Data Thinning vs Data Splitting

Hansen Zhang, Spring 2024

Mentor: Ethan Ancell

Setting Up: LASSO Regression

- The most common way of fitting a linear model is through ordinary least squares (OLS), which involves minimizing the expression below:

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 .$$

However, this can result in a model which contains superfluous variables!

- LASSO performs variable selection: it is a method of fitting our linear model which shrinks “less relevant” coefficients in the linear model exactly to 0, thus removing them entirely from the model.
- This selects for a set of predictors which are “most” relevant to the linear relationship
- Variable selection helps to identify the covariates which most meaningfully have a relationship with the response

$$\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{j=1}^{p}\beta_j x_{ij})^2 + \lambda \sum_{j=1}^{p}|\beta_j|$$

Problem: “Double Dipping”

“Double Dipping” is the act of using the same data to select a regression model and test/validate the model [Kriegeskorte 2009, Neufeld 2023].

- Following on our LASSO regression, if we wish to perform inference for **the selected coefficients**, it becomes problematic because we cannot do inference using the same data that we selected our variables with…
- This will result in invalid inference!

Solutions to Double Dipping

Splitting - the most common way to overcome this is to partition or “split” our data into two sets of observations: a training set and a test set.

Thinning - a newer method for creating independent training and test sets based upon a resampling technique and membership within certain classes of distributions. This method works especially well when there are relatively few observations.

A simple modeling strategy

$$\begin{aligned} Y_i &= f(X_1, X_2, \ldots, X_p) + \epsilon_i \\ &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_p X_{ip} + \epsilon_i \end{aligned}$$

where Y_i is the observed outcome and X_{ij} the realization of the jth predictor of the ith data point:

$$d_i = (X_1, X_2, \ldots, X_p, Y)$$

Notice, we can rewrite the above linear relationship as

$$Y_i = \beta^T X_i = \begin{bmatrix} \beta_1 & \beta_2 & \ldots & \beta_p \end{bmatrix} \begin{bmatrix} 1 \\ X_{i1} \\ X_{i2} \\ \vdots \\ X_{ip} \end{bmatrix} + \epsilon_i$$

The additional random noise, ϵ_i, is also a random variable, which we will assume is normally distributed with mean 0:

$$\epsilon \sim N(0, \sigma^2) \perp X_i$$

Thus, by the sum of independent normal variables, we have that:

$$Y_i = \beta^T X_i + \epsilon \sim N\left(\beta^T X_i, \sigma^2\right)$$

How to do data thinning with this distribution

Then, from Table 2 (Neufeld et al. 2023)

$$Y_i^{(tr)} \sim N(\epsilon\beta^T X_i, \epsilon\sigma^2)$$

$$Y_i^{(te)} \sim N((1-\epsilon)\beta^T X_i, (1-\epsilon)\sigma^2)$$

*Note: The epsilon above is not the same as the random noise in the previous slide

Efficiency of Inference on Beta

- Does thinning or splitting result in more efficient inference?
- To test this, we run a simulation in R:

Steps (for split and thinned, separately)

1. Fix arbitrary sparse beta (the vector of coefficients) and fix an arbitrary X
2. Sample an observed Y = (Y1, Y2, …, Yn) by simulating additive random noise
3. Perform variable selection using train set
4. Perform inference on only the LASSO-selected beta variables (95% CI) using test set
5. Repeat steps 2-5 many times

Results

After running the simulation in R using 1000 repetitions, 50 predictors and 200 observations, we have that:

- When doing selective inference naively (not using splitting or thinning) yields average coverage of 0.89, which disagrees with our nominal coverage rate of 0.95. (This is evidence that we have invalid inference if we don't split or thin)
- With thinning and splitting, the average coverage rate was 0.9466 and 0.9427 respectively (valid inference)
- However, the average lengths of the confidence intervals from thinning and splitting differed greatly at 0.45 and 0.55 respectively. (We get more efficient inference with thinning.)

Conclusion

- *When our target of inference depends on our data*, we need to use a information division technique like splitting or thinning to achieve valid inference. However…
- The results imply that data thinning is superior to data splitting when it comes to inferential efficiency

Thank You

References

James G, Witten D, Hastie T, Tibshirani R. An Introduction to Statistical Learning : With Applications in R. Springer; 2013.2.

Neufeld A, Dharamshi A, Gao LL, Witten D. Data thinning for convolution-closed distributions. arXiv.org. doi:https://doi.org/10.48550/arXiv.2301.072763.

Kriegeskorte N, Simmons WK, Bellgowan PSF, Baker CI. Circular analysis in systems neuroscience: the dangers of double dipping. Nature Neuroscience. 2009;12(5):535-540. doi:https://doi.org/10.1038/nn.2303