Homework 4: Applied Probability and Statistics

University of Arizona CSC 380: Principles of Data Science

Homework due at 11:59pm on September 23

This assignment will strengthen your understanding of the fundamental concepts in applied probability and statistics. The questions in this assignment will build on lecture material as well as the assigned readings from (Wasserman, L. 2004. "All of Statistics").

Deliverables Submit your responses as a PDF along with any code as separate files. **Show all work along with answers.** This is for your benefit as incorrect answers may receive partial credit if the work demonstrates understanding.

Problem 1: Maximum Likelihood Estimation (2 points)

I would like to build a simple model to predict how many students are likely to come to my office hours this semester. Because this is an arrival process, I will model the number of arrivals during office hours as Poisson distributed. Recall that the Poisson is a discrete distribution over the number of arrivals (or events) in a fixed time-frame. The Poisson distribution has a probability mass function (PMF) of the form,

$$\text{Poisson}(x; \lambda) = \frac{1}{x!}\lambda^x e^{-\lambda}.$$

The parameter λ is the *rate* parameter, and represents the expected number of arrivals $\mathbb{E}[x] = \lambda$. To fit the model I will need to estimate the rate parameter using some data.

a) *During my last three office hours I received $X_1 = 10, X_2 = 11, X_3 = 8$ students. Write the logarithm of the joint probability distribution $\log p(X_1, X_2, X_3; \lambda)$.*

b) *Compute the maximum likelihood estimate (MLE) of the rate parameter λ^{MLE} which maximizes the joint probability in part (a). The model is concave and so the MLE can be computed by finding the zero-derivative solution. Make sure to show all of your calculations. How many arrivals should I expect at my next office hours under this model?*

c) *I have assigned a particularly challenging homework which has led to a lot of students $X_4 = 25$ arriving at my office hours. Compute the MLE again, but include this new training point. How has the model changed with this new data?*

Problem 2: Bayesian Model and MAP Estimation (4 points)

In the last problem we observed the impact of including an extreme outlier $X_4 = 25$ on the MLE rate estimate. Since I don't expect this to occur often, I could simply discard X_4. However, throwing out data is bad practice. A better approach is to model the prior belief on λ by defining a Bayesian model. We will use a Gamma prior with parameters $k = 29$ and $\theta = 1/3$. The Gamma PDF has the form,

$$\text{Gamma}(\lambda \mid k, \theta) = Z(k, \theta)\lambda^{k-1}e^{-\frac{\lambda}{\theta}}.$$

The term $Z(k, \theta)$ is a normalizing constant that does not depend on λ and can be ignored for our purposes.

a) *Write the logarithm of the joint probability* $\log p(X_1, X_2, X_3, X_4, \lambda \mid k, \theta)$*. You may ignore constant terms that do not depend on* λ*.*

b) *Computer the maximum a posteriori (MAP) estimate of the rate parameter* λ*. To do this, compute the zero-derivative of the log-probability in part (a). What is* λ^{MAP}*? In comparing this estimate to the previous MLE, what can you conclude about introducing a prior distribution on* λ*? For full credit make sure to show your calculation of the zero derivative.*

c) *We want to visualize the change from the prior to posterior. Gamma-Poisson is a conjugate pair, meaning that the posterior is a Gamma distribution with posterior,*

$$p(\lambda \mid x_1, \ldots, x_N) = Gamma\left(\lambda \middle| k + \sum_i x_i, \frac{\theta}{N\theta + 1}\right)$$

*Using **matplotlib.pyplot.plot** plot the prior and posterior together, on the same plot, by evaluating the Gamma PDF with appropriate parameters at each location. To do this, use **numpy.arange** or **numpy.linspace** define a vector of values from 0 to 30 densely spaced and evaluate the PDFs at each location. You can use **numpy.random.gamma** to evaluate your PDFs. Make sure to label your axes and include a legend to indicate which is the prior and which is the posterior. See the Matplotlib documentation at https://matplotlib.org/stable/api/_as_gen/matplotlib.pyplot.legend.html for instructions and examples on how to include a legend.*