CSE 5523: Problem Set 2

Due via Carmen dropbox at 11:59 PM 9/27.

1. A probabilistic fraud detector produces the following recall (R) and precision (P) at different probability thresholds τ:

τ	R	P
.9	.2	.9
.5	.6	.6
.1	.9	.2

   The system is used to flag claims for fraud investigation. Previous studies have shown that 10% of all claims are actually fraudulent.

   (a) [9 pts.] If each true or false positive costs $100 to investigate and each false negative costs $1000 in fraudulent payments, what will be the estimated per-claim cost at each threshold τ?

   (b) [1 pt.] Which threshold τ should be used?

2. Use Kullback-Leibler divergence to evaluate the effectiveness of:

   (a) [5 pts.] a uniform distribution over {apple,pear,kiwi} as a predictor of data distributed according to the below distribution, and

   (b) [5 pts.] the below distribution as a predictor of data distributed according to the uniform distribution over {apple,pear,kiwi}.

apple	pear	kiwi
.5	.4	.1

3. PROGRAMMING:

   (In general for your programming problems you should hand in the following:

   - a copy of each program file you write,
   - a representative sample of each input file you use,
   - a representative sample of each output you produce.

   Your programs should be as short as possible.)

   (a) [5 pts.] Write a program called 'score.py' that takes the following as command-line arguments in the following order:

      i. a filename of a csv file containing true responses for some classifier task in one column with any column name, and

      ii. a filename of a csv file containing estimated responses for that same classifier task in one column with any column name,

Your program should output a csv file containing one column called 'score' containing a one for each item if the item is estimated correctly and a zero if it is not. Run the decision tree classifier from the lecture notes on the 'investor-train.csv' data and 'investor-test.csv' data in the lecture notes and score the results.

(b) [10 pts.] Write a program called 'overeighty.py' that that takes the following command-line argument:

i. a filename of an input csv file with a single column called 'score' containing ones (for correct responses) and zeros (for errors).

Your program should internally estimate a distribution over the underlying accuracy of a model that produces those scores, then output the probability that the model that produced these scores is really underlyingly at least 80% accurate. You may base your program on the Bayesian binomial test program in the lecture notes.

(c) [5 pts.] Use your program to evaluate the decision tree classifier in the lecture notes. Based on the investor data, what is the probability that the decision tree is over 80% accurate?

4. PROGRAMMING:

(a) [15 pts.] Write a program called 'numerical-decision-tree.py' that takes the following as command-line arguments in the following order:

i. a filename of a training csv file, containing data in columns for one hidden variable followed by any number of observed variables, delimited by commas, and

ii. a filename of a test csv file, containing data in columns for these same observed variables, delimited by commas, with headers matching the relevant columns of the training file.

Your program should perform decision tree training on the data in the training file and print out an estimate for the hidden variable of each test item in a single-column csv file with the same column name, ***except*** that it should treat the 'birthyr' variable as numerical rather than categorical, and therefore split items ***above or below*** some value (rather than ***at or not at*** some value). You may base your program on the decision tree program in the lecture notes, and you may retain the depth limits in the sample code of that program.

(b) [5 pts.] Compare your modified decision tree learner to the one from lecture notes 5 using the binomial significance test code from lecture notes 4. What is the probability that your new learner is better?