CSE 544, Spring 2022: Probability and Statistics for Data Science

Assignment 6: Bayesian Inference and Regression Due: 04/29, 8:15pm, 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 $X_1, X_2, \ldots, X_n$ be distributed as Normal(θ, σ^2), where σ is assumed to be known. You are also given that the prior for θ is Normal(a, b^2).

(a) Show that the posterior of θ is Normal(x, y^2), such that: (6 points)

$x = \frac{b^2\bar{X}+se^2 a}{b^2+se^2}$ and $y^2 = \frac{b^2 se^2}{b^2+se^2}$; 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.dat file (on class website) for this question. The file contains 100 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 $\sigma = 5$ (meaning $\sigma^2 = 25$). Let the prior be the standard Normal (mean 0, variance 1). Read in the 1st row of q2.dat and compute the new posterior. Now, assuming this posterior is your new prior, read in the 2nd row of q2.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? (6 points)

(b) Repeat part (a) but assuming $\sigma = 20$; provide all steps, the table, etc. What is the difference between the outputs of (a) and (b)? What is the reason for this difference, in your opinion? (3 points)

(c) Repeat plotting the graph in part (a) assuming $\sigma = 5$ but this time use all the 100 rows. Do not draw the table but note down the mean and variance of the posterior after 100 steps. What trend do you observe in the shape of the plot as more rows are used? (3 points)

(d) Repeat part (c) but use the 1st row for all the 100 steps. What difference do you observe in mean and variance of (c) and (d)? What is the reason for this? (3 points)

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 (Total 10 points)

Around Feb 2020, governments worldwide started announcing lockdowns to restrict the spread of SARS-COV-2. Delivery companies like Amazon have gotten into the spotlight for becoming more valuable during the pandemic as people started utilizing their services more frequently. To verify this, let us have a look at the stock price of Amazon before and during the pandemic. The dataset q4.csv on the class website contains the stock price of the company between Jan 2012 and Dec 2021. Report all answers and figures in your submission. Submit your code as q4.py.code.

(a) Using the complete dataset, perform simple linear regression (time vs. stock price, including β_0 term), plot the original data and the regression fit, and report the SSE. (4 points)

(b) Using the data from 17 Jan 2012 to 31 Jan 2020, predict the stock price on 1st November 2021 using simple linear regression. How much is the percentage difference between the prediction and the real value? Show the prediction result and the SSE. Plot the original data and the regression fit. (2 points)

(c) Let us see how the stock price will increase in the future considering the same factors will continue for 10 more years. Using only the data after Sept 2020, predict the stock price on 1st January 2031 using simple linear regression. Show the prediction result and SSE. Plot the original data and the regression fit. Add the regression fit from (b) into the same plot. (2 points)

(d) Using the data from Jan 2012 to Jan 2020, predict the stock price on 1st January 2026 using simple linear regression. What is the percentage difference between the predictions in (c) and (d)? (1 point)

(e) What is your inference from the results of (b), (c), and (d)? (1 point)

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

The rent of a flat changes based on many factors. The q5.csv file (on the course webpage) contains a dataset of rents in Manhattan, including several parameters: (1) #bedrooms (integers 0-5), (2) #bathrooms (integers 0-5), and (3) square foot area (sqft) of the flat (large range of integers). Also, as the view from an apartment attracts more tenants, factors like (4) floor number of the apartment are also included (integers 0-83). Some apartments also include other amenities (0 or 1) which might also affect the rent (look for column names starting with "has"). The dataset has 3539 rows (rental records) and 18 columns (apartment attributes and facilities). Submit your code as q5.py and show your answers in the pdf.

(a) Using MLR, find the linear relationship between the rent and the first 3 parameters. Use 80% of the rows (randomly chosen) to train, and the remaining to test. Report your linear equation, and the SSE of your test set. (3 points)

(b) Now repeat (a) along with the 4th parameter. Report your linear equation, and the SSE of your test set. (2 points)

(c) Now repeat (b) along with the amenities. Report the SSE of your test set. (2 points)

(d) Choose 1 other column from the dataset which you think is also important and is not included in (a), (b), and (c) above. Repeat (c) along with this new column. Report your linear equation, and the SSE of your test set. Explain why you selected this very column. (1 point)

(e) What are your observations based on the SSEs obtained for (a), (b), (c) and (d)? (2 points)

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 $\mu, \sigma, \Phi(\ \ )\ and\ p$. (4 points)