DS 102 Discussion 5

Wednesday, March 2, 2022

1. **The Posterior Predictive Distribution for Ordinary Linear Regression**

   In lecture, we learned about a method to assess the validity of fitted Bayesian GLMs called *posterior predictive checks*. This approach uses the posterior predictive distribution (PPD) to compare data that the model would predict (sometimes called "replicate" data) to the data we actually observed. If the model is a good fit for the data, then the replicate data should look similar to the observed data.

   The PPD is the conditional distribution for this new replicate data $\tilde{y}$, conditioned on the data we observed. It's described by the following formula, which uses the fact that our model for the data y typically is based on some unobserved θ:

$$p(\tilde{y}|y) = \int_{\theta \in \Theta} p(\tilde{y}|\theta)\, p(\theta|y)\, d\theta$$

   The terms within the integral are familiar to us: the first is the likelihood of the new replicate data, and the second is the posterior distribution for θ given the data we actually observed. The bounds of the integral are over all possible values of θ.

   In lecture, we saw an example of how to simulate replicate values from the PPD via `PyMC3`. In this problem, we'll discuss how to apply this technique in the setting of Ordinary Linear Regression, where we can find the PPD analytically.

   In Ordinary Linear Regression, we assume that each observation is independent and have equal variance, expressed by:

$$y|\beta, \sigma \sim \mathcal{N}(X\beta, \sigma^2 I)$$

   where $y \in \mathbb{R}^n$, $X \in \mathbb{R}^{n \times k}$, $\beta \in \mathbb{R}^k$, and I is the $n \times n$ identity matrix. [1]We will use a uniform prior over regression parameters,

$$p(\beta, \sigma^2) \propto \sigma^{-2}$$

   This prior says that every possible value of β and σ^2 is equally likely. This is an example of an *improper prior*, because if β can be any vector in $\mathbb{R}^k$ and σ^2 any scalar in $\mathbb{R}^+$, the prior distribution does not integrate to 1. Fortunately, we are still allowed to use it for Bayesian inference as long as the posterior distribution is valid. [2]

[1]Question Source: Gelman, A. (2013). Bayesian Data Analysis. Chapman & Hall/CRC.

[2]Reading on Improper Priors: Ewing (2020, March 16). Improper Prior — Ben Ewing: What is an improper prior? https://improperprior.com/posts/2020-03-16-what-is-an-improper-prior/

(a) *Posterior Predictive Simulation*

Write out the steps involved in a posterior predictive check for a Bayesian Ordinary Linear Regression model.

The posterior predictive distribution for Ordinary Linear Regression is a Normal distribution due to properties of linear combinations of Gaussians [3]. In the following subparts, we will find the parameters of this Normal distribution.

(b) *Deriving the Mean of the PPD for OLR*

Show that $\mathbb{E}[\tilde{y}|\sigma, y] = \tilde{X}\hat{\beta}$. Interpret your result in words.

[3] Proof: https://www.cs.toronto.edu/~rgrosse/courses/csc411_f18/slides/lec19-slides.pdf

(c) (optional) *Deriving the Variance of the PPD for OLR*

Show that $\mathrm{Var}[\tilde{y}|\sigma, y] = \sigma^2 \left(I + \tilde{X}(X^T X)^{-1} \tilde{X}^T \right)$.

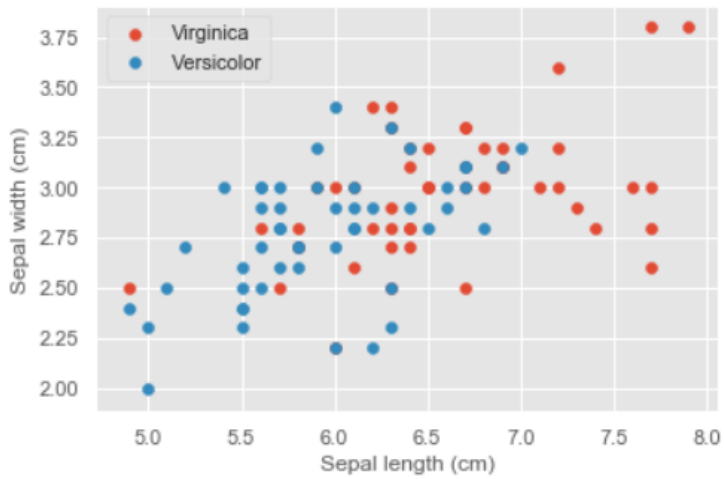
Thus, we have shown that the analytical form of the Posterior Predictive Distribution for Ordinary Linear Regression is:

$$\mathbb{P}\left[\tilde{y}|\sigma, y\right] \sim \mathcal{N}\left(\tilde{X}\hat{\beta}, \sigma^2 \left(I + \tilde{X}(X^T X)^{-1} \tilde{X}^T \right) \right)$$

2. **Interpreting the Logistic Regression Model**

In this problem, we fit a logistic regression model on a subset of the famous iris dataset. We have 100 samples of iris flowers, and measure their sepal length, sepal width, petal length and petal width (sepals are the small, green growths at the base of a flower). The response labels are whether they belong to the *Virginica* species (1) or the *Versicolor* species (0).

Let's say we first fit a Logistic regression model to predict the iris species, using only the sepal features. Then, our data is represented in the following plot:



(a) *Reformulating Logistic Regression*

From lecture, we have seen that Logistic Regression applies the sigmoid inverse link function to map a linear predictor $x_i^T\beta$ to probabilities in the following way:

$$\sigma(x_i^T\beta) = \frac{1}{1+e^{-x_i^T\beta}} = p \in (0,1)$$

where p is the probability of the i-th data point belonging to class 1. Reformulate Logistic Regression in terms of the *log* link function.

(b) *Interpreting a coefficient of a Logistic model*

Suppose after fitting the aforementioned logistic regression model, you observe the following output:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                species   No. Observations:                  100
Model:                            GLM   Df Residuals:                       97
Model Family:                Binomial   Df Model:                            2
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -55.163
Date:                Mon, 22 Feb 2021   Deviance:                       110.33
Time:                        00:47:22   Pearson chi2:                     100.
No. Iterations:                     4
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const         -13.0460      3.097     -4.212      0.000     -19.117      -6.975
sepal_width     0.4047      0.863      0.469      0.639      -1.286       2.096
sepal_length    1.9024      0.517      3.680      0.000       0.889       2.916
===============================================================================
```

Assuming that the model is correct, use the derivation in Part (a) to write a one sentence interpretation for the logistic model with respect to sepal length. What happens to the interpretation if the model is misspecified?

(c) *Goodness-of-Fit metrics for Frequentist GLMs*

We now build another logistic model which additionally includes petal width as a feature. You are presented with the following summary output:

```
                 Generalized Linear Model Regression Results
==============================================================================
Dep. Variable:                species   No. Observations:                  100
Model:                            GLM   Df Residuals:                       96
Model Family:                Binomial   Df Model:                            3
Link Function:                  logit   Scale:                          1.0000
Method:                          IRLS   Log-Likelihood:                -12.951
Date:                Mon, 22 Feb 2021   Deviance:                       25.902
Time:                        00:47:22   Pearson chi2:                     32.6
No. Iterations:                     8
Covariance Type:            nonrobust
===============================================================================
                  coef    std err          z      P>|z|      [0.025      0.975]
-------------------------------------------------------------------------------
const         -20.2873      8.055     -2.519      0.012     -36.075      -4.499
sepal_width    -4.8233      2.097     -2.300      0.021      -8.933      -0.714
sepal_length    1.2951      1.089      1.189      0.234      -0.839       3.430
petal_width    15.9227      3.981      4.000      0.000       8.121      23.725
===============================================================================
```

Which model has a better fit? How can you tell?

(d) *Understanding the Data Generating Process*

Looking at your finding from Part (c), your friend argues that the `petal_width` feature has a strong predictive effect that makes the second model better, so that must mean it's the only important factor differentiating the Virginica and Versicolor species. Explain why this argument is flawed.

3. **Bootstrap and the Sample Maximum**

Let $X_1, X_2, ..., X_n$ represent i.i.d draws from a Uniform$[0, 1]$ distribution. We wish to use the bootstrap to understand the sampling distribution of the maximum,

$$M_n = \max\{X_1, X_2, ..., X_n\}$$

We will use $X_1^*, X_2^*, ..., X_n^*$ to denote the bootstrap resamples.

(a) *Finding the Distribution of the Sample Maximum*
Compute $\mathbb{P}[M_n \leq t]$. Use this to compute the density of M_n.

(b) *Accuracy of Bootstrap Max Estimates*
Let $M_n^* = \max\{X_1^*, X_2^*, ..., X_n^*\}$. Find $\mathbb{P}[M_n^* = M_n]$.

(c) *Quality of Bootstrap Approximation of M_n*
Is the distribution of M_n^* a good approximation for the distribution of M_n? Why is this result to be expected?
Hint: Use the fact that $\lim_{n\to\infty}\left(1-\frac{1}{n}\right)^n = e^{-1}$.