CSCI 5521: Intro to Machine Learning (Spring 2015)

Homework 1 Due 2/15/15 11:59 PM CDT

1. (**30 points**) Let $\mathcal{H}$ be the hypothesis space of circles with center $c \in \mathbb{R}^2$ and radius r, and $d_{\mathcal{H}}$ be the VC-dimension of $\mathcal{H}$. The prediction functions in $\mathcal{H}$ are of the form $\text{sign}(\|x - c\| - r)$. Derive $d_{\mathcal{H}}$ and prove your answer.

2. (**30 points**) Let $P(x|C)$ denote a Bernoulli density function for a class $C \in \{C_1, C_2\}$ and $P(C)$ denote the prior,

   (a) Given the priors $P(C_1)$ and $P(C_2)$, and the Bernoulli densities specified by $p_1 \equiv p(x = 1|C_1)$ and $p_2 \equiv p(x = 1|C_2)$, derive the classification rules for classifying a sample x into C_1 and C_2 based on the posteriors $P(C_1|x)$ and $P(C_2|x)$. (Hint: give rules for classifying $x = 0$ and $x = 1$.)

   (b) Consider D-dimensional independent Bernoulli densities specified by $p_{ij} \equiv p(x_j = 1|C_i)$ for $i = 1, 2$ and $j = 1, 2, \ldots, D$. Derive the classification rules for classifying a sample x into C_1 and C_2. It is sufficient to give your rule as a function of x.

   (c) Follow the definition in 2(b) and assume $D = 2$, $p_{11} = 0.7$, $p_{12} = 0.2$, $p_{21} = 0.4$, and $p_{22} = 0.8$. For three different priors ($P(C_1) = 0.1, 0.5, 0.9$ and $P(C_2) = 1 - P(C_1)$), calculate the posterior probabilities $P(C_1|x)$ and $P(C_2|x)$. (Hint: Calculate the probabilities for all possible samples $(x_1, x_2) \in \{(0, 0), (0, 1), (1, 0), (1, 1)\}$).

3. (**40 points**) Using your answers from Question 2 and the provided training, validation, and test datasets, write a Matlab program to calculate the maximum likelihood estimation on the training set. With the learned Bernoulli distributions and different priors ($P(C_1) = 0.1, 0.2, ..., 0.9$ and $P(C_2) = 1 - P(C_1)$) to classify the samples in the validation set using your classification rules. Finally, choose the prior that gives the lowest error rate on the validation set to classify the samples in the test set. Print to the terminal (stdout) a table of error rates for all the priors on the validation set and the error rate on the test set.

Programming Questions

All programming questions must be written in Matlab, no other programming languages will be accepted. The code must be able to be executed from the terminal command prompt on the cselabs machines. Each function must take the inputs in the order specified and display the textual output via the terminal. For each part, you can submit additional files/functions

(as needed) which will be used by the main functions specified below. Put comments in your code so that one can follow the key parts and steps. **Please follow the rules strictly. If we cannot run your code, you will receive no credit.**

- **Question 3:**
  - Training function in Bayes_learning.m: *Bayes_Learning*(training_data: a 800x100 matrix of training data, validation_data: a 100x100 matrix of validation data). The function returns the outputs (p1: learned Bernoulli parameters of the first class, p2: learned Bernoulli parameters of the second class, pc1: best prior of the first class, pc2: best prior of the second class). It must also print to the terminal (stdout) a table of error rates of all priors.
  - Testing function in Bayes_testing.m: *Bayes_Testing*(test_data: a 100x100 matrix of test data, p1: the learned Bernoulli paramter of the first class, p2: the learned Bernoulli paramter of the second class, pc1: the learned prior of the first class, pc2: the learned prior of the second class). The function must print to the terminal (stdout) the error rate on the test dataset.
- **Error rate:** Error rate is the percentage of wrongly classified data points divided by the total number of classified data points.

Report

Solutions to Questions 1 and 2 must be included in a report. The table of error rates on the validation set and the error rate on the test set for Question 3 must also be included in the report.

Instructions

- **Things to submit:**
  1. hw1_sol.pdf: A document which contains the report with solutions to all questions.
  2. Bayes_Learning.m and Bayes_Testing.m: Code for Question 3.
  3. Any other files, except the data, which are necessary for your code.
- **Submit**: All material must be submitted electronically via Moodle.