STA302 - Lecture 8

Cedric Beaulac

June 6, 2019

Introduction

Introduction

- Test#2 is done!
- Core content is lecture 3,4,5,6 and 7.
- Lecture 8 and 9 are discussing modern problems and they open up the course to more advanced topics.
- Lecture 10 : No new content. Probably a review lecture ? (Truth is, I have to give the final exam to the department next Thursday)

Today's plan

- Today we will introduce variable selection methods and keep an eye on the current state of research in data analysis. We will :
  - Do a quick review of our model
  - Discuss overfitting and large date set issues
  - Introduce basic variable selection methods

Review

Review of the model

- We have established a basic model that allow p predictors x_j to influence our prediction for a given response y.
- $\mathbf{y} = \mathbf{X}\beta + \mathbf{e}$, where $\mathbf{X}$ is $n \times (p+1)$ matrix of observed predictor values, $\mathbf{y}$ a $n \times 1$ vector of response values, β the $p \times 1$ vector of parameters and $\mathbf{e} \sim MVN(0, \sigma^2 I)$.
- We estimate β with $\hat{\beta}_{MLE} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$.
- It implies $\hat{\beta} \sim N(\beta, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$.
- Remember, the unbiased estimate for σ^2 is $\sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-(p+1))$

Big Data

Large data sets

- BIG DATA (in capital letters) used to be the most trendy word in data science.
- Defining big data is complicated but some of the challenges caused by big data are :
  - Data storage
  - Data analysis
  - Data visualisation
  - Information privacy

Large data sets

- *Big data usually includes data sets with sizes beyond the ability of commonly used software tools*
- We can define multiple axis for which a data set can be considered *large* :
  - Volume/Tall data : Large number of observations (large n)
  - Wide data : Large number of predictors (large p)
  - Variety : Multiple styles of data from texts to images to audio and video files.
  - Velocity : The speed at which the data is generated.

Large numbe of predictors

- We will not solve every big data problems today (we might never fix them actually)
- We will introduce some proposed solution to take care of one specific problem : Large number of predictors (large p).
- I will try to convince you that it is problem. (For linear regression model, that's not flattering)
- Then we will discuss *variable selection*

Large number of predictors

- *When do we have a large number of predictors ?*
- We can simply have a large number of parameters (modern style data set, netflix, facebook, hospital records, genetic codes, etc. . .)
- Want to investigate many interaction terms and interaction of high order (which increses the number of parameters).
- Want to add multiple polynomial terms (which increses the number of parameters).

Large number of predictors : Why is it a problem ?

- **It reduces interpretation.** Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Even though Occam's razor's principal is debatable, everyone agree that it is easier to explain a simpler model and that in order to get the big picture, we are willing to sacrifice small details.
- **It increases the variance of the estimates.** Having more parameters to estimate increases the variace of the estimators. This is reflected in $s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n-(p+1))$.

Large number of predictors : Why is it a problem ?

- **It is more prone to overfitting.** The more parameters, the more complexe the model can be which increases the chances that we fit *too much* the data set to the detriment of generalization abilities.
- **It increases the chances of collinearity issues.** Collinearity is caused by having too many variables poviding similar information. The more predictors you have in the model the higher the chances are that some provide similar information.

Large number of predictors : Why is it a problem ?

- We have already talked about interpretability, large variance and collinearity in previous lectures.
- But what is *overfitting* ?

Overfitting

Overfitting

- Overfitting happens when we detect a pattern in the data set that does not exist for new observations.
- When we allow a model to capture all sorts of complexe relationships using many predictors, the model might capture some relationships that only appears by luck in the data set.
- These complexe relation do not represent the *reality* and thus our model will have a poor performances on new observation.
- We say that a model overfits when it offers poor generalization abilities (generalization: performances of the model on non-observed points).

Overfitting

- In other words, $\hat{\beta}$ does not represent the general relationship between $\mathbf{x}$ and y for all $\mathbf{x}$ and y but rather something specific to the sampled data set.
- This is truly undesireable because we often have to deal with a small data set but the purpopse of a model is to understand a more global relationship.
- We wish to understand this relationship to solve new problems, for prediction purposes and for insightful decision making.
- A model that overfits can not help us with these problems.

Overfitting

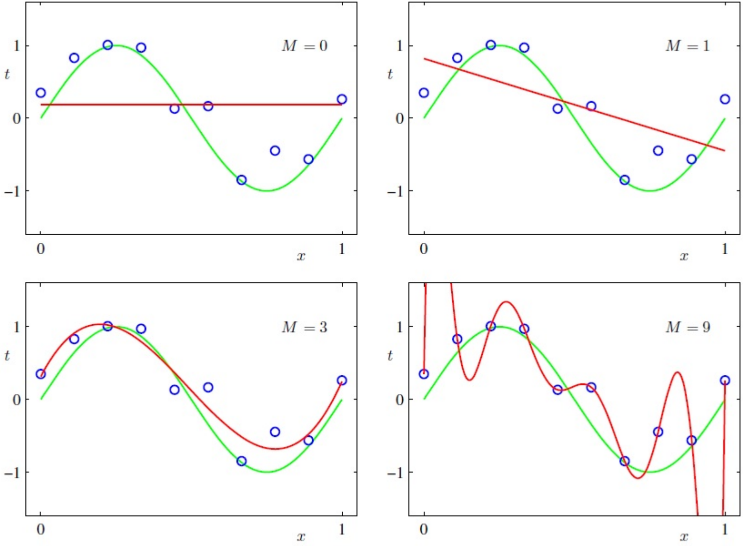

Figure 1: Allowing polynomial terms of high order can cause overfitting

Overfitting

- The figure above shows an overfitting issue that occurs when we utilize a polynomial fit of high order.
- We do end up with a model that perfectly explains the data set.
- But with $M = 3$ our predictions would be much better.
- Overall, $M = 3$ would provide a much better generalization. or 1
- (For $M = 0$ or 1 we would say the model *underfit* and it would show in the residuals plot)

Overfitting

- It was typical in statistics to define overfitting as a function of model complexity (number of parameters p) and number of observations (n) as shown in the picture above.
- Machine learning came up with models that contain way more parameters than observations (Neural Networks) and forced a new definition for overfitting.

Overfitting

- It is common to define a training (data) set and a test (data) set.
- The training set is used for training: in our problem it is the matrix of predictors $\mathbf{X}$ and the vector of response $\mathbf{y}$ we used to build the estimates :$\hat{\beta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$
- The test set is used to asses the performances of the fitted model on *new* observations. It contains observation that were not used to fit $\hat{\beta}$.
- We say a model overfits if it has been fitted to have too good performances on the training set to the detriment of test set performances.
- The gap between the performance on test set and training set can be used to assess overfitting problems.

Overfitting

- We can also observe symptoms of overfitting by selecting a performance metric and comparing its value on the training set to its value on the test set as we change the model complexity.
- The mean squared error (MSE) or the maximum likelihood are examples of reasonnable metrics.
- For the example illustrate above, here is the MSE over the training set a test set as the degree of the polynomial fit changes.

Overfitting

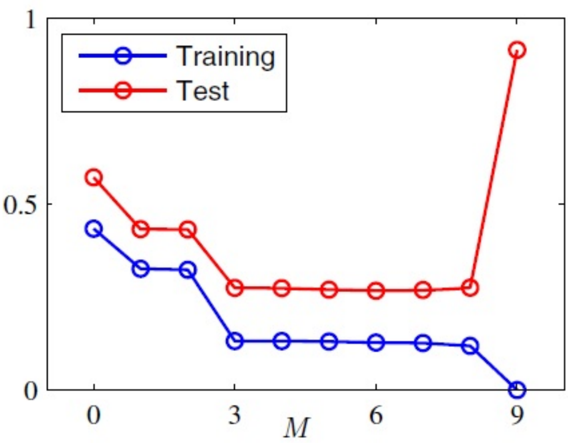

Figure 2: Allowing polynomial terms of high order can cause overfitting

Overfitting

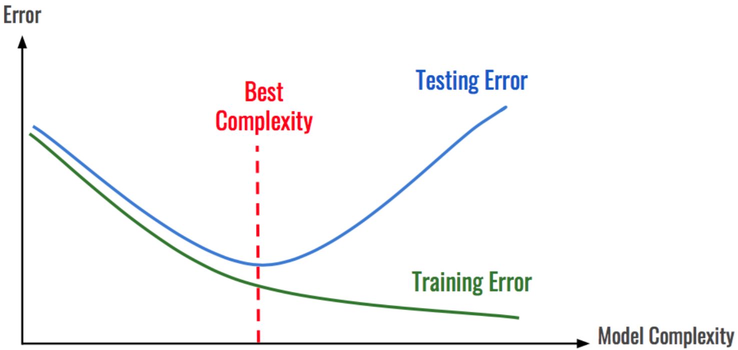

Figure 3: Thank you google

Overfitting : Conclusion

- Overfitting is a big concern in today's research.
- It is one of the main culprit of the lack of reproducibility in many experiments.
- A model with a large amount of parameters (large p) increases the chances of overfitting.

Variable Selection

Large number of predictors : Why is it a problem ?

- **It reduces interpretation.** Occam's Razor states that among several plausible explanations for a phenomenon, the simplest is best. Even though Occam's razor's principal is debatable, everyone agrees that it si easier to explain a simpler model and that in orderto get the big picture, we are willing to sacrifice small details.
- **It increases the variance of the estimates.** Having more parameters to estimate increases the variace of the estimators. This is reflected in $s^2 = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2/(n - (p+1))$.

Large number of predictors : Why is it a problem ?

- **It is more prone to overfitting.** The more parameter, the more complexe the model can be which increases the chances that we fit *too much* the data set to the detriment of generalization abilities.
- **It increases the chances of collinearity issues.** Collinearity is caused by having too many variables poviding similar information. The more predictors you have in the model the higher the chances are that some provide similar information.

Variable selection

- In this lecture we will introduce classic variable selection tools.
- We will define them (mathematically and intuitively).
- We will finally see how to use them in R
- Next Tuesday (LAST LECTURE!) we will see modern approaches to variable selection and shrinkage methods.

Variable selection

- Let's begin by considering simple metrics of performance.
- R^2 coefficient and the log-likelihood of our model.
- We could be tempted to select the set of variables that maximizes the likelihood or the R^2 coefficient but this can only lead to overfitting.
- As we introduce new predictors (or interactions etc) to the model, the R^2 coefficient and the log-likelihood can only increase (IMPORTANT).
- The motivation behind the commonly used metrics is to use the R^2 coefficient or the log-likelihood but to penalizes for high number of parameters.

Variable selection : Adjusted R^2.

- Some of you might have noticed the R^2-adjusted in the linear regression output table in R.
- Remember:

$$R^2 = \frac{SSreg}{SST} = 1 - \frac{SSE}{SST}$$

- R^2-adjusted includes a penalty per parameters :

$$\text{Adjusted } R^2 = 1 - \frac{SSE/(n-p-1)}{SST/(n-1)}$$

- A large Adjusted R^2 indicates a good improvement over $\bar{y}$.

Akaike information criterion (AIC)

- The AIC is a likelihood-based metric with a penalty for the number of parameters.

$$AIC = 2p - 2l$$

where p is the number of parameter and l the log-likelihood of the current model.

- The smallest the AIC is the better the model is.

Bayesian information criterion (BIC)

- The BIC is also a likelihood-based metric with a penalty for the number of parameters.

$$BIC = \log(n)p - 2l$$

where p is the number of parameter and l the log-likelihood of the current model.

- The smallest the BIC is the better the model is.

AIC vs BIC

- BIC has a stronger penalty for the number of parameter ($\log(n)$ > 2 as long as $n > 7$).
- It sometimes lead the BIC to select underdevelopped model but makes AIC more prone to select a model that overfits.
- Hard to say which one is the best, and while considering the adjusted R^2 we have many metrics which is great, but they might tell different stories, which is less desirable.

Variable selection

- We can now use either of the metrics we just defined to compare two models and select the *best* one without any fear of simply selecting the one with the highest amount of parameters.
- When comparing two models with these tools we make sure that if they differ by only a parameter or two, the added parameters must be worth a *price* in order to be selected.
- We make sure that when we add a parameter to the model, the improvement to the model is worth the price (in interpretability etc) of the extra parameter.
- How to compare multiple models ? More precisely, how to establish the list of models to compare?

Variable selection

- The first idea that comes to mind is : *let's try ALL the models*.
- If we have p predictors, it would lead to 2^p different models.
- It might be reasonnable to fit all of the 2^p models and select the one with the highest adjusted R^2 of lowest AIC/BIC.
- Remember that when comparing models wih the same number of parameters all of the metric will select the same model but when the list of models contains models of different complexities, these metrics might select different models.

Variable selection : hierarchical models

- It would seems natural to establish some kind of path through the models when comparing them.
- When working with *hierarchical* models, it is pretty easy to define a natural path through the models.
- Polynomial fits are an example of model wiht a natural hierarchy. A polynomial fit with x^2 is of higher order than a fit with x only.
- One way to proceed would be to start with only x, then include x^2 and compare the two models with one of our metrics.
- We can increase the order of the model as long as the adjusted R^2 (AIC/BIC) keeps increasing (decreasing) or as long as the added terms are significants.

Variable selection : hierarchical models

- A similar process can be done with interactions.
- Assuming we have three predictors. We should start with a model with no iteraction.
- Then add all the second order terms : the three interaction terms and the three quadratic terms.
- Finally add the third order terms : the three-way interaction term and the three cubic terms.

Variable selection : stepwise subsets

- Remember that with p parameters, there exist a toal of 2^p models.
- Let's now introduce two techniques to examine a sequential subset of the 2^p possible regression models.
- Both of these techniques are well-know and extremely popular but they were also heavily criticise.

Variable selection : Forward selection

- Forward selection is a stepwise subset technique that starts with the simplest model (no predictors) and sequentially add predictors to the model.
- At every step, for all predictors that are not in the model, we check their p-value if they were added to the model and add the predictor that would have the smallest p-value.
- We stop the process when R^2 (AIC/BIC) decreases (increases) or when all parameters were added to the model.

Variable selection : backward elimination

- Backward elimination start with all of the possible predictors.
- At every step, for all predictors in the model, we take out the predictor with the largest p-value.
- We stop the process when R^2 (AIC/BIC) decreases (increases) or when all parameters were taken out of the model.

Variable selection : stepwise subsets

- These two technique share similar pros :
  - They are easy to use
  - They are intuitive
  - They are computationnaly cheap (they are both a lot cheaper than looking through all subsets)

Variable selection : stepwise subsets

- But they also share major weaknesses
  - They don’t use p-value appropriately
  - By testing model sequentially we might stop before finding the best model.
  - The selection procedure disturbs inference and prediction (well that’s a big problem).

Variable selection : Post-selection inference

...the selection process changes the properties of the estimators as well as the standard inferential procedures such as tests and confidence intervals. The regression coefficients obtained after variable selection are biased.

- The p-value are usually much smaller.
- And the t-statistic and F-statistic can be missleading.

Variable selection : Post-selection inference

- *Regardless of sample size, the model selection step typically has a dramatic effect on the sampling properties of the estimators that can not be ignored. In particular, the sampling properties of post-model-selection estimators are typically significantly different from the nominal distributions that arise if a fixed model is supposed.*
- *As a consequence, naive use of inference procedures that do not take into account the model selection step (e.g., using standard t-intervals as if the selected model had been given prior to the statistical analysis) can be highly misleading.*
- Leeb and Potscher (2005, page 22)

Variable selection : Post-selection inference

- Post-selection inference has been a big problem in the early 2000's
- Slowly, solutions were found; mostly they relly on computing the conditional distibution of the parameters (conditionnaly on the selection process)

Variable selection : Conclusion

- Modern data set are extremely large, both in number of observations n and number of parameters p.
- Having a too large number of parameters seems like a good thing (and should be. . . .) but can be problematic in linear regression.
- To reduce the number of predictors, variable selection schemes were established.
- To begin we need to make sure our evaluation metrics account for the number of predictors.
- Then we need to establish the list of models to test and an efficient way to go through that list.
- Variable selection techniques can fix most of the issues related with large p but they cause a new problem; the inference is unreliable.

Practice Problems

- A modern approach to regression with R ch.7 : 1,2 and 3 (Most of the solutions are in the book directly)
- HARD PROBLEMS The elements of statistical learning ch.3 : 8 and 9

External Sources

- A modern approach to regression with R ch.7
- Linear Models with R ch.8
- The elements of statistical learning ch.3 Click here
- First two sections of this link