CS 760: Machine Learning © Fall 2020

Homework 3: Logistic Regression

INSTRUCTOR: DANIEL L. PIMENTEL-ALARCÓN DUE 10/22/2020

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

In this homework you will use logistic regression to determine whether you would have survived the Titanic sinking. 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 classifier that determines/predicts whether an individual would survive or not. Each subproblem is worth 15 points.

Problem 3.1. Write your own gradient ascent code to maximize the log-likelihood function:

$$\ell(\boldsymbol{\theta}) \;:=\; \sum_{i=1}^{N} \log \left[\left(\frac{1}{1+e^{-\boldsymbol{\theta}^\top \mathbf{x}_i}} \right)^{y_i} \left(\frac{1}{1+e^{\boldsymbol{\theta}^\top \mathbf{x}_i}} \right)^{1-y_i} \right].$$

Submit your code in an appendix. Use your code to identify the maximizer $\hat{\boldsymbol{\theta}}$ for the Titanic dataset.

(a) What step size value/strategy did you use?

(b) How long did it take your computer to converge?

(c) What value did you obtain for $\hat{\boldsymbol{\theta}}$?

(d) What value did you obtain for $\ell(\hat{\boldsymbol{\theta}})$?

(e) Derive the asymptotic distribution of $\hat{\boldsymbol{\theta}}$.

Problem 3.2. Given a new sample with feature vector $\mathbf{x}$,

(a) Derive the MLE of the log-odds $\hat{\omega}$?

(b) Derive the asymptotic distribution of $\hat{\omega}$?

Problem 3.3. Build your own feature vector $\mathbf{x}$. According to this, and the results you obtained above

(a) Would you have survived the Titanic sinking? Briefly explain your conclusion.

(b) Derive a 95% confidence interval for your log-odds? (In other words, find τ for $\alpha = 0.05$)

(c) Based on this, would you say that your answer from (a) is fairly certain, or it could have gone either way? Briefly explain.

Problem 3.4. Based on the asymptotic distribution of $\hat{\boldsymbol{\theta}}$,

(a) Derive a test with significance level $\alpha = 0.05$ to determine whether a feature is significant.

(b) Using your test, which features do you conclude are significant?

(c) If you changed the most significant feature in your feature vector, would your survival prediction change?