CSE 544, Spring 2021: Probability and Statistics for Data Science

Assignment 6: Bayesian Inference and Regression Due: 05/06, 1:15pm, via Blackboard

(6 questions, 70 points total)

I/We understand and agree to the following:

(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.

(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.

(write down the name of all collaborating students on the line below)

---

1. Posterior for Normal **(Total 10 points)**

Let $X_1, X_2, \dots, X_n$ be distributed as Normal(θ, σ²), where σ is assumed to be known. You are also given that the prior for θ is Normal(a, b²).

(a) Show that the posterior of θ is Normal(x, y²), such that: (6 points)

$x = \frac{b^2\bar{X}+se^2a}{b^2+se^2}$ and $y^2 = \frac{b^2se^2}{b^2+se^2}$; where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $se^2 = {\sigma^2}/{n}$.

(Hint: less messier if you ignore the constants, but please justify why you can ignore them)

(b) Compute the (1-α) posterior interval for θ. (4 points)

2. Bayesian Inference in action (Total 15 points)

You will need the q2_sigma3.dat and q2_sigma100.dat files for this question; these files are on the class website. Each file contains 5 rows of 100 samples each. Refer back to Q 1 (a); you can use its result even if you have not solved that question. Submit all python code for this question with suitable filenames.

(a) Assume that σ = 3 (meaning $\sigma^2 = 9$). Let the prior be the standard Normal (mean 0, variance 1). Read in the 1st row of q2_sigma3.dat and compute the new posterior. Now, assuming this posterior is your new prior, read in the 2nd row of q2_sigma3.dat and compute the new posterior. Repeat till the 5th row. Please provide your steps here and draw a table with your estimates of the mean and variance of the posterior for all 5 steps (table should have 5 rows, 2 columns). Also plot each of the 5 posterior distributions on a single graph and attach this graph. What do you observe? (7 points)

(b) Now assume that σ = 100 and repeat part (a) above but with q2_sigma100.dat. Assume the same prior of a standard Normal. Provide the table and final graph. What do you observe? (7 points)

(c) Based on the comparison of answers of (a) and (b), what can you conclude? (1 point)

3. Regression Analysis (Total 7 points)

Assume Simple Linear Regression on *n* sample points (Y_1, X_1), (Y_2, X_2), …, (Y_n, X_n); that is, Y = $\beta_0 + \beta_1 X + \varepsilon_i$, where $E[\varepsilon_i] = 0$.

(a) Using the estimates of β derived in class, show that:

$\widehat{\beta_1} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$ and $\widehat{\beta_0} = \bar{Y} - \widehat{\beta_1}\bar{X}$, where $\bar{X} = (\sum_{i=1}^{n} X_i)/n$ and $\bar{Y} = (\sum_{i=1}^{n} Y_i)/n$. (2 points)

(b) Show that the above estimators, given X_is, are unbiased (Hint: Treat X's as constants) (5 points)

4. More on Regression and Time series analysis (Total 10 points)

US media says that the number of Americans ages 65 and older is projected to nearly double from 52 million in 2018 to 95 million by 2060. However, this is not sufficient to show that the US population is aging. To be precise, we also need to know how the total population will grow. In this question, you are going to use q4.csv (on course website) which contains the US population and the population of 65+ year olds from 1980 to 2019. Report all answers and figures in your submission; you do not need to submit any code.

(a) For US total population and 65+ years olds population, using simple linear regression (population vs. year, include β_0 term), plot the original data and the regression fit, and calculate the SSE respectively. Are they all suitable for linear regression? (Which is or which is not?) (4 points)

(b) Using the data from 1980 to 2018, predict the population of 65+ years old in 2060. Show the linear regression equation, prediction result, and SSE. Then, do the same thing but using data only from 2008 to 2018. Which prediction should you trust more? Is the media right with the 65+ population doubling? (4 points)

(c) If we want to predict the ratio of 65+ population, there are two ways to do this. One is to compute the ratio from 2008 to 2018 and then predict the result in 2019. The other is first predict the total population for 2019 and 65+ years old population for 2019 (can use result in (b)) respectively using data from 2008 to 2018, and then compute the ratio for 2019. Predict the ratio in both ways. Which way is more accurate for 2019? Why? (Hint: You made different assumptions in these two methods. Compare the linearity of 65+ population and the ratio.) (2 points)

5. Multiple Linear Regression (MLR) (Total 10 points)

The admission chance when students apply for the Masters program depends on many factors. In q5.csv (uploaded on the course webpage), there are several parameters included, 1) GRE score (out of 340), 2) TOEFL Score (out of 120), 3) University Rating (out of 5), 4) SOP (Statement of Purpose Strength, out of 5), 5) LOR (Letter of Recommendation Strength, out of 5), 6) GPA (Undergraduate GPA, out of 10), 7) Research (0 or 1). And the corresponding chance of admit ranges from 0 to 1. The dataset has 500 rows (samples) and 8 columns (chance of admit and 7 features). Submit your code as q5.py and show your answers in the pdf.

(a) Using MLR, find the linear relationship between chance of admit and the 7 features listed above (input are 7 factors while output is chance of admit). Do not include the intercept term β_0. Use the first 400 rows to train, and the last 100 rows to test. Report your linear equation, and the SSE of your test set. (3 points)

(b) Now we use less features, only TOEFL, SOP and LOR. Do not include the intercept term β_0. Do the same thing as (a). Report your linear equation, and the SSE of your test set. (3 points)

(c) Now we only use GRE and GPA. Do the same thing as (a). Do not include the intercept term β_0. Report your linear equation, and the SSE of your test set. (3 points)

(d) What are your observations based on the SSEs obtained for (a), (b), and (c)? (1 point)

6. Bayesian hypothesis testing (Total 18 points)

You are tired of studying probs and stats and have finally decided to give up your current life and turn to your one true passion – farming. Lucky for you, there is lot of farmland on Long Island, and you have your heart set on a particular farm that is available for purchase. However, you do not know whether the soil in the farm is good or not. Say the soil in the farm is a discrete random variable H and it can only take values in the set $\{0, 1\}$, where 0 represent good soil and 1 represents bad soil. We transform this as a hypothesis test as follows: $\boldsymbol{H_0}: H = 0$ and $\boldsymbol{H_1}: H = 1$. Let the prior probability $P(\boldsymbol{H_0}) = P(H = 0) = p$ and $P(\boldsymbol{H_1}) = P(H = 1) = 1 - p$. The water content in the soil depends upon the type of soil. If we assume water content to be a RV W, then $f_W(w|H = 0) = N(w; -\mu, \sigma^2)$ and $f_W(w|H = 1) = N(w; \mu, \sigma^2)$. To test which of the two hypotheses is correct, you take n samples of the soil from different patches of the farm and measure the water content metric of each sample; the resulting data sample set is $\boldsymbol{w} = \{w_1, w_2, w_3 \dots, w_n\}$. Assume that the samples are conditionally independent given the hypothesis/soil type.

(a) If we denote the hypothesis chosen as a RV C where $C \in \{0, 1\}$, then according to MAP (Maximum a posteriori), we have $C = \begin{cases} 0 & if\ P(H = 0|\boldsymbol{w}) \geq P(H = 1|\boldsymbol{w}) \\ 1 & otherwise \end{cases}$. This implies that the hypothesis H=0 is chosen (referring to C=0) when P(H=0|w) ≥ P(H=1|w). Derive a condition for choosing the hypothesis that soil in the farm is of type is 0, in terms of $p, \mu\ and\ \sigma$. (4 points)

(b) Write a python function **MAP_descision()** in a script named Q6_b.py, where your function takes as input (i) the list of observations $\boldsymbol{w}$, and (ii) the prior probability of $\boldsymbol{H_0}$, and returns the chosen hypothesis (value of C) according to the MAP criterion. Report the result for the 10 different instances of observations from the q6.csv dataset and for each prior probability p = [0.1, 0.3, 0.5, 0.8] for the value of (μ, σ^2) = (0.5, 1.0). Each column is one set of observations. (10 points)

Example output format:

```
For P(H_0) = 0.1, the hypotheses selected are :: 0 1 0 1 0 0 1 0 0 1
For P(H_0) = 0.3, the hypotheses selected are :: 1 1 0 1 1 0 0 0 0 1
For P(H_0) = 0.5, the hypotheses selected are :: 1 1 0 1 1 0 0 0 0 1
For P(H_0) = 0.8, the hypotheses selected are :: 1 1 0 1 1 0 0 0 0 1
```

(c) Denoting the hypothesis selected as a RV C where $C \in \{0, 1\}$, the average error probability via the MAP criterion is given by $\boldsymbol{AEP} = P(C = 0|H = 1)P(H = 1) + P(C = 1|H = 0)P(H = 0)$. Given the observations $\boldsymbol{w} = \{w_1, w_2, w_3 \dots, w_n\}$, derive $\boldsymbol{AEP}$ in terms of $\mu, \sigma, \Phi(\quad)\ and\ p$. (4 points)