University of Pittsburgh
CS1675 Intro to Machine Learning
Professor Milos Hauskrecht

Handout 8
March 7, 2019

Problem assignment 7

Due: Thursday, March 21, 2019

In this assignment we continue our investigation of the "Pima" dataset. As in the previous assignment, you can download the dataset (*pima.txt*) and its description (*pima_desc.txt*) from the course web page. In addition to the complete dataset *pima.txt*, you have *pima_train.txt* and *pima_test.txt* you will need to use for training and testing purposes. The dataset has been obtained from the UC Irvine machine learning repository:
$http://www1.ics.uci.edu/\sim mlearn/MLRepository.html$.

Problem 1. Decision trees

The decision tree approach is yet another classification method we covered in the course. The method builds a tree by recursively splitting the training set using one of the attributes by optimizing the gain with respect to some impurity measure.

- Part a. The script *run_DT.m* shows how to train, display, and apply the decision tree in Matlab. The script first builds a default tree with minimal restrictions on its size, and after that the tree obtained by restricting the number of nodes in the tree. Please run and familiarize yourself with the code. What do you think, which tree is better for prediction, the unrestricted or restricted tree? Why? Should we always try to backprune it?
- Part b. Experiment with the decision tree function fitctree.m and its optional parameters, modifying the algorithm and the tree built. Report the results of your investigations in the report by listing the settings used for the tree learning algorithm and obtained results. You can find the different settings in the matlab help documents.

Problem 2. 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 3. Bayesian belief networks: foundations

The Bayesian belief networks framework relies on conditional independence to represent complex joint probability distributions. We say that X and Y are conditionally independent given Z, if the following two expressions hold:

1. $P(X, Y|Z) = P(X|Z)P(Y|Z)$
2. $P(X|Y, Z) = P(X|Z)$.

Show that 1 implies 2 and vice versa.