GENERAL ASSEMBLY

MODEL SELECTION

Brian Chung

LESS THAN PERFECT BOOK RAFFLE

```
students = [‘Doug’,’Dylan’,’Logan’,’Thomas’]
np.random.choice(students)
```

FROM LAST TIME

Questions:

- Other variants of K Means (Why k means or k medians)
- Non-convex data for K Means
- Choosing enough centroids
- How many features to use in k means?

MODEL SELECTION AGENDA

I. LINEAR MODEL SELECTION
II. CRITERION
III. PROJECT / FREE TIME

I. LINEAR MODEL SELECTION

CHOOSING FACTORS IN A LINEAR MODEL

Q: In linear regression, what is the purpose of feature selection?

CHOOSING FACTORS IN A LINEAR MODEL

Q: In linear regression, what is the purpose of feature selection?

A: Better prediction accuracy as well as model interpretability

CHOOSING FACTORS IN A LINEAR MODEL

Q: When we have many features that we think might or might be important to our prediction problem, what are some ways we can choose the “right” features?

CHOOSING FACTORS IN A LINEAR MODEL

Q: When we have many features that we think might or might be important to our prediction problem, what are some ways we can choose the “right” features?

A: There are a couple methods, including one such method we’ve already learned

CHOOSING FACTORS IN A LINEAR MODEL

Q: When we have many features that we think might or might be important to our prediction problem, what are some ways we can choose the “right” features?

A: There are a couple methods, including one such method we’ve already learned

- Shrinkage (Ridge Regression, LASSO)

CHOOSING FACTORS IN A LINEAR MODEL

Q: When we have many features that we think might or might be important to our prediction problem, what are some ways we can choose the "right" features?

A: There are a couple methods, including one such method we've already learned

- Shrinkage (Ridge Regression, LASSO)
- Subset Selection (**Today**)

CHOOSING FACTORS IN A LINEAR MODEL

Q: When we have many features that we think might or might be important to our prediction problem, what are some ways we can choose the “right” features?

A: There are a couple methods, including one such method we’ve already learned

- Shrinkage (Ridge Regression, LASSO)
- Subset Selection (**Today**)
- Dimension Reduction (**Future: Principle Components Analysis**)

SUBSET SELECTION

There are a few main methods used in subset selection of features:

- Best Subset Selection
- Forward Stepwise Selection
- Backward Stepwise Selection

As well as a few different metrics used in measuring “goodness”:

- AIC
- BIC
- Adjusted R^2

BEST SUBSET SELECTION

Assume you begin with **p** potential features, of which you believe **k** of these features are pertinent to your problem.

In best subset selection, we iterate through all potential subsets of (p choose k) factors, evaluating them using a common metric (AIC, BIC, Adjusted R^2)

Ultimately, we choose the best set of **k** factors that have been tested through cross validation

BEST SUBSET SELECTION – ALGORITHM

1. Define M_0 to be the *null model*, indicating a model with zero features. The model predicts the response average for every observation
2. For k = 1, 2, …, p:
    1. Fit all (p choose k) models that contain exactly ***k*** features
    2. Pick the model among these (p choose k) models that has the highest adjusted R^2. Call this model M_k
3. Amongst the models from M_0 to M_k , choose the single best model with the highest cross validated adjusted R^2.

FORWARD STEPWISE SELECTION

That's great...except what if you have 100 features? Thats (100 choose 1) + (100 choose 2) + (100 choose 3) + ... + (100 choose 100) different subsets to try. It's computationally extremely unfeasible for large values of **p**.

So, what if we try an augmented approach. Let's start with a blank model, and choose the best single feature to add onto this model. Then, let's choose the next best feature to add onto this new model. Almost like cherry picking the best "next additional feature" to add.

Ultimately, we'll have M_0, M_1, M_2, ..., M_k models, where each is building upon the last.

STEPWISE FORWARD SELECTION – ALGORITHM

1. Define M_0 to be the *null model*, indicating a model with zero features. The model predicts the response average for every observation
2. For k = 1, 2, …, p:
    1. Fit all models that augment M_{k-1} with a single additional feature.
    2. Pick the model among these models with the highest adjusted R^2 and call this M_k
3. Amongst the models from M_0 to M_k , choose the single best model with the highest cross validated adjusted R^2.

BACKWARD STEPWISE SELECTION

1. Define M_p to be the *full model*, indicating a model with every feature. The model predicts based on the fitted coefficients.
2. For k = p, p-1, p-2, ..., 1:
    1. Fit all models that augment M_k by removing a single feature.
    2. Pick the model among these models with the highest adjusted R^2 and call this M_k
3. Amongst the models from M_0 to M_k , choose the single best model with the highest cross validated adjusted R^2.

FORWARD OR BACKWARDS?

Forward and backward selection will often give similar but slightly different results.

Personally, I choose forward selection when I want the simplest possible model, and choose backwards selection when I believe the ‘true’ model is close to what I have right now

The main issue with either stepwise selection methods is that of optimality. We use a heuristic to choose features, but these may not be the most ideal **set** of features

I. LINEAR MODEL SELECTION
II. CRITERION

OTHER SELECTION CRITERION

Recall that the *adjusted* R^2 penalizes the goodness of fit for adding many features or variables.

OTHER SELECTION CRITERION

Recall that the *adjusted* R^2 penalizes the goodness of fit for adding many features or variables.

However, for models that are fit using maximum likelihood (this include linear regression as well as logistic regression), you can also perform model selection by minimizing the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).

OTHER SELECTION CRITERION

Recall that the *adjusted* R^2 penalizes the goodness of fit for adding many features or variables.

However, for models that are fit using maximum likelihood (this include linear regression as well as logistic regression), you can also perform model selection by minimizing the Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC).

Information criterion essentially measure the "loss of information" when moving from some true model “f” to a model “g”. While we do not know the true model “f”, we can compare two different models g_1 and g_2 using AIC or BIC.

AKAIKE INFORMATION CRITERION

The AIC is defined as -2 times the log likelihood of the model plus 2 times the number of features

Given two models with different AIC values, the preferred model is the one with a lower AIC

$$\mathrm{AIC(M)} = -2\ln(L) + 2k$$

AKAIKE INFORMATION CRITERION

The AIC is defined as -2 times the log likelihood of the model plus 2 times the number of features

Given two models with different AIC values, the preferred model is the one with a lower AIC

$$\mathrm{AIC(M)} = -2\ln(L) + 2k$$

AIC(M): Akaike Information Criterion for model M
ln(L): Log likelihood for model and data
k: number of features

BAYESIAN INFORMATION CRITERION

The BIC is defined as -2 times the log likelihood of the model plus the natural log of the number of observations times the number of features

Given two models with different BIC values, the preferred model is the one with a lower BIC

$$\mathrm{BIC(M)} = -2\ln(L) + \ln(N) * k$$

BAYESIAN INFORMATION CRITERION

The BIC is defined as -2 times the log likelihood of the model plus the natural log of the number of observations times the number of features

Given two models with different BIC values, the preferred model is the one with a lower BIC

$$\mathrm{BIC(M)} = -2\ln(L) + \ln(N) * k$$

BIC(M): Bayesian Information Criterion for model M
ln(L): Log likelihood for model and data
ln(N): log of number of observations
k: number of features

AIC VS BIC

The equations are similar, but are born from different justifications.

- AIC believes that all models are incorrect, and only looks for the best approximation.
- BIC believes that the “true model” is a simplification of the current set of features.
- Generally, AIC tends to favor overfitting, while BIC tends to favor underfitting

$$\mathrm{AIC(M)} = -2\ln(L) + 2k$$

$$\mathrm{BIC(M)} = -2\ln(L) + \ln(N) * k$$

SO WHICH DO WE USE IN MODEL SELECTION?

Generally speaking, when using AIC vs BIC for stepwise selection, both should yield similar results.

Statsmodels provides all the metrics including AIC and BIC.

Let's look at a revised example of Forward Stepwise Selection using the AIC (you can replace AIC with BIC). Just remember that now, instead of maximizing adjusted R^2 we want to minimize AIC or BIC.

STEPWISE FORWARD SELECTION – ALGORITHM

1. Define M_0 to be the *null model*, indicating a model with zero features. The model predicts the response average for every observation
2. For k = 1, 2, …, p:
    1. Fit all models that augment M_{k-1} with a single additional feature.
    2. Pick the model among these models with the **lowest AIC** and call this M_k
3. Amongst the models from M_0 to M_k , choose the single best model with the **lowest** cross validated **AIC**.

I. LINEAR MODEL SELECTION
II. CRITERION
III. PROJECT / REVIEW / ANYTHING

THAT'S IT!

- Exit Tickets: DAT1 - Lesson 10 - Model Selection