CS 760: Machine Learning © Spring 2024

Homework 5: Nearest Neighbors & Naive Bayes

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN DUE 03/13/2024

DO NOT POLLUTE! AVOID PRINTING, OR PRINT 2-SIDED MULTIPAGE.

In this homework you will use nearest neighbors and naive bayes to determine whether you would have survived the Titanic sinking, and compare your results to those obtained with previous algorithms. 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	···	Passenger 887
Survived	**0**	**1**	**1**	···	**0**
Passenger Class	3	1	3	···	3
Gender	0	1	1	···	0
Age	22	38	26	···	32
Siblings/Spouses	1	1	0	···	0
Parents/Children	0	0	0	···	0
Fare	7.25	71.2833	7.925	···	7.75

Each subproblem is worth 10 points.

Problem 5.1. Nearest Neighbors

(a) Write your own code to implement your favorite variant of K-nearest neighbors that seems appropriate to predict whether you would have survived the titanic sinking or not. Explain your choice/reasoning, and submit your code in an appendix.

(b) What measure of distance did you use, and why do you think this is this a good idea?

(c) Build your own feature vector $\mathbf{x}$. For $\mathrm{K} = 1, 2, \ldots, \mathrm{N}$, would you have survived the titanic sinking? Describe your results with a plot.

(d) In light of this, what do you think would be the best choice of K, and why?

(e) Describe how you could assess the reliability (confidence) of your results?

Problem 5.2. Naive Bayes

(a) Write your own code to implement Naive Bayes. Submit your code in an appendix.

(b) How did you model each variable (e.g., Bernoulli, Multinomial, Gaussian), and why?

(c) Build your own feature vector $\mathbf{x}$. According to your Naive Bayes classifier, would you have survived the titanic sinking?

(d) Describe how you could assess the reliability (confidence) of your results?

Problem 5.3. Out of all the methods that you have used so far for this dataset, which would you prefer, and why?

Problem 5.4. In Example 10.4 in the Lecture Notes, would you classify the new email as *spam*, or *ham*?

Problem 5.5. In Example 10.5 in the Lecture Notes, would you conclude that the killer is male or female?