Week 5 Exercise Questions: Maximum Likelihood Estimation and Logistic Regression

March 16, 2023

1. (2021 AI2 Exam question) As a data scientist in a telecommunication company, your task is to analyse a customer dataset to predict whether a customer will terminate his/her contract. The dataset consists of around 8000 customer records, each consisting of one binary dependent variable Y, indicating whether the customer terminates the contract ($Y = 1$) or not ($Y = 0$), and 19 independent variables, which include the customer's information, e.g., age, subscription plan, extra data plan, etc., and the consumer behaviour such as average numbers of calls and hours per week. Since your boss needs some actionable insights to retain customers, you decided to use interpretable machine learning methods. Design your interpretable machine learning method by answering the following questions:

   (a) You have implemented a feature selection algorithm based on mutual information to select the most informative features from the 19 independent variables. To validate the implementation of your mutual information calculation function, you use a small subset of the data to calculate mutual information manually. You select one independent variable 'subscription plan', denoted as S, which takes two values, $S \in \{1, 2\}$. Please use the following Probability Mass Function table

$p(S, Y)$	$S = 1$	$S = 2$
$Y = 0$	$\frac{2}{12}$	$\frac{5}{12}$
$Y = 1$	$\frac{2}{12}$	$\frac{3}{12}$

   to calculate

   - Entropies $H(S)$ and $H(Y)$
   - Conditional entropies $H(S|Y)$ and $H(Y|S)$
   - Joint entropy $H(S, Y)$
   - Mutual information $I(S; Y)$

   Show all your working. Discuss what mutual information means and whether this feature will be selected or not. (**6 marks**)

   (b) After applying your algorithm you selected two variables: 1) extra data plan E, which is a binary random variable that indicates whether the customer subscribes to the extra data plan ($E = 1$) or not ($E = 0$); and 2) averaged hours used per week H, which is a continuous random variable. You then built a logistic regression model

to classify customers into 'low risk' or 'high risk' of terminating the contract. The fitted model is

$$\log\left(\frac{p}{1-p}\right) = -0.77 + 0.23H - 1.18E$$

- Given a customer who has the extra data plan ($E = 1$) and spent on average 0.5 hours per week, calculate the odds and the probability the customer will terminate the contract ($Y = 1$). (**4 marks**)
- Using this fitted model, explain to your boss what actions should be taken to retain customers. (**10 marks**)

2. We flipped a coin 100 times. Given that there were 55 heads, use the maximum likelihood estimation to find for the probability p of heads on a single toss.

3. Let X be independent and identically distributed (i.i.d.) Poisson (λ) distributed. Find the maximum likelihood estimator for λ, $\hat{\lambda}$. Calculate an estimate using this estimator when, $x_1 = 1$, $x_2 = 2$, $x_3 = 5$, $x_4 = 3$.