CS 1675 Spring 2020: Test 02

Assigned April 1, 2020; Due: April 2, 2020

Shibo Xing

Submission time: April 2, 2020 at 5:00PM EST

Collaborators You are **not** permitted to work with anyone else for Test 02. All of the work must be your own.

Instructions

You must submit a PDF document to CourseWeb. That PDF can be created from the provided .Rmd template, or with another environment, such as Microsoft Word, LaTeX, or Google Docs. If you use the .Rmd template you are allowed to create as many equations blocks as you need, and add as much as markdown text as you need to answer the questions.

Problem 1

You are interested in predicting a continuous response, y, based on two continuous inputs, x_1 and x_2. You will use a Gaussian likelihood between the response and the mean trend, μ. You will use independent Gaussian priors on the coefficients of the mean trend function (the $\boldsymbol{\beta}$ parameters), with common prior mean μ_β and common prior standard deviation τ_β. You will use an Exponential prior on the noise, σ, and assume that the mean trend coefficients are independent of the noise, σ. The complete probability model, including the mean trend functional relationship, is given to you below.

$$y_n \mid \mu_n, \sigma \sim \text{normal}\left(y_n \mid \mu_n, \sigma\right)$$

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} x_{n,1}^2$$

$$\boldsymbol{\beta} \mid \mu_\beta, \tau_\beta \sim \prod_{d=0}^{D=2} \left(\text{normal}\left(\beta_d \mid \mu_\beta, \tau_\beta\right)\right)$$

$$\sigma \mid \lambda_\sigma \sim \text{Exp}\left(\sigma \mid \lambda_\sigma\right)$$

1a) (2 points)

Does this model include a non-linear relationship between the mean trend and the inputs?

solution: Yes, it does.

1b) (2 points)

Is this a linear model? Why or why not?

solution: This is an linear model since the beta parameters are still linearly related to the linear predictor.

1c) (2 points)

Assume that we collected $N = 25$ observations of the two inputs and the continuous response. Write out the mean trend expression for the 3rd observation, $\mu_{n=3}$. Use the notation that the n-th observation of the d-th input is denoted as $x_{n,d}$.

$$\mu_{n=3} = logit^{-1}(\beta_0 + \beta_1 x_{3,1} + \beta_2 x_{3,2} x_{3,1}^2)$$

1d) (4 points)

Write out the first two rows of the design matrix $\mathbf{X}$. Assume that the column associated with the intercept is included in the design matrix.

$$1, x_{1,1}, x_{1,1}x_{1,2}$$
$$1, x_{2,1}, x_{2,1}x_{2,2}$$

1e) (2 points)

Write out the expression for the mean trend of the 3rd observation as an inner-product in vector/matrix notation. Assume that the mean trend coefficients are in a column vector, $\boldsymbol{\beta}$.

$$\mu_{n=3} = logit^{-1}(X_{3,:}\beta)$$

1f) (2 points)

Instead of treating σ as unknown, as was specified in the problem statement, assume that σ is known, and thus it does not need to be learned with the $\boldsymbol{\beta}$ parameters. Write out the expression for the posterior covariance matrix on the mean trend coefficients, assuming this is the case.

$$cov(\beta) = \sigma^2 * X^T X$$

1g) (6 points)

Write out the contribution of the 3rd observation to the matrix sum-of-squares.

$$X_{3,:}^T X_{3,:}$$

1h) (2 points)

What would the value of the first column, first row be equal to in the matrix sum-of-squares accounting for all 25 observations?

$$X_{:,1}^T X_{:,1}$$

Problem 2

You will continue working with the model specified in Problem 1. You will consider that σ is unknown and uses the Exponential prior as stated in the Problem 1 problem statement.

2a) (2 points)

If you would fit the model in Problem 1 with the Laplace Approximation, why is it a good idea to transform σ to a parameter φ with a log-transformation?

solution: The MVN we can generate based on a Laplace Approximation is strictly symmetric of which the sigma value ranges from negative infinity to positive infinity. Yet the in log posterior the sigma is not allowed to be negative, and thus ne will need some transformation like logit function to enforce the bound on the unbounded noise sampled from MVN.

2b) (2 points)

We are also interested in a model with the following mean trend expression:

$$\mu_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2} + \beta_3 x_{n,1} x_{n,2} + \beta_4 x_{n,1}^2 x_{n,2} + \beta_5 x_{n,1} x_{n,2}^2 + \beta_6 x_{n,1}^3 x_{n,2} + \beta_7 x_{n,1} x_{n,2}^3$$

The prior on the mean trend coefficients and σ are the same as those used for the model in Problem 1. You fit both models to the 25 observations using the Laplace Approximation. You calculate the R-squared distribution associated with each model based on the 25 training points. The R-squared distribution for the model from Problem 1 is lower than the R-squared distribution for the model given in this problem.

Based on these results, is the model from Problem 1 definitely worse than the model in this problem? Why or why not?

solution: Not necessarilly, we haven't gotten to know any testing errors this model might encounter. In the cases with simple trends this type of high level model might results in overfitting and model from Problem 1 might be superior.

2c) (2 points)

The Laplace Approximation allows us to estimate the log-Evidence. The log-Evidence includes a penalty term, in addition to the term associated with goodness-of-fit. How are models penalized based upon the Laplace Approximation's estimate to the log-Evidence?

solution: According to Bayes Factors formula, the smaller the log[Evidence] is, the less plausible is a model. log[Evidence] consists of the sum of multiple terms, among which the $\frac{1}{2} log[2H|_\theta]$ has a negative sign, indicating that the log-determinant hessian matrix also known as the curvature of the log-posterior surface will penalizes the model performance.

2d) (2 points)

Would you expect the penalty term to be lower or higher for the model from Problem 1 relative to the model introduced in Problem 2? Why?

solution: The complex model would be expected to have a higher penalty term, which is the log determinant of the hessian matrix and can be calculated as the summation of the local curvatures of all likelihoods. The hessian matrix would become larger and thus the determinant will have higher value as the number of beta parameters increases. Thus, more complex models have higher penalties.

2e) (2 points)

If you decided to use k-fold cross-validation to compare the two models, how many times would each observation be used to evaluate the model performance if we used 5-fold cross-validation?

solution: 4 times

2f) (2 points)

If we use 5-fold cross-validation, how many times would we train and test each model?

solution: 5 times

2g) (2 points)

If we use 5-fold cross-validation how can we assess which model is better if we are interested in the RMSE?

solution: We can calculate the RMSE on each fold and plot those RMSE values for each spline in a box plot. By comparing RMSE values for each spline we can tell how much differently a model performs on different folds and what the average RMSE over all folds is. And thus, we can pick a model with small variability and small average RMSE.

2h) (2 points)

How many times would we train each model if we used Leave-One-Out (LOO) cross-validation, instead of 5-fold cross-validation?

solution: We would train the model the same number of times as the number of observations.

2i) (2 points)

Whether we use an information criterion or cross-validation to identify the best model, what are trying to guard against?

solution: We are trying to guard against using an overfitted model.

Problem 3

You will now model a binary outcome, y, based on two inputs, x_1 and x_2. The binary outcome takes on two possible states or classes. A value of $y = 1$ corresponds to an event of interest, and a value of $y = 0$ represents that we did not observe the event. You will fit a model for the event probability, μ, as a function of the two inputs.

You will use the following model written in the probability model format below.

$$y_n \mid \text{size} = 1, \mu_n \sim \text{Bernoulli}\left(y_n \mid \mu_n\right)$$

$$\mu_n = \text{logit}^{-1}\left(\eta_n\right)$$

$$\eta_n = \beta_0 + \beta_1 x_{n,1} + \beta_2 x_{n,2}$$

$$\boldsymbol{\beta} \mid \mu_\beta, \tau_\beta \prod_{d=0}^{D=2} (\beta_d \mid \mu_\beta, \tau_\beta)$$

3a) (2 points)

Is the model specified in the problem statement to Problem 3 a linear model?

solution: No, it is not.

3b) (2 points)

As you can see in the model specification, the variable η_n is related to the event probability, μ_n, via a transformation or *link* function. Specifically, you are using the logit-transformation to relate μ_n to η_n.

Would you still use the logit-transformation even if you do not fit the model with the Laplace Approximation? Yes, I would. Since, our model this time is a Bernoulli model, we calulate the probability of each input through bernoulli's function.

3c) (4 points)

Write out the first 2 rows of the design matrix, $\mathbf{X}$, associated with the model defined in Problem 3.

$$1, x_{1,1}, x_{1,2}$$
$$1, x_{2,1}, x_{2,2}$$

3d) (3 points)

Write out the expression for the 2nd observation's log-likelihood. You may keep the expression in terms of the outcome, $y_{n=2}$, and the event probability, $\mu_{n=2}$.

$$log[p(y_n|\mu_n)] = y_n log[\mu_n] + (1 - y_n) log[1 - \mu_n]$$

3e) (2 points)

The Hessian matrix of the log-likelihood depends on the weighting matrix, $\mathbf{S}$. Describe in words what the weighting matrix contains along its main diagonal.

solution: the diagonal consists of the probability of the bernoulli weight of each observation.

3f) (8 points)

Write out the contribution of the 2nd observation to the weighted sum-of-squares matrix.

$$ \begin{aligned} x_{2,:} x_{2,:} = \ 1, x_1, x_2,\ x_1, x_1^2, x_1x_2 \ x_2, x_1x_2, x_2^2

\end{aligned} $$

3g) (2 points)

After fitting the model, you will make predictions on a hold-out set. You are first interested in calculating the accuracy of the model. The model, however, does not predict a class directly. It predicts a probability of the event. The predicted probability is compared to a threshold value in order to convert the prediction into a discrete class (event vs non-event). If the predicted probability is greater than the specified threshold, the prediction is classified as the event.

The confusion matrix is calculated comparing each predicted classification to the observed class. What metric tells us the fraction of times the model is correct when the event is observed?

solution: the accuracy.

3h) (2 points)

The ROC curve is a graphical tool for assessing the performance of a binary classifier. What threshold value is used to construct the ROC curve?

solution: 0.5 for both true positive and false positive rates. (the decision threshold used to build the confusion matrix?)

3i) (2 points)

The ROC curve can be summarized as a single quantitative metric by integrating the area under the curve (AUC). What value for the AUC corresponds to random guessing?

solution: 0.5. Since, true positive rate is equal to false positive rate.

3j) (2 points)

If our model is struggling to correctly predict the event, can we just decrease the threshold value to improve performance? Why or why not?

solution: No we cannot. If we do that the decision threshold would have less correctness.

3k) (2 points)

Now assume that we used Leave-One-Out (LOO) cross-validation to assess model performance and we are interested in Accuracy. What are the possible Accuracy values our model could achieve on the test set in each fold?

Problem 4

We have discussed models for continuous responses and models for binary outcomes. However, there are many different types of models that we have not introduced in this course. For example, we can build models focused on predicting the number of events (referred to as counts) over a fixed interval of time. These types of models are useful for modeling rare events. A basic approach to modeling count data is with Poisson regression, which gets its name from the fact that the Poisson distribution is used as the likelihood.

In this problem, you are interested in trying to predict the number of events over a fixed time interval, y, based on a single input, x.

The Poisson likelihood for the n-th observation is given to you below.

$$y_n \mid \mu_n \sim \text{Poisson}\left(y_n \mid \mu_n\right) = \frac{\mu_n^{y_n} \times \exp\left(-\mu_n\right)}{y_n!}$$

The variable μ_n is referred to as the rate, and is the expected value of a Poisson distributed random variable. The Poisson distribution has the unique (and somewhat limiting property) that the expected value is also equal to the variance. Thus, $\text{var}\left(y_n\right) = \mu_n$.

4a) (4 points)

Write out the expression for the log-likelihood of the n-th observation. Keep the expression in terms of the observed counts, y_n, and the rate μ_n.

$$\begin{aligned} log(Poisson(y_n|\mu_n)) &= log[\mu_n^{y_n} e^{-\mu_n}] - log[y_n!] \\ &= log[\mu_n^{y_n}] + log[e^{-\mu_n}] - log[y_n!] \\ &= y_n log[\mu_n] - \mu_n - log[y_n!] \end{aligned}$$

4b) (3 points)

In Poisson regression, we model the rate μ_n. Thus, as with logistic regression we are actually modeling the expected value, not the outcome itself. The rate however is not modeled directly. As with logistic regression a transformation or *link* function is used to relate the rate, μ_n, with a transformed variable, η_n:

$$\eta_n = g\left(\mu_n\right)$$

The rate is recovered from η_n via the inverse link function:

$$\mu_n = g^{-1}\left(\eta_n\right)$$

The most commonly used link function for Poisson regression is the log-link:

$$\eta_n = \log\left[\mu_n\right]$$

What is the inverse link function which allows back-transforming from η_n to μ_n?

$$\mu = exp(\eta_n)$$

4c) (12 points)

You will work with a simple relationship between the single input, x, and the variable η_n. The deterministic function will be that the input is linearly related to the η_n:

$$\eta_n = \beta_0 + \beta_1 x_n$$

Derive the expression for the partial first derivative of the n-th log-likelihood with respect to β_1.

$$\begin{aligned}\frac{d(y_n log[\mu_n] - \mu_n - log[y_n!])}{d\beta_1} &= \frac{dy_n log[\mu_n]}{d\beta_1} - \frac{d\mu_n}{d\beta_1} - \frac{dlog[y_n!]}{d\beta_1} \\ &= \frac{dy_n log[\mu_n]}{d\beta_1} - \frac{d\mu_n}{d\beta_1} \\ &= \frac{dy_n log[exp(\beta_0 + \beta_1 x_n)]}{d\beta_1} - \frac{e^{\beta_0+\beta_1 x_n}}{d\beta_1} \\ &= \frac{dy_n(\beta_0 + \beta_1 x_n)}{d\beta_1} - \frac{e^{\beta_0+\beta_1 x_n}}{d\beta_1} \\ &= y_n x_n - e^{\beta_0+\beta_1 x_n} \cdot x_n\end{aligned}$$

4d) (10 points)

Derive the expression for the partial second derivative of the log-likelihood with respect to β_1.

$$\begin{aligned}\frac{d^2(y_n log[\mu_n] - \mu_n - log[y_n!])}{d\beta_1^2} &= \frac{d(y_n x_n - e^{\beta_0+\beta_1 x_n} \cdot x_n)}{d\beta_1} \\ &= x_n \cdot e^{\beta_0+\beta_1 x_n} \cdot x_n = x_n^2 \cdot e^{\beta_0+\beta_1 x_n}\end{aligned}$$