Data 401
Model Selection

Dennis Sun

October 10, 2016

1. The Class So Far
2. Stepwise Regression
3. Lasso

1. The Class So Far
2. Stepwise Regression
3. Lasso

The Class So Far

The Class So Far

- Between basis expansions and interaction terms, the number of candidate variables is very large.

The Class So Far

- Between basis expansions and interaction terms, the number of candidate variables is very large.
- Given a set of models, we know how to compare them (AIC, BIC, cross validation).

The Class So Far

- Between basis expansions and interaction terms, the number of candidate variables is very large.
- Given a set of models, we know how to compare them (AIC, BIC, cross validation).
- Suppose we have p variables. How many possible models are there?

The Class So Far

- Between basis expansions and interaction terms, the number of candidate variables is very large.
- Given a set of models, we know how to compare them (AIC, BIC, cross validation).
- Suppose we have p variables. How many possible models are there? **Answer:** 2^p

1 The Class So Far

2 Stepwise Regression

3 Lasso

Forward Stepwise

Forward Stepwise

1. Start with a model with no variables.

Forward Stepwise

1. Start with a model with no variables.
2. Try adding each variable one by one to the model.

Forward Stepwise

1. Start with a model with no variables.
2. Try adding each variable one by one to the model.
3. Add the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)

Forward Stepwise

1. Start with a model with no variables.
2. Try adding each variable one by one to the model.
3. Add the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)
4. Next, try adding the remaining variables to the model one by one to the model.

Forward Stepwise

1. Start with a model with no variables.
2. Try adding each variable one by one to the model.
3. Add the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)
4. Next, try adding the remaining variables to the model one by one to the model.
5. Add the variable that improves the fit the most.

Forward Stepwise

1. Start with a model with no variables.
2. Try adding each variable one by one to the model.
3. Add the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)
4. Next, try adding the remaining variables to the model one by one to the model.
5. Add the variable that improves the fit the most.
6. Repeat this process until it is no longer possible to improve the fit.

Forward Stepwise

1. Start with a model with no variables.
2. Try adding each variable one by one to the model.
3. Add the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)
4. Next, try adding the remaining variables to the model one by one to the model.
5. Add the variable that improves the fit the most.
6. Repeat this process until it is no longer possible to improve the fit.

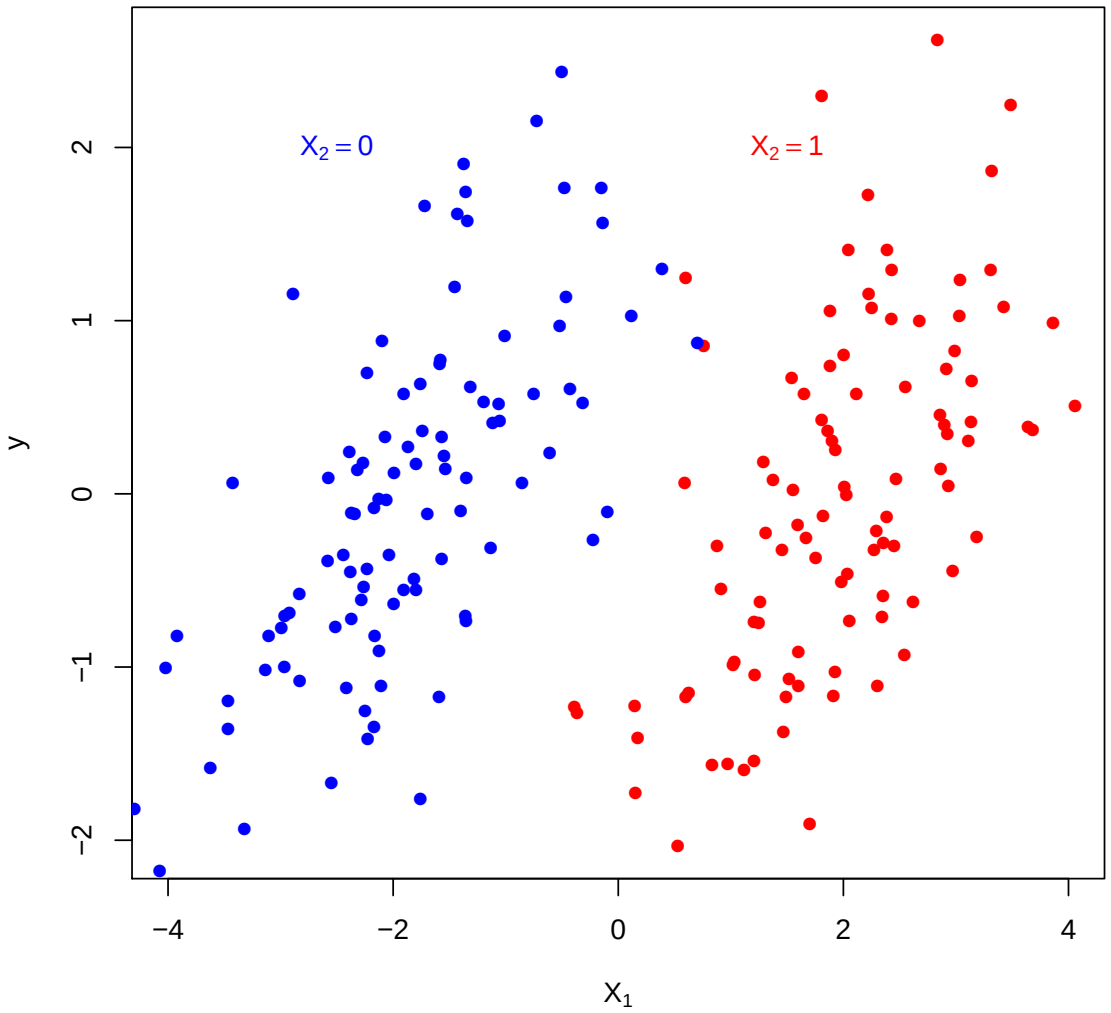
At the first stage, we consider p models. At the second stage, we consider $p-1$ models. And so on. In total, we consider:

$$p + (p-1) + (p-2) + ... + 1 = O(p^2) \text{ models}$$

Problems with Forward Stepwise

Problems with Forward Stepwise

It is possible for two variables, neither of which are useful on their own, to be useful jointly.



Backward Stepwise

Backward Stepwise

1. Start with *all* variables in the model.

Backward Stepwise

1. Start with *all* variables in the model.
2. Try deleting each variable one by one from the model.

Backward Stepwise

1. Start with *all* variables in the model.
2. Try deleting each variable one by one from the model.
3. Delete the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)

Backward Stepwise

1. Start with *all* variables in the model.
2. Try deleting each variable one by one from the model.
3. Delete the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)
4. Next, try deleting the remaining variables in the model one by one.

Backward Stepwise

1. Start with *all* variables in the model.
2. Try deleting each variable one by one from the model.
3. Delete the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)
4. Next, try deleting the remaining variables in the model one by one.
5. Delete the variable that improves the fit the most.

Backward Stepwise

1. Start with *all* variables in the model.
2. Try deleting each variable one by one from the model.
3. Delete the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)
4. Next, try deleting the remaining variables in the model one by one.
5. Delete the variable that improves the fit the most.
6. Repeat this process until it is no longer possible to improve the fit.

Backward Stepwise

1. Start with *all* variables in the model.
2. Try deleting each variable one by one from the model.
3. Delete the variable that improves the fit the most (as measured by AIC, cross-validation, etc.)
4. Next, try deleting the remaining variables in the model one by one.
5. Delete the variable that improves the fit the most.
6. Repeat this process until it is no longer possible to improve the fit.

At the first stage, we consider p models. At the second stage, we consider $p-1$ models. And so on. In total, we consider:

$$p + (p-1) + (p-2) + ... + 1 = O(p^2) \text{ models}$$

Problems with Backward Stepwise

If $p > n$ (i.e., more unknowns than equations), it may not be possible to fit *all* the variables to the model.

1 The Class So Far

2 Stepwise Regression

3 Lasso

Lasso

The lasso (Tibshirani 1996) finds $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p))^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Lasso

The lasso (Tibshirani 1996) finds $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p))^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

The penalty $\lambda \sum_{j=1}^{n} |\beta_j|$ will encourage the solution to be **sparse** (mostly zeroes).

Lasso

The lasso (Tibshirani 1996) finds $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p))^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

The penalty $\lambda \sum_{j=1}^{n} |\beta_j|$ will encourage the solution to be **sparse** (mostly zeroes).

What does this remind you of? How does it differ?

Lasso

The lasso (Tibshirani 1996) finds $\boldsymbol{\beta}$ that minimizes

$$\sum_{i=1}^{n}(y_i - (\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p))^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

The penalty $\lambda \sum_{j=1}^{n} |\beta_j|$ will encourage the solution to be **sparse** (mostly zeroes).

What does this remind you of? How does it differ?

Unlike AIC/BIC, we don't have to try all the possible combinations. This is a convex optimization problem that can be solved efficiently.

In-Class Exercise

```
from sklearn.linear_model import Lasso, lasso_path
```

In-Class Exercise

Use the Lasso to determine the best order polynomial to predict price from city-mpg. Either try different values of λ manually or fit the entire path of λ values (selecting the best one by cross-validation).

Problems with the Lasso

$$\sum_{i=1}^{n}(y_i - (\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p))^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

Problems with the Lasso

$$\sum_{i=1}^{n}(y_i - (\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p))^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- It selects the wrong model when the predictor variables are too correlated.

Problems with the Lasso

$$\sum_{i=1}^{n}(y_i - (\beta_0 + x_{i1}\beta_1 + ... + x_{ip}\beta_p))^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

- It selects the wrong model when the predictor variables are too correlated.
- It penalizes not only the number of non-zero coefficients but also their size, so it will tend to shrink the coefficients.

  Some people fix this problem by taking the model that Lasso selects and then refitting linear regression to just those variables.