Bayesian Analysis Lab

Dataset

ab_data.csv, mimic3d.csv, patients.csv

Due Date: October 28, 23:59:59 EST.

Problem Description

Bayesian methods are very useful when you have specific prior information and/or when you need to quantify the uncertainty in your estimates. In this lab we'll get into some situations where probabilities are useful outputs from a model.

The datasets come from the following sources:

- https://www.kaggle.com/zhangluyuan/ab-testing
- https://www.kaggle.com/drscarlat/predict-hospital-length-of-stay-classification/data (information no longer available)

Tasks:

1. Suppose the marketing people are testing a new web page, with the hope of increasing the conversion rate (proportion of visitors who sign up or take some other action). We'll imagine that you collected the data in the ab_data.csv file which lists user visits with whether it was the new page or old page, and whether there was a conversion. Explore and visualize the data.

2. Instead of using p-values, do a Bayesian AB test. You can define and update independent priors on the old and new conversion rates and arrive at respective posterior distributions. Try a prior of Beta(alpha=2, beta=20) for the old rate, which represents what has been observed in the past.

3. Start with a subset of 100 data points and perform inference. Find the posterior probability that the new page has a higher conversion rate. Hint: use random samples from the independent posteriors to estimate the probability. Update the posteriors with another 100 data points. At what data size do the priors become irrelevant?

4. Sometimes we don't just want to estimate a dependent variable, we want a probability distribution for it. E.g. if your life expectancy is 80 years you might want to know whether it's 50-50 between 0 years and 160 years or another distribution. Load the second dataset which lists the length of stay in a hospital ("LOSdays") along with a number of other variables. Explore this second dataset.

5. Predict the length of the hospital stay for the patients in the dataset *patients.csv* by conducting a Bayesian linear regression analysis. What's the probability of staying longer than 2 days and therefore definitely missing work? You may want to use normal priors for simplicity.