CSE 544, Spring 2020: Probability and Statistics for Data Science

Assignment 6: Bayesian Inference and Regression Due: 04/29, 2:30pm, via google forms
(5 questions, 60 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)

$x = \frac{b^2\bar{X}+se^2a}{b^2+se^2}$ 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) Derive the estimates of β when minimizing the sum of squared errors and 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 $\bar{Y} = (\sum_{i=1}^{n} Y_i)/n$. (4 points)

(b) Show that the above estimators are unbiased (Hint: Treat X’s as constants) (3 points)

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

In this problem, we use the data from Azure trace; refer to q4.dat dataset on the website. The file contains 576 values. Each value represents the number of VMs running in a data center for a 5 minutes interval. Thus, the data spans exactly 2 days. Report all answers in your submission; you do not need to submit any code.

(a) Split the dataset into 4 equal parts. For each quarter of the data, using simple linear regression (include β_0 term), plot the original data and the regression fit (using the corresponding quarter of data as training), and calculate the SSE in all 4 cases. (5 points)

(b) Split the dataset into 2 equal parts. Use the first half of the data as the training set. Predict the data points for the second half of the data using exponential moving average (α=0.5), auto regression (p=3), and seasonal average (s=288). For each technique report the average errors across all the 288 predictions. Note that you may have to use predicted data for training. From the original data, use only the first 288 values as part of training (you can augment them with predictions for 289th point, 290th point, etc.), and use the final 288 points for computing the error. (5 points)

5. 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 $\boldsymbol{w} = \{w_1, w_2, w_3 \ldots, w_n\}$. 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 $C = \begin{cases} 0 \ \ if \ P(H = 0|\boldsymbol{w}\,) \geq P(H = 1|\boldsymbol{w}\,) \\ 1 \qquad\qquad\qquad\qquad otherwise \end{cases}$. 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 Q5_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 q5.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). (10 points)

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 $\boldsymbol{w} = \{w_1, w_2, w_3 \ldots, w_n\}$, derive $\boldsymbol{AEP}$ in terms of $\mu, \sigma, \Phi(\quad) \ and \ p$. (4 points)