Introduction to Machine Learning for Social Science

Class 6: LASSO Regression

Rochelle Terman

Postdoctoral Fellow
Center for International Security Cooperation
Stanford University

January 25th, 2018

Some Review

Evaluating Model Fit

- Evaluate fit with gold standard data
- In sample: dependent variable of model
- Out of sample: held out data

Assessing Classification Performance

Measures of classification performance

	Actual Label	
Guess	Yea	Nay
Yea	True Yea	False Yea
Nay	False Nay	True Nay

Assessing Classification Performance

Measures of classification performance

	Actual Label	
Guess	Yea	Nay
Yea	True Yea	False Yea
Nay	False Nay	True Nay

$$\text{Accuracy} = \frac{\text{TrueYea} + \text{TrueNay}}{\text{TrueYea} + \text{TrueNay} + \text{FalseYea} + \text{FalseNay}}$$

Assessing Classification Performance

Measures of classification performance

	Actual Label	
Guess	Yea	Nay
Yea	True Yea	False Yea
Nay	False Nay	True Nay

$$\text{Accuracy} = \frac{\text{TrueYea} + \text{TrueNay}}{\text{TrueYea} + \text{TrueNay} + \text{FalseYea} + \text{FalseNay}}$$

$$\text{Precision} = \frac{\text{True Yea}}{\text{True Yea} + \text{False Yea}}$$

Assessing Classification Performance

Measures of classification performance

	Actual Label	
Guess	Yea	Nay
Yea	True Yea	False Yea
Nay	False Nay	True Nay

$$\text{Accuracy} = \frac{\text{TrueYea} + \text{TrueNay}}{\text{TrueYea} + \text{TrueNay} + \text{FalseYea} + \text{FalseNay}}$$

$$\text{Precision} = \frac{\text{True Yea}}{\text{True Yea} + \text{False Yea}}$$

$$\text{Recall} = \frac{\text{True Yea}}{\text{True Yea} + \text{False Nay}}$$

Assessing Classification Performance

Measures of classification performance

	Actual Label	
Guess	Yea	Nay
Yea	True Yea	False Yea
Nay	False Nay	True Nay

$$\text{Accuracy} = \frac{\text{TrueYea} + \text{TrueNay}}{\text{TrueYea} + \text{TrueNay} + \text{FalseYea} + \text{FalseNay}}$$

$$\text{Precision} = \frac{\text{True Yea}}{\text{True Yea} + \text{False Yea}}$$

$$\text{Recall} = \frac{\text{True Yea}}{\text{True Yea} + \text{False Nay}}$$

$$F = \frac{2 \times \text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Supervised Learning $\rightsquigarrow$ Text analysis

Components to Supervised Learning Method

Components to Supervised Learning Method

1) Set of categories

Components to Supervised Learning Method

1) Set of categories
   - Credit Claiming, Position Taking, Advertising
   - Positive Tone, Negative Tone
   - Pro-war, Ambiguous, Anti-war

Components to Supervised Learning Method

1) Set of categories
    - Credit Claiming, Position Taking, Advertising
    - Positive Tone, Negative Tone
    - Pro-war, Ambiguous, Anti-war
2) Set of hand-coded documents

Components to Supervised Learning Method

1) Set of categories
   - Credit Claiming, Position Taking, Advertising
   - Positive Tone, Negative Tone
   - Pro-war, Ambiguous, Anti-war
2) Set of hand-coded documents
   - Coding done by human coders
   - Training Set: documents we'll use to learn how to code
   - Validation Set: documents we'll use to learn how well we code

Components to Supervised Learning Method

1) Set of categories
   - Credit Claiming, Position Taking, Advertising
   - Positive Tone, Negative Tone
   - Pro-war, Ambiguous, Anti-war
2) Set of hand-coded documents
   - Coding done by human coders
   - Training Set: documents we'll use to learn how to code
   - Validation Set: documents we'll use to learn how well we code
3) Set of unlabeled documents

Components to Supervised Learning Method

1) Set of categories
   - Credit Claiming, Position Taking, Advertising
   - Positive Tone, Negative Tone
   - Pro-war, Ambiguous, Anti-war
2) Set of hand-coded documents
   - Coding done by human coders
   - Training Set: documents we'll use to learn how to code
   - Validation Set: documents we'll use to learn how well we code
3) Set of unlabeled documents
4) Method to extrapolate from hand coding to unlabeled documents

Analyzing News Stories

New York Times Annotated Corpus

Analyzing News Stories

New York Times Annotated Corpus
November 1-3, 2004 (Day Before, Of, And After General Election)

Analyzing News Stories

New York Times Annotated Corpus

November 1-3, 2004 (Day Before, Of, And After General Election)

We've preprocessed the data $\rightsquigarrow$ Create a Document-Term Matrix

Goal: predict article from National desk (1) or other desk (0)
Method: LASSO Regression
Evaluation:

1) In Sample Accuracy
2) Out of Sample Accuracy

Key Terms:

- Overfitting
- Regularization
- LASSO
- Mean Squared Error (MSE)
- Bias-Variance Trade-off
- Cross validation

Key R Functions and Terms

- `glmnet`, `cv.glmnet`

Document-Term Matrices

$$X = \begin{array}{llllll} & \text{Word1} & \text{Word2} & \text{Word3} & \ldots & \text{WordP} \\ \text{Doc1} & 1 & 0 & 0 & \ldots & 3 \\ \text{Doc2} & 0 & 2 & 1 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ \text{DocN} & 0 & 0 & 0 & \ldots & 5 \end{array}$$

Document-Term Matrices

$$
\boldsymbol{X} = \begin{array}{llllll}
 & \text{Word1} & \text{Word2} & \text{Word3} & \ldots & \text{WordP} \\
\text{Doc1} & 1 & 0 & 0 & \ldots & 3 \\
\text{Doc2} & 0 & 2 & 1 & \ldots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \\
\text{DocN} & 0 & 0 & 0 & \ldots & 5
\end{array}
$$

$\boldsymbol{X} = N \times P$ matrix

- N = Number of documents

Document-Term Matrices

$$
\boldsymbol{X} = \begin{array}{llllll}
 & \text{Word1} & \text{Word2} & \text{Word3} & \dots & \text{WordP} \\
\text{Doc1} & 1 & 0 & 0 & \dots & 3 \\
\text{Doc2} & 0 & 2 & 1 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \\
\text{DocN} & 0 & 0 & 0 & \dots & 5
\end{array}
$$

$\boldsymbol{X} = N \times P$ matrix

- $N =$ Number of documents
- $P =$ Number of features

Document-Term Matrices

$$\boldsymbol{X} = \begin{array}{llllll} & \text{Word1} & \text{Word2} & \text{Word3} & \dots & \text{WordP} \\ \text{Doc1} & 1 & 0 & 0 & \dots & 3 \\ \text{Doc2} & 0 & 2 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ \text{DocN} & 0 & 0 & 0 & \dots & 5 \end{array}$$

$\boldsymbol{X} = N \times P$ matrix

- $N =$ Number of documents
- $P =$ Number of features
- $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{iP})$

Document-Term Matrices

$$
\boldsymbol{X} = \begin{array}{llllll}
 & \text{Word1} & \text{Word2} & \text{Word3} & \dots & \text{WordP} \\
\text{Doc1} & 1 & 0 & 0 & \dots & 3 \\
\text{Doc2} & 0 & 2 & 1 & \dots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \\
\text{DocN} & 0 & 0 & 0 & \dots & 5
\end{array}
$$

$\boldsymbol{X} = N \times P$ matrix

- $N =$ Number of documents
- $P =$ Number of features
- $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{iP})$

Let $p = (\Pr(\text{Desk}_i = 1))$

$$\text{logit}(p) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_P X_P$$

Regression with Many Predictors

Rules for regression coefficients to exist:

Regression with Many Predictors

Rules for regression coefficients to exist:

1) Number of observations (N) $>$ Number of predictors (P)

Regression with Many Predictors

Rules for regression coefficients to exist:

1) Number of observations (N) $>$ Number of predictors (P)
2) Predictors $\rightsquigarrow$ distinct (i.e. not highly correlated)

Regression with Many Predictors

Rules for regression coefficients to exist:

1) Number of observations (N) $>$ Number of predictors (P)
2) Predictors $\rightsquigarrow$ distinct (i.e. not highly correlated)

If (1) and (2) are close to false, predictions become highly variable

Regression with Many Predictors

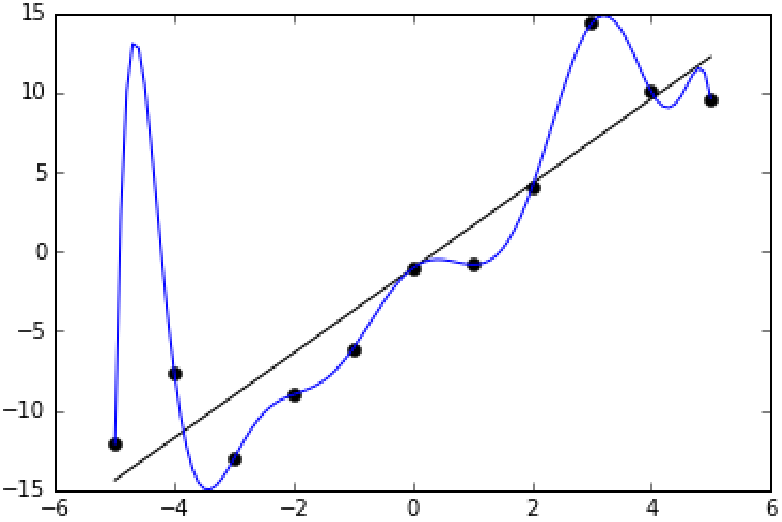
Rules for regression coefficients to exist:

1) Number of observations (N) $>$ Number of predictors (P)
2) Predictors $\rightsquigarrow$ distinct (i.e. not highly correlated)

If (1) and (2) are close to false, predictions become highly variable

$\rightsquigarrow$ overfitting.

Overfitting



R Code

Overcoming overfitting

1.) Reduce number of features.
   - Manually select which features to keep.
   - Model selection algorithm.

Overcoming overfitting

1.) Reduce number of features.
   - Manually select which features to keep.
   - Model selection algorithm.

2.) Regularization
   - Keep all features, but shrink magnitude / value of parameters close to zero (ridge).
   - Keep all features, but shrink magnitude / value of (some) parameters to zero (lasso).

LASSO Regression

LASSO Regression

LASSO ("least absolute shrinkage and selection operator"): a regularization procedure that shrinks regression coefficients toward zero.

LASSO Regression

LASSO ("least absolute shrinkage and selection operator"): a regularization procedure that shrinks regression coefficients toward zero.

Document i, $(i = 1, \ldots, N)$

LASSO Regression

LASSO ("least absolute shrinkage and selection operator"): a regularization procedure that shrinks regression coefficients toward zero.

Document i, $(i = 1, \ldots, N)$
Count vector $\boldsymbol{x}_i$

LASSO Regression

LASSO ("least absolute shrinkage and selection operator"): a regularization procedure that shrinks regression coefficients toward zero.

Document i, $(i = 1, \ldots, N)$
Count vector $\boldsymbol{x}_i$
Labels Y_i

LASSO Regression

LASSO ("least absolute shrinkage and selection operator"): a regularization procedure that shrinks regression coefficients toward zero.

Document i, $(i = 1, \ldots, N)$
Count vector $\boldsymbol{x}_i$
Labels Y_i
Linear regression: Choose $\beta's$ to minimize sum of squared residuals

$$\boldsymbol{\beta}_{\text{OLS}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{N} (Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2$$

LASSO Regression

LASSO ("least absolute shrinkage and selection operator"): a regularization procedure that shrinks regression coefficients toward zero.

Document i, $(i = 1, \ldots, N)$
Count vector $\boldsymbol{x}_i$
Labels Y_i
LASSO Regression: Choose $\beta's$ to minimize sum of squared residuals and penalty on size of coefficients

LASSO Regression

LASSO ("least absolute shrinkage and selection operator"): a regularization procedure that shrinks regression coefficients toward zero.

Document i, $(i = 1, \ldots, N)$
Count vector $\boldsymbol{x}_i$
Labels Y_i
LASSO Regression: Choose $\boldsymbol{\beta}'s$ to minimize sum of squared residuals and penalty on size of coefficients

$$\boldsymbol{\beta}_{\text{LASSO}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{N} (Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2 + \lambda \underbrace{\sum_{p=1}^{P} |\beta_p|}_{\text{penalty}}$$

LASSO Regression

LASSO ("least absolute shrinkage and selection operator"): a regularization procedure that shrinks regression coefficients toward zero.

Document i, $(i = 1, \ldots, N)$
Count vector $\boldsymbol{x}_i$
Labels Y_i
LASSO Regression: Choose $\boldsymbol{\beta}'s$ to minimize sum of squared residuals and penalty on size of coefficients

$$\boldsymbol{\beta}_{\text{LASSO}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{N} (Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2 + \lambda \underbrace{\sum_{p=1}^{P} |\beta_p|}_{\text{penalty}}$$

What does λ do?

LASSO Penalty: Algebra

Why does LASSO shrink coefficients to 0?

LASSO Penalty: Algebra

Why does LASSO shrink coefficients to 0?

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

LASSO Penalty: Algebra

Why does LASSO shrink coefficients to 0?

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Penalty for Coefficients:

LASSO Penalty: Algebra

Why does LASSO shrink coefficients to 0?

Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Penalty for Coefficients:

$$\sum_{j=1}^{2} |\beta_j| \quad = \quad \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$$

LASSO Penalty: Algebra

Why does LASSO shrink coefficients to 0?

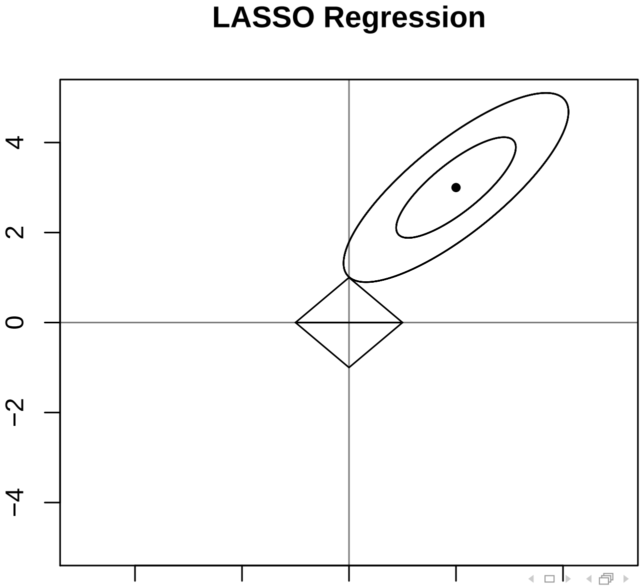
Contrast $\beta = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}})$ and $\tilde{\beta} = (1, 0)$

Penalty for Coefficients:

$$\sum_{j=1}^{2} |\beta_j| = \frac{1}{\sqrt{2}} + \frac{1}{\sqrt{2}} = \sqrt{2}$$

$$\sum_{j=1}^{2} |\tilde{\beta}_j| = 1 + 0 = 1$$

LASSO Penalty: Geometry



R Code!

Methods/Metrics for:

1) Choosing λ
2) Assessing model performance