CS412 Machine Learning, Spring 2024: Homework 2

March 21, 2024

Instructions

- Please submit your code as a notebook and a PDF. Name your submission as CS412-HW2-YourName.pdf and CS412-HW2-YourName-Code.ipynb, where you substitute your first and last names into the file-names in place of 'YourName'.
- You are required to code in Python. In preparing your notebook, please ensure that your code is well-organized and not confined to a single cell. Use separate cells for different sections of the code. This organization will help in making your notebook more readable.
- The responsible TA for this homework is Atra Zeynep Bahçeci. But first ask your questions on the Homework Forum.
- If you are submitting the homework late (see the late submission policy in the syllabus), we will grade your homework based on the time stamp of submission and your remaining late days.

1 MLE and MAP [40 pts]

You are part of a research team that focuses on a rare species of butterfly, whose pre-rain flight patterns have been observed to correlate with rainfall amounts. The rate at which the butterfly flaps its wings, *flutterness*, affects the rainfall amounts. You have developed a statistical model, called the *Flutter Distribution*, to encapsulate this phenomenon. This model leverages parameter, α, to quantify the observed flutterness, x. The Flutter Distribution is delineated by the probability density function (PDF) where $x \in [0, 1]$ and $\alpha > 0$:

$$p(x|\alpha) = \alpha(1-x)^{\alpha-1}$$

1.a [20 pts] Based on a dataset $X = \{x_1, x_2, \ldots, x_n\}$ of flutterness measurements, determine is the Maximum Likelihood Estimate (MLE) the parameter α.

1.b [20 pts] Your team determined the prior of α as $p(\alpha) = \lambda\alpha^{\lambda-1}e^{-\lambda\alpha}$ where $\lambda > 0$. Find the Maximum a Posteriori (MAP) estimation of the parameter α based on $p(\alpha)$.
Show all the steps of your calculations for full points and partial grading. You can type your answers (LaTeX, MathType, MS Word Equations, etc.), or add your hand-written answers as images to the pdf. For hand-written solutions, make sure your handwriting is legible, else **your answer will not be graded**.

2 Probability and Bayes' Theorem: Kaggle ML & DS Survey [60 pts]

Dataset: 2017 Kaggle Machine Learning (ML) & Data Science (DS) Survey comprises over 16,000 responses from Kaggle's industry-wide survey on the state of data science and machine learning. The dataset, provided to you as cs412-hw2-data.zip, has 5 files:

- `schema.csv`: a CSV file with survey schema. This schema includes the questions that correspond to each column name in both the `multipleChoiceResponses.csv` and `freeformResponses.csv`.

- `multipleChoiceResponses.csv`: Respondents' answers to multiple choice and ranking questions. These are non-randomized and thus a single row does correspond to all of a single user's answers.
- `freeformResponses.csv`: Respondents' freeform answers to Kaggle's survey questions. These responses are randomized within a column, so that reading across a single row does not give a single user's answers.
- `conversionRates.csv`: Currency conversion rates (to USD).
- `RespondentTypeREADME.txt`: This is a schema for decoding the responses in the "Asked" column of the `schema.csv` file.

Using the `.csv` files from cs412-hw2-data.zip, answer the following questions on Jupyter Notebook or Google Colab. **You do not need to add your answers to the pdf**. Use a separate cell for your code/answer for each question. You do not need to use any machine learning libraries, once you have loaded the data you can answer all questions via Pandas and/or NumPy methods.

2.a [5 pts] What is the probability that a respondent is currently employed as a `Programmer` given they use `C/C++` at work?

2.b [5 pts] What is the probability that a respondent is a `Data Scientist` given they have majored in `computer science`, `mathematics or statistics`?

2.c [5 pts] What is the probability that a respondent works in the `Technology` industry given that they earn more than `40,000 USD` annually?

2.d [5 pts] What is the joint probability of a respondent being `over 30 years` old and having a at least a `Bachelor's degree`?

2.e [5 pts] What is the probability that a respondent is a `Data Scientist` who majored in *Computer Science*, `Mathematics or statistics`?

2.f [5 pts] What is the joint probability that a respondent is from `France`, earns less than `100,000 USD` annually, and uses `Cross-Validation` `Often` or `Most of the time`?

2.g [5 pts] What is the probability that a respondent uses `C/C++` at work given that they are employed as a `Programmer`? (Hint: Use your findings from Question 2a).

2.h [10 pts] Given the probability of a respondent wearing glasses is 0.15, and the probability of a respondent wearing glasses given they have a PhD is 0.25, find the probability of a respondent having a PhD given that they wear glasses.

2.i [15 pts] Do you think that the probabilities given in Question 2h (probability of a respondent wearing glasses, probability of a respondent wearing glasses given they have a PhD) are correct? Why / why not? Prove your point using the dataset. (Hint: Use the Bayes' Theorem).