CS 760: Machine Learning © Fall 2020

Homework 4: Decision Trees

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN DUE 10/29/2020

GO GREEN. AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

In this homework you will use decision trees to determine whether you would have survived the Titanic sinking, and compare your results to those obtained with logistic regression. To find out, we will use the titanic dataset (`titanic_data.csv`), containing the following information about 887 passengers: 1) whether they survived or not (1 = survived, 0 = deceased), 2) passenger class, 3) gender (0 = male, 1 = female), 4) age, 5) number of siblings/spouses aboard, 6) number of parents/children aboard, and 7) fare:

	Passenger 1	Passenger 2	Passenger 3	$\cdots$	Passenger 887
Survived	**0**	**1**	**1**	$\cdots$	**0**
Passenger Class	3	1	3	$\cdots$	3
Gender	0	1	1	$\cdots$	0
Age	22	38	26	$\cdots$	32
Siblings/Spouses	1	1	0	$\cdots$	0
Parents/Children	0	0	0	$\cdots$	0
Fare	7.25	71.2833	7.925	$\cdots$	7.75

Our goal is to construct a decision tree that determines/predicts whether an individual would survive or not. Each subproblem is worth 10 points.

Problem 4.1. To make things easier, transform each of your features into a binary variable. Describe the transformation that you will use for each feature, and explain your reasoning.

Problem 4.2. Given vectors $\mathbf{x}_j, \mathbf{y} \in \{0,1\}^N$ containing the j^{th} feature and the response of N i.i.d. samples, write your own code to estimate the mutual information $I(x_j, y)$. Submit your code in an appendix.

Problem 4.3. Given vectors $\mathbf{x}_1, \ldots, \mathbf{x}_D, \mathbf{y} \in \{0,1\}^N$ containing information about D features, and the response $\mathbf{y}$ of N i.i.d. samples, use your mutual information code above to write your own code to build a decision tree. Submit your code in an appendix. What stopping criteria did you use?

Problem 4.4. Use your code above to build a decision tree for the titanic dataset. Display the tree you obtained.

Problem 4.5. Write your own code to perform 10-fold cross-validation to assess the accuracy of your tree. Submit your code in an appendix. What accuracy did you obtain?

Problem 4.6. Build your own feature vector $\mathbf{x}$. According to this and your decision tree, would you have survived the Titanic sinking?

Problem 4.7. Write your own code to build a random forest with 5 decision trees, using a random subset of 80% of the samples for each tree.

(a) Display the 5 trees you obtained.

(b) Use your code above to perform 10-fold cross-validation for your random forest. What accuracy do you obtain?

(c) According to your random forest, would you have survived the Titanic sinking?

Problem 4.8. Write your own code to build a random forest with 6 decision trees, each excluding one of the features in your dataset.

(a) Display the 6 trees you obtained.

(b) Use your code above to perform 10-fold cross-validation for your random forest. What accuracy do you obtain?

(c) According to your random forest, would you have survived the Titanic sinking?

Problem 4.9. Do all your predictions agree with each other, and with your prediction using logistic regression? Which method would you prefer to use, and why?