150B/355B
Introduction to Machine Learning for Social Science
TA Section 4

Haemin Jee and Tongtong Zhang

February 2, 2018

Road Map

1. Overfitting

Road Map

1. Overfitting
2. Lasso Regression

Road Map

1. Overfitting
2. Lasso Regression
   - Cross Validation

Road Map

1. Overfitting
2. Lasso Regression
   - Cross Validation
   - Bias-Variance Tradeoff

Road Map

1. Overfitting
2. Lasso Regression
   - Cross Validation
   - Bias-Variance Tradeoff
3. R code in lectures this week

Document-Term Matrices

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

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

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

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

Overfitting

Overfitting means your model fits the training data too well such that it starts to "memorize" the training data rather than "learn" to generalize from its trend.

Overfitting

Overfitting means your model fits the training data too well such that it starts to "memorize" the training data rather than "learn" to generalize from its trend.

When overfitting happens, small perturbation in the training data can lead to substantial change in your model coefficients and thereby, lead to substantial change in your the predictions on the test set.

Overfitting

Reasons for overfitting:

Overfitting

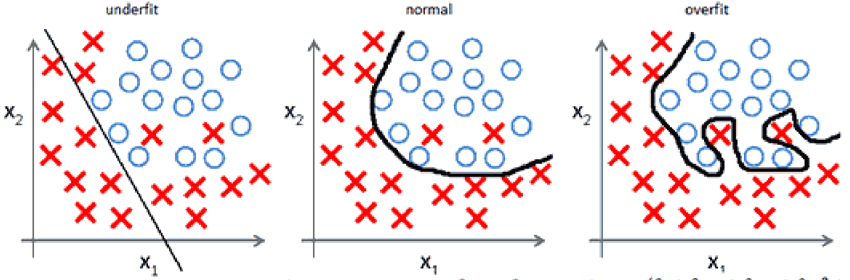
Reasons for overfitting:

1. Number of observations (N) $<$ Number of predictors (P)

Overfitting

Reasons for overfitting:

1. Number of observations (N) < Number of predictors (P)



$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$

$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2 + \theta_6 x_1^2 x_2 + \theta_7 x_1 x_2^2 + \theta_8 x_1^2 x_2^2 + \theta_9 x_1^3 + \dots)$

Overfitting

Reasons for overfitting:

Overfitting

Reasons for overfitting:

2. Predictors are highly correlated

Overfitting

Reasons for overfitting:

2. Predictors are highly correlated -the error term (random noise) will have substantial effect on the model.

Overfitting

Reasons for overfitting:

2. Predictors are highly correlated -the error term (random noise) will have substantial effect on the model.

Suppose we have two highly correlated predictors, X_1 and X_2.

```
base<-1:100
set.seed(12345)
X1<-base+rnorm(100,0,0.01)
#both x1 and x2 have the same base with a little noise
X2<-base+rnorm(100,0,0.01)
cor(X1,X2)

## [1] 0.9999999
```

Overfitting

Reasons for overfitting: Predictors are highly correlated

Overfitting

Reasons for overfitting: Predictors are highly correlated

The true model is $Y = X1 + X2 + \epsilon$, where ϵ is some random noise.

Overfitting

Reasons for overfitting: Predictors are highly correlated

The true model is $Y = X1 + X2 + \epsilon$, where ϵ is some random noise.

```
set.seed(1234)
Y<-X1+X2+rnorm(100,0,1) #1st random sample
lm(Y~X1+X2)$coefficients

## (Intercept)          X1          X2
##  -0.6004386   4.0852277  -2.0765643

set.seed(123456) # at a different seed, error term is different
Y<-X1+X2+rnorm(100,0,1) #2nd random sample with small changes in Y
lm(Y~X1+X2)$coefficients

## (Intercept)          X1          X2
##   0.2576408 -14.6233034  16.6191533
```

Overfitting

Reasons for overfitting: Predictors are highly correlated

The true model is $Y = X1 + X2 + \epsilon$, where ϵ is some random noise.

```
set.seed(1234)
Y<-X1+X2+rnorm(100,0,1) #1st random sample
lm(Y~X1+X2)$coefficients

## (Intercept)          X1          X2
##  -0.6004386   4.0852277  -2.0765643

set.seed(123456) # at a different seed, error term is different
Y<-X1+X2+rnorm(100,0,1) #2nd random sample with small changes in Y
lm(Y~X1+X2)$coefficients

## (Intercept)          X1          X2
##   0.2576408 -14.6233034  16.6191533
```

With only small perturbation in the error term (Y), we got very different coefficients!

LASSO Regression

Overcoming Overfitting $\rightsquigarrow$ LASSO Regression

LASSO Regression

Overcoming Overfitting $\rightsquigarrow$ LASSO Regression

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

LASSO Regression

Overcoming Overfitting $\rightsquigarrow$ 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

Overcoming Overfitting $\rightsquigarrow$ 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

Overcoming Overfitting $\rightsquigarrow$ 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

Overcoming Overfitting $\rightsquigarrow$ 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

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

LASSO Regression

Overcoming Overfitting $\rightsquigarrow$ 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,
subject to a constraint on coefficients:

$$\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} \left(Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i\right)^2, \text{subject to} \sum_{p=1}^{P} |\beta_p| \leq t$$

LASSO Regression

Overcoming Overfitting $\rightsquigarrow$ 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, subject to a constraint on coefficients:

$$\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} \left(Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i \right)^2, \text{subject to} \sum_{p=1}^{P} |\beta_p| \leq t$$

Re-write in the Lagrangian form:

LASSO Regression

Overcoming Overfitting $\rightsquigarrow$ 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, subject to a constraint on coefficients:

$$\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} (Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2, \text{subject to} \sum_{p=1}^{P} |\beta_p| \leq t$$

Re-write in the Lagrangian form:

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

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the mean squared error (MSE) / Loss function.

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the mean squared error (MSE) / Loss function.

$$
\begin{aligned}
\widehat{\boldsymbol{\beta}}^{\lambda} &= \text{Coefficients at } \lambda \\
\widehat{p}_{i,\lambda} &= \Pr(Y_i = 1 | \boldsymbol{X}_i, \widehat{\boldsymbol{\beta}}^{\lambda}) \rightarrow \text{Prediction}
\end{aligned}
$$

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the mean squared error (MSE) / Loss function.

$$
\begin{aligned}
\widehat{\beta}^{\lambda} &= \text{Coefficients at } \lambda \\
\widehat{p}_{i,\lambda} &= \Pr(Y_i = 1 | \boldsymbol{X}_i, \widehat{\beta}^{\lambda}) \rightarrow \text{Prediction} \\
\text{MSE} &= \frac{\sum_{i=1}^{N} (Y_i - \widehat{p}_{i,\lambda})^2}{N}
\end{aligned}
$$

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the MSE on out-of-sample data

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the MSE on out-of-sample data

In practice, try out a set of λ. For each candidate λ,

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the MSE on out-of-sample data

In practice, try out a set of λ. For each candidate λ,

1. Get $\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} \left(Y_i - \beta \cdot \boldsymbol{x}_i\right)^2 + \lambda \sum_{p=1}^{P} |\beta_p|$ in the training data

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the MSE on out-of-sample data

In practice, try out a set of λ. For each candidate λ,

1. Get $\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} \left(Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i \right)^2 + \lambda \sum_{p=1}^{P} |\beta_p|$ in the training data
2. Evaluate the model on the test data, record MSE

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the MSE on out-of-sample data

In practice, try out a set of λ. For each candidate λ,

1. Get $\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} (Y_i - \beta \cdot \boldsymbol{x}_i)^2 + \lambda \sum_{p=1}^{P} |\beta_p|$ in the training data
2. Evaluate the model on the test data, record MSE
3. Repeat steps 1-2 for each candidate λ

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\beta}$

How do we choose λ?

Find λ that minimizes the MSE on out-of-sample data

In practice, try out a set of λ. For each candidate λ,

1. Get $\beta_{\text{LASSO}} = \text{argmin}_{\beta} \sum_{i=1}^{N} (Y_i - \beta \cdot \boldsymbol{x}_i)^2 + \lambda \sum_{p=1}^{P} |\beta_p|$ in the training data
2. Evaluate the model on the test data, record MSE
3. Repeat steps 1-2 for each candidate λ
4. Pick λ^* that returns the lowest MSE

LASSO Regression

Each λ corresponds to a set of coefficients $\hat{\boldsymbol{\beta}}$

How do we choose λ?

Find λ that minimizes the MSE on out-of-sample data

In practice, try out a set of λ. For each candidate λ,

1. Get $\beta_{\text{LASSO}} = \text{argmin}_{\boldsymbol{\beta}} \sum_{i=1}^{N} (Y_i - \boldsymbol{\beta} \cdot \boldsymbol{x}_i)^2 + \lambda \sum_{p=1}^{P} |\beta_p|$ in the training data
2. Evaluate the model on the test data, record MSE
3. Repeat steps 1-2 for each candidate λ
4. Pick λ^* that returns the lowest MSE

In practice, we use cross-validation to complete these procedures

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, ..., Group K	Group 1
2	Group 1, Group3, Group 4, ..., Group K	Group 2
3	Group 1, Group 2, Group 4, ..., Group K	Group 3
$\vdots$	$\vdots$	$\vdots$
K	Group 1, Group 2, Group 3, ..., Group K - 1	Group K

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, . . ., Group K	Group 1
2	Group 1, Group3, Group 4, . . ., Group K	Group 2
3	Group 1, Group 2, Group 4, . . ., Group K	Group 3
$\vdots$	$\vdots$	$\vdots$
K	Group 1, Group 2, Group 3, . . ., Group K - 1	Group K

Strategy:

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, . . ., Group K	Group 1
2	Group 1, Group3, Group 4, . . ., Group K	Group 2
3	Group 1, Group 2, Group 4, . . ., Group K	Group 3
$\vdots$	$\vdots$	$\vdots$
K	Group 1, Group 2, Group 3, . . ., Group K - 1	Group K

Strategy:

- Randomly divide data into K groups

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, ..., Group K	Group 1
2	Group 1, Group3, Group 4, ..., Group K	Group 2
3	Group 1, Group 2, Group 4, ..., Group K	Group 3
$\vdots$	$\vdots$	$\vdots$
K	Group 1, Group 2, Group 3, ..., Group K - 1	Group K

Strategy:

- Randomly divide data into K groups
- Train data on $K-1$ groups.

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, ..., Group K	Group 1
2	Group 1, Group3, Group 4, ..., Group K	Group 2
3	Group 1, Group 2, Group 4, ..., Group K	Group 3
$\vdots$	$\vdots$	$\vdots$
K	Group 1, Group 2, Group 3, ..., Group K - 1	Group K

Strategy:

- Randomly divide data into K groups
- Train data on $K - 1$ groups.
- Predict values for K^{th}

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, ..., Group K	Group 1
2	Group 1, Group3, Group 4, ..., Group K	Group 2
3	Group 1, Group 2, Group 4, ..., Group K	Group 3
$\vdots$	$\vdots$	$\vdots$
K	Group 1, Group 2, Group 3, ..., Group K - 1	Group K

Strategy:

- Randomly divide data into K groups
- Train data on $K - 1$ groups.
- Predict values for K^{th}
- Summarize performance with loss function:

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, . . ., Group K	Group 1
2	Group 1, Group3, Group 4, . . ., Group K	Group 2
3	Group 1, Group 2, Group 4, . . ., Group K	Group 3
⋮	⋮	⋮
K	Group 1, Group 2, Group 3, . . ., Group K - 1	Group K

Strategy:

- Randomly divide data into K groups
- Train data on $K - 1$ groups.
- Predict values for K^{th}
- Summarize performance with loss function:
  - MSE, Accuracy, Prediction error, ...

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, . . ., Group K	Group 1
2	Group 1, Group3, Group 4, . . ., Group K	Group 2
3	Group 1, Group 2, Group 4, . . ., Group K	Group 3
$\vdots$	$\vdots$	$\vdots$
K	Group 1, Group 2, Group 3, . . ., Group K - 1	Group K

Strategy:

- Randomly divide data into K groups
- Train data on $K - 1$ groups.
- Predict values for K^{th}
- Summarize performance with loss function:
    - MSE, Accuracy, Prediction error, ...
- Final choice: model with optimal performance / CV score

LASSO Regression in Practice

K-fold Cross Validation

Iteration	Training	Validation ("Test")
1	Group2, Group3, Group 4, ..., Group K	Group 1
2	Group 1, Group3, Group 4, ..., Group K	Group 2
3	Group 1, Group 2, Group 4, ..., Group K	Group 3
⋮	⋮	⋮
K	Group 1, Group 2, Group 3, ..., Group K - 1	Group K

Strategy:

- Randomly divide data into K groups
- Train data on $K - 1$ groups.
- Predict values for K^{th}
- Summarize performance with loss function:
  - MSE, Accuracy, Prediction error, ...
- Final choice: model with optimal performance / CV score

Common K's: 5-fold, 10-fold, N (LOOCV)

LASSO Regression

Why don't we choose λ by evaluating the model on in-sample/ training data?

LASSO Regression

Why don't we choose λ by evaluating the model on in-sample/ training data?

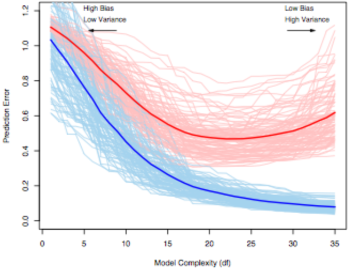


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

As we add more features to the model,

- Bias decreases $\rightarrow$ we have perfect in-sample fit

LASSO Regression

Why don't we choose λ by evaluating the model on in-sample/ training data?

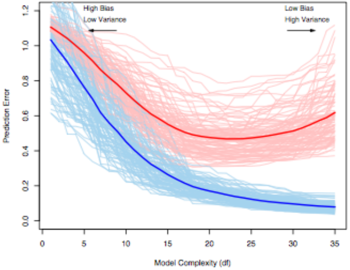


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\mathrm{err}}$*, while the light red curves show the conditional test error* $\mathrm{Err}_{\mathcal{T}}$ *for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\mathrm{E}[\overline{\mathrm{err}}]$.

As we add more features to the model,

- Bias decreases $\rightarrow$ we have perfect in-sample fit
- Variance increases $\rightarrow$ our model is too specific to the training data and we have bad out-of-sample fit

LASSO Regression

Why don't we choose λ by evaluating the model on in-sample/ training data?

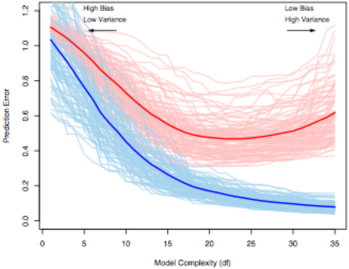


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$*, while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

As we add more features to the model,

- Bias decreases $\rightarrow$ we have perfect in-sample fit
- Variance increases $\rightarrow$ our model is too specific to the training data and we have bad out-of-sample fit
- LASSO finds the sweet spot between bias and variance:

LASSO Regression

Why don't we choose λ by evaluating the model on in-sample/ training data?

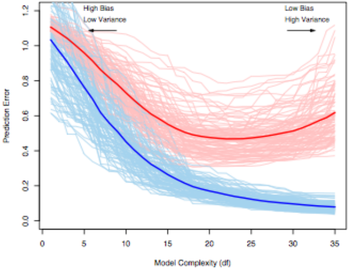


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$, *while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

As we add more features to the model,

- Bias decreases $\rightarrow$ we have perfect in-sample fit
- Variance increases $\rightarrow$ our model is too specific to the training data and we have bad out-of-sample fit
- LASSO finds the sweet spot between bias and variance:
  - Introduce a bit bias by constraining β small

LASSO Regression

Why don't we choose λ by evaluating the model on in-sample/ training data?

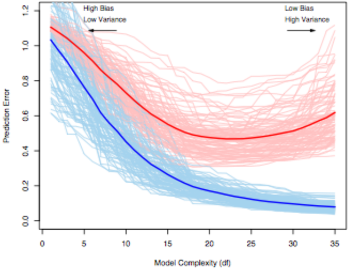


FIGURE 7.1. *Behavior of test sample and training sample error as the model complexity is varied. The light blue curves show the training error* $\overline{\text{err}}$, *while the light red curves show the conditional test error* $\text{Err}_{\mathcal{T}}$ *for 100 training sets of size 50 each, as the model complexity is increased. The solid curves show the expected test error* Err *and the expected training error* $\text{E}[\overline{\text{err}}]$.

As we add more features to the model,

- Bias decreases $\rightarrow$ we have perfect in-sample fit
- Variance increases $\rightarrow$ our model is too specific to the training data and we have bad out-of-sample fit
- LASSO finds the sweet spot between bias and variance:
  - Introduce a bit bias by constraining β small
  - Reduce variance by minimizing MSE on out-of-sample data.

R code!

Big Picture Review

What we have been doing?

Big Picture Review

What we have been doing?

- Task: build a model of label $\sim$ features using training data (hand-labeled) to predict the label of out-of-sample data.

Big Picture Review

What we have been doing?

- Task: build a model of label $\sim$ features using training data (hand-labeled) to predict the label of out-of-sample data.
- To do so, we need to determine the features / predictors in our model.

Big Picture Review

What we have been doing?

- Task: build a model of label $\sim$ features using training data (hand-labeled) to predict the label of out-of-sample data.
- To do so, we need to determine the features / predictors in our model.
- While choosing features, we want to avoid overfitting (too many features, too little generalizability)

Big Picture Review

What we have been doing?

- Task: build a model of label $\sim$ features using training data (hand-labeled) to predict the label of out-of-sample data.
- To do so, we need to determine the features / predictors in our model.
- While choosing features, we want to avoid overfitting (too many features, too little generalizability)
- Solution: LASSO regression

Big Picture Review

What we have been doing?

- Task: build a model of label $\sim$ features using training data (hand-labeled) to predict the label of out-of-sample data.
- To do so, we need to determine the features / predictors in our model.
- While choosing features, we want to avoid overfitting (too many features, too little generalizability)
- Solution: LASSO regression
  1. Pick λ^* that minimizes MSE on test set

Big Picture Review

What we have been doing?

- Task: build a model of label $\sim$ features using training data (hand-labeled) to predict the label of out-of-sample data.
- To do so, we need to determine the features / predictors in our model.
- While choosing features, we want to avoid overfitting (too many features, too little generalizability)
- Solution: LASSO regression
  1. Pick λ^* that minimizes MSE on test set
  2. Choose β given λ^*

Big Picture Review

What we have been doing?

- Task: build a model of label $\sim$ features using training data (hand-labeled) to predict the label of out-of-sample data.
- To do so, we need to determine the features / predictors in our model.
- While choosing features, we want to avoid overfitting (too many features, too little generalizability)
- Solution: LASSO regression
  1. Pick λ^* that minimizes MSE on test set
  2. Choose β given λ^*
  3. We have our model!