outcomes in all 5 trials.

Problem 4. Probabilities: Bayes theorem

A pharmaceutical company has developed a nearly accurate test for the disease A. The accuracy of the test is 99%, that is, with probability 0.99 it gives the correct result (the same probability for disease-positive-test and no-disease-negative-testcombinations are assumed) and only in 1% of tested cases (probability 0.01) the result is wrong. The incidence of the disease in the population is 0.01% (probability 0.0001). Compute the probability that

somebody from wide population who has tested positive indeed suffers from the disease. Would you recommend the test to be widely adopted?

Problem 5. Uniform distribution

Assume a uniform distribution $p(x|a,b) = \frac{1}{b-a}$ where $x \in [a,b]$.

- (a) Show that the distibution is properly normalized (that is, integral over its possible values equals 1)
- (b) Derive the mean of the distribution.

Problem 6. Bernoulli trials

Assume we have conducted a coin toss experiment with 100 coin flips. The results of the experiment are in file 'coin.txt' where 1 means a head and 0 means a tail. Assume that θ represents the probability of observing a head.

- (a) What is the ML estimate of θ?
- (b) Assume the prior on θ is defined by a Beta distribution $Beta(\theta|1,1)$. Plot and report both the prior and the posterior distributions on θ.
- (c) Calculate and report the MAP estimate of θ based on the posterior from part b. Show (plot) both the MAP estimate, and the expected value of θ on the plot of the posterior of θ you have generated in part b.
- (d) Repeat parts b and c by assuming that the prior on θ follows $Beta(\theta|4,2)$.