midtermq

October 31, 2024

1 Midterm

1.1 Part I

A researcher wishes to fit a linear regression model with several predictors. She is considering three loss functions:

1. $loss = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \lambda^2 \sum_{j=1}^{p} \beta_j^2 + 5\lambda$
2. $loss = \sum_{i=1}^{n} |y_i - \hat{y}_i| + \sum_{j=1}^{p} |\beta_j|$
3. $loss = \prod_{i=1}^{n} \frac{(y_i - \hat{y}_i)^2}{y_i^2}$

For each of these options:

1. Give an intuitive explanation for this choice of loss function. How does it express the desire for accurate predictions of a quantitative variable?
2. Do you see any possible issues with this loss function? What assumptions have to be true about the data for this loss function to be viable?
3. Find an equation for the *gradient* of the loss function. (A general equation for the partial derivative at β_j will suffice.)
4. Give a brief code outline (psuedocode) to show the procedure you would use to calculate the "best" β's according to this loss function. You code to does not need to run; however, it **does** need to be specific about the inputs as well as the equations. For example:

Sufficient:

```
def ols(x, y):
    sx = std dev of x
    sy = std dev of y

    mx = mean of x
    my = mean of y

    rxy = correlation of x and y

    beta_1 = sy/sx * rxy
    beta_0 = my - beta_1 * mx

    return beta_0, beta_1
```

1.2 Part II

1.2.1 Question A

A researcher is trying to fit a linear regression model using a LASSO penalty. To choose a good value of the penalty parameters, λ, she decides to take the following approach:

- For each of the many possible λ values:
  - Fit the model using that λ
  - Find the predicted values from the model, $\hat{y}_1 ... \hat{y}_n$
  - Calculate the residuals, $r_i = y_i - \hat{y}_i$
  - Calculate the sum of the squared residuals
- Then we choose the value of λ that resulted in the model with the smallest sum of squared residuals.

Discuss this strategy: Do you think it is a good idea? Why or why not? Do you have any suggestions to make it more efficient, more justifiable, or more correct?

1.2.2 Question B

Your fame as a data scientist has spread far and wide, and university hires you to investigate faculty happiness. They supply you with two datasets:

1. 1000 emails sent from faculty accounts in January 2022 all of which have been analyzed by language experts and given a "happiness score" on a scale of 0 to 100.
2. 10,000 emails sent from faculty accounts in February to December 2022, which have not been analyzed.

The university cannot afford to hire their language analysts for all 10,000 emails, but they can bring them back for another batch of around 1000 if you ask them to.

Your client would like you to create a predictive model from the 1000 email dataset, that can be used to assess the happiness of the 10,000 email dataset. They would then like you to tell them if faculty happiness was increasing, decreasing, of staying the same across 2022.

Propose a modeling process to address this question

You should include:

- A (very brief) description of how you might **pre-process** the data
- A **model specification**, and **why** you think that model would be a good choice for this task. This can be a model we studied, an existing model we haven't studied, or you can "invent" something–but you must include some discussion of why you think that choice is good/reasonable for this scenario.
- A **loss function** that you will use to fit the model, and **why** this loss function correctly expresses your desires for your "best" model.
- A **metric** you will use to report your model's abilities and/or to tune hyperparameters, and **why** this metric is a good measure of "model success" in this case. This metric should not be R-squared, MAE, or MSE. (Those would be reasonable in this case, but I want you to find/invent a different one and justify it.)

Note: You do *not* need to concern yourself in this question with the computational feasibility of your model, loss, or metric. I am only looking for your to translate the "real world" needs of the

scenario into mathematical decisions.