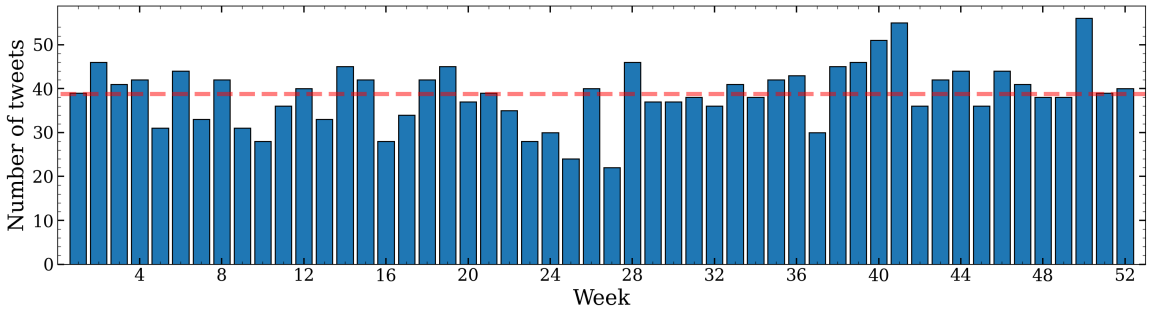
Homework 2

Course: Introduction to applied data science (PHYS247) Spring 2020
TA: Nima Chartab Due Date: April 27, 11:59 p.m.

Problem 1: Bayesian inference

In this problem, you are given a "tweet_counts.csv" file which includes Bob's weekly activity on twitter during the last year. The first and second columns are the number of week and weekly tweet counts, respectively. The figure below, visualizes the data where the count of tweets per week is shown as a function of week number. The horizontal red line shows the average number of weekly tweets over the last year.



a) Read tweet_counts.csv file and define two variables which represent the week number and weekly tweet counts. What is the average number of weekly tweets over the last year? This is the horizontal line, I have plotted in the figure above.

b) Use `matplotlib.pyplot.bar` to create the same figure shown above.

Do you think that Bob's tweeting habit has changed over time? We are going to answer this question in the Bayesian framework. Imagine that his tweeting habit suddenly changed at week S (W_S). Consider that the count of tweets can be modeled with Poisson distribution.

$$\mathrm{P(k{=}tweet\ count)}{=}\frac{\lambda^k}{k!}e^{-\lambda}$$

where λ is a constant that controls the shape of the distribution. λ shows the expected value for number of tweets. As Bob's tweeting pattern changed at W_S, so λ changed suddenly at that point.

$$\lambda = \begin{cases} \lambda_1 & \text{if} \quad W < W_S \\ \lambda_2 & \text{if} \quad W \geq W_S \end{cases}$$

Do you have any prior belief on λ_1 and λ_2? I think it is unlikely that someone (including Bob) posts very large number of tweets every week. So, my prior belief is that the probability of λ should decrease with the increase of λ. This belief as a prior can be modeled with an exponential distribution,

$$\mathrm{P}(\lambda){=}\alpha e^{-\alpha\lambda}$$

where α is another constant that emerges from our prior belief.

c) Prove explicitly that the expected value of λ is $1/\alpha$. Hint: $\mathbb{E}(\lambda) = \int_0^\infty \lambda P(\lambda) d\alpha$

d) Take the average value of λ from part "a" to estimate α. Plot the estimated distribution function ($\mathbf{P}(\lambda)$ vs λ).

Do we have any prior belief about W_S? Consider that we have no specific prior on this. So, let's take a uniform distribution as a prior for W_S. The normalized uniform prior can be written as

$$P(W_S)=\frac{1}{52}$$

Now we are ready to answer the main question of the homework. Before moving forward, one quick note: here we took the same prior for λ_1 and λ_2. One can consider different priors. For example, if we believe that sometime around week 30 the pattern changed, then we can define different values of α considering separate averages for λ, one for $W < 30$ and the other for $W \geq 30$. But, be aware that this will have a minimal effect on our final conclusion.
Our model has 3 parameters, λ_1, λ_2 and W_S. We are going to find the posterior of these three parameters to infer that how much we believe in the change in Bob's tweeting pattern and when is the most likely week for this change?

e) Use `numpy.linspace` to create two variables with 1×50 array in the interval of 25 and 50. These variables are defined as the model space that we want to search for posterior of $\lambda 1$ and $\lambda 2$. You have also defined an array in part a which shows the week number. This provides a space for W_S. Consider all these points as a 3-D mesh-grid, how can we find posterior for each point in the 3-D space given all the information you have? Use Bayes' theorem to elaborate your method in detail.

f) Write a code to find marginalized-posterior for $\lambda 1$, $\lambda 2$ and W_S. Plot posteriors for $\lambda 1$ and $\lambda 2$ in the same figure and create a bar plot for posterior of W_S in a separate figure. Running your code for this part may take a long time since you compute posterior for every single point in your model space. However, we will learn more efficient way, Markov chain Monte Carlo (MCMC), later in the next homework.

g) How is your belief updated about a sudden change in Bob's tweeting habit? Can you estimate the week when tweeting pattern changed? Use marginalized 2-D posteriors of $\lambda 1$ and $\lambda 2$ to obtain $P(\lambda_2 - \lambda_1 > 5)$. This shows the probability that Bob's weekly tweet counts has increased by five at some point.