APPLIED MATH 115 HOMEWORK 1

This homework covers material in week one and two of the class. You are allowed to work in groups of up to three people from the class on this homework – every individual should submit their homework but please list your group members, and write a summary of 'author contributions' to the homework that everyone in the group agrees upon.

The learning goals of homework is several-fold:

(1) Understand the mathematics underlying maximum likelihood approximation.
(2) Apply this to analyzing both synthetic and real data for a series of games like the world series, estimating the probability that the best team win.
(3) Critically examine variants of Mosteller's model, probing how the model could be improved both for baseball
(4) Think about how to fit the probability that the best team wins given data on the team from the current season. This builds and completes work you started in section.
(5) Understanding how to extend this model to calculate probable outcomes of tournaments. Do this with synthetic data for a tournament that we provide.

Problem 1: *Maximum Likelihood Estimation*

Suppose the World Series were best of 3 games. Assume like Mostellar that each game is a Bernoulli trial with probability p. Assume that we see n_0 events where the better team wins 0 games, n_1 events where the better team wins 1 game and n_2 events where the better team wins 2 games.

(1) Following the derivation in class, derive the Likelihood of our observations.
(2) Analytically compute the maximum likelihood for p, again following the derivation in class.
(3) Attached to this homework is data that we have generated for 44 such *synthetic* competitions. Each competition reports the number of games the losing team wins. From this data, use your derivation for MLE to compute your estimate for p.
(4) Additionally, following the code we used in class and in section, carry out a numerical calculation of the maximum likelihood estimate for p and compare to the analytical derivation.

Date: January 23, 2024.

Problem 2 *Variants on Mostellar*

(1) Run the 44 game world series and plot the distribution of times that the losing team wins using Mosteller's number $p = 0.65$. Now do the same thing and run it 4400 times. Does the fraction of games that the losing team won change significantly by changing the number of Series played? Were Mosteller's measurements for the first fifty years of the twentieth century lucky?
(2) In 1976, baseball players were given the right to become free agents, which meant that they could switch teams much more easily. Analyze the probability that the better team wins the world series *after* 1976. If it does not hold, please formulate a hypothesis of what a new model might be. Note: We have provided a csv file of the win/loss for *all* world series on the course web site and github.
(3) In the paper, Mosteller also discussed a model in which "p" changed from year to year. Modify the code from class to see if this makes much of a difference. A simple way to do this is to add a small random number to p so that it changes from game to game.
(4) Identify some reason that you are angry at Mosteller's Model. Think about how you would change the model to assuage your anger.

Problem 3: *Adding more features to compute p*

In Problem 1, you estimated the probability p of the better team winning the series. Now, let's explore how additional features in the data could enhance the accuracy of p prediction. In the lecture 2 notebook, we introduced a dataset from Major League Baseball that includes a set of features about teams that go beyond games won and lost, including strikeouts, double plays and so forth. The goal of this problem is to create a simple model that predicts the probability a team wins the world series.

(1) Explain what the code in the notebook does. Include in your explanation the following concepts:
   (a) What does the **dropna** command do in the line transforming the features?
   (b) What is the point of the dataset split into train and text?
   (c) What is the logistic regression fit predicting?
   (d) Explain accuracy, confusion matrix,precision and recall?

(2) Create and explain the feature importance plot, identifying the features that influence the probability of the better team winning. Does this make sense?

(3) Invent new features and find out if they are more informative for predictions. [Hint: think about some of the stats you hear about when people discuss baseball, e.g. batting average, winning streaks, run differentials....]. Explain why you've chosen these features and transformations and see if the turn out to be important.

(4) How could you use this type of analysis to improve Mostellar's model? *Note: Just propose ideas, you do not need to carry them out*

Problem 4: *Simulating a tournament*

Consider a tournament with 4 teams. You are given a draw. you are also given the previous games that the four teams have played against each other.

(1) From the previous games, estimate p_{ij}, the probability that team i will beat team j in a given game.

(2) Given your estimates of p_{ij}, carry out 1000 simulations of the draw. Determine: which team is most likely to win the tournament, and how likely the worst team is to win the tournament.

(3) Predicting sports draws is a serious business. As one example the website fivethirtyeight.com always has a forecast of March Madness, the NCAA Division 1 basketball draws. They describe their methodology here. **do something with this**

Problem 5 *Apply these ideas to a tennis match*
Use the mathematical formalism that we have introduced to write out a series of steps to create a mathematical model for a tennis match. *you are only required to write out the series of steps, not implement the steps.*

(1) Make a model for an individual game, where there is a probability that a player wins a point is a parameter. Note that whoever serves typically has a higher probability of winning a point. Note that there is a serious complication over the world series, which is that a player must win by two points. How can you deal with this?
(2) Given the probability that a player wins a game, turn this into the probability that they will win a set. Assume tiebreakers are not allowed, and just assume that (like in the case of the game) that a player must win by two games.
(3) A match is 2 out of 3 sets.

Extending these ideas and fitting them to actual data would be a great first mini project