Ridge and Lasso - Hyperparameters

CS109A Introduction to Data Science

Pavlos Protopapas, Natesh Pillai

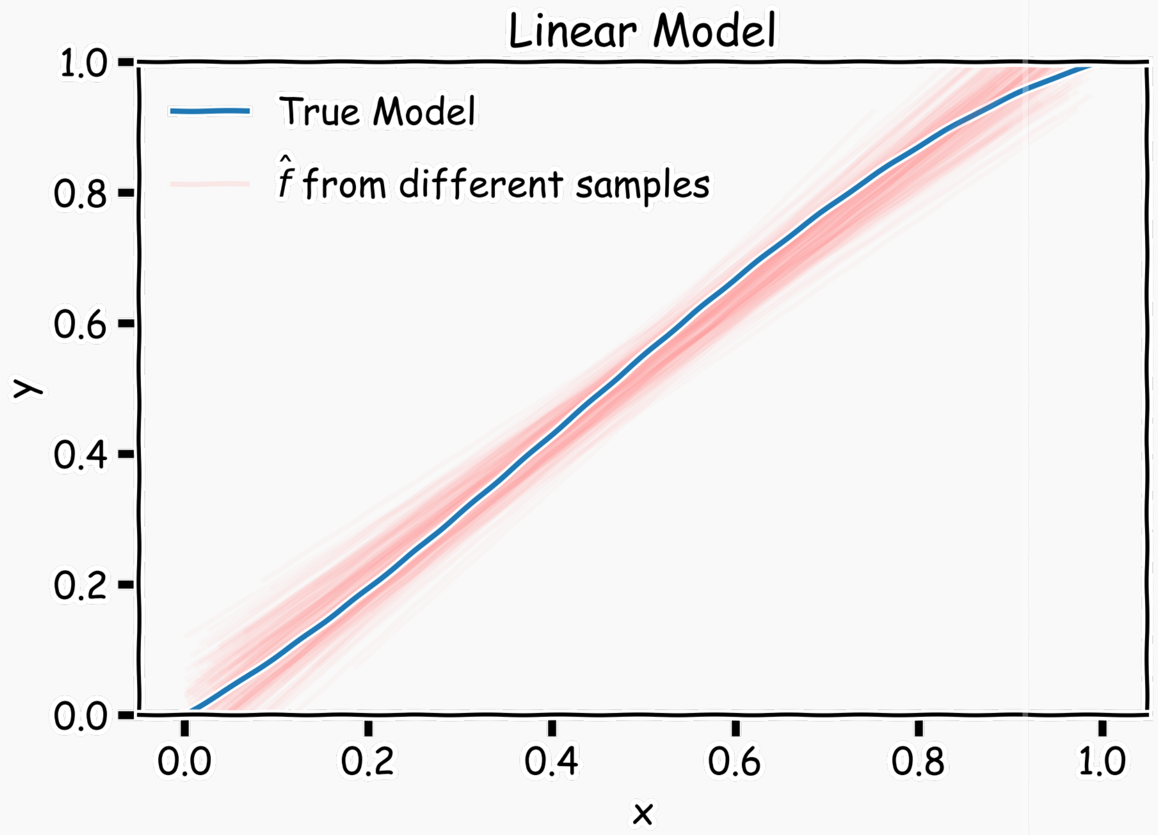
Outline

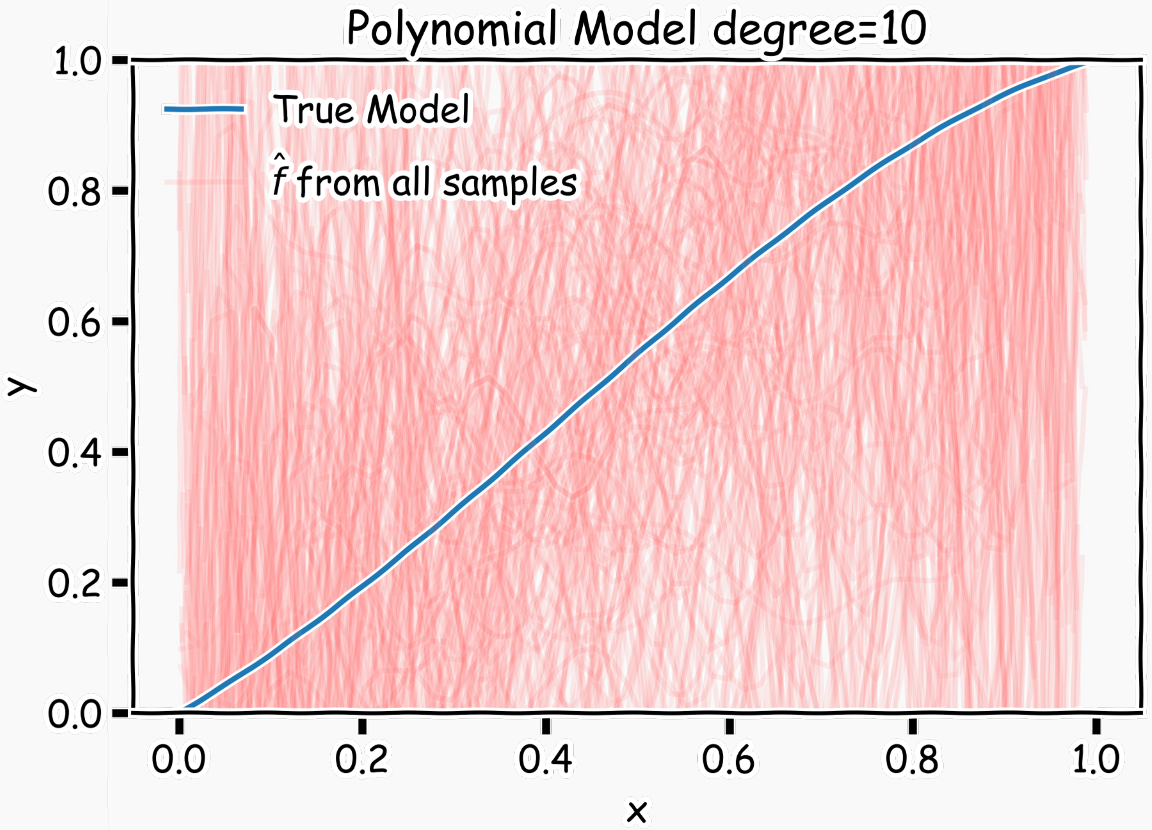
- Q&A from lecture 5:
  - Train/Validation/Test
  - Scaling
- Generalization Error, Bias Variance Tradeoff
- Regularization
  - Lasso and Ridge

Bias vs Variance

Left: 2000 best fit straight lines, each fitted on a different 20 point training set.

Right: Best-fit models using degree 10 polynomial

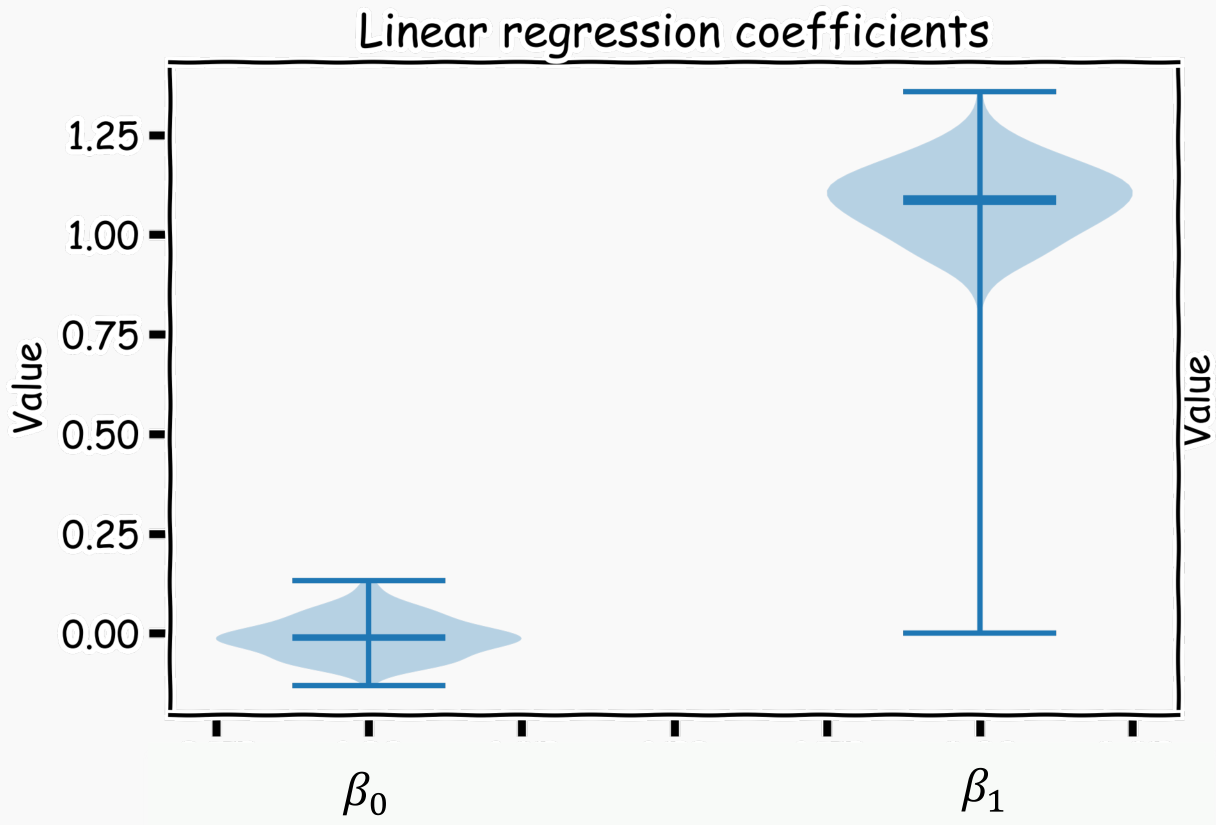




Bias vs Variance

Left: Linear regression coefficients

Right: Poly regression of order 10 coefficients



Model Selection

?

Model selection is the application of a principled method to determine the complexity of the model, e.g., choosing a subset of predictors, choosing the degree of the polynomial model etc.

A strong m … id
overfitting

How do we discourage extreme values in the model parameters?

- there are
    - the
    - the polynomial degree is too high
    - too many cross terms are considered
- the coefficients values are too **extreme**

Regularization

What we want

Low model error.

Minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2$$

Discourage extreme values in model parameters.

Minimize:

Regularization

What we want

Low model error.

Minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2$$

Discourage extreme values in model parameters.

Minimize:

$$L_{reg} = \begin{cases} \sum_{j=1}^{J}\beta_j^2 \\ \sum_{j=1}^{J}|\beta_j| \end{cases}$$

Regularization

What we want

Low model error.

Minimize:

$$\frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2$$

Discourage extreme values in model parameters.

Minimize:

$$L_{reg} = \begin{cases} \sum_{j=1}^{J} \beta_j^2 \\ \sum_{j=1}^{J} |\beta_j| \end{cases}$$

How do we combine these two objectives?

Regularization

What we want

Low model error.

Minimize:

Discourage extreme values in model parameters.

Minimize:

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + L_{reg}$$

Regularization

What we want

Low model error.

Minimize:

Discourage extreme values in model parameters.

[...]mize:

λ is the **regularization parameter**. It controls the relative importance between model error and regularization term

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda\ L_{reg}$$

Regularization

What we want

Low model error.
Low model error.

Discourage extreme values in model parameters.

mize:

$\lambda = 0$: equivalent to simple linear regression

$\lambda = \infty$: yields a model with $\beta's$ =0

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda\, L_{reg}$$

Regularization

What we want

Low model error.

Minimize:

Discourage extreme values in model parameters.

Minimize:

How do we determine λ?

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda L_{reg}$$

Regularization

What we want

Low model error.

Minimize:

Discourage extreme values in model parameters.

Minimize:

Cross Validation!

$$\mathcal{L}_{REG} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda L_{reg}$$

Regularization: LASSO Regression

What we want

Low model error.

Minimize:

Discourage extreme values in model parameters.

Minimize:

Note that $\sum_{j=1}^{J} |\beta_j|$ is the l_1 norm of the vector $\boldsymbol{\beta}$

$$\mathcal{L}_{LASSO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

Regularization: **LASSO** Regression

What we want

Low model error.

Discourage extreme values in
parameters.

ize:

No need to regularize the bias, β_0
Why?

$$\mathcal{L}_{LASSO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda\sum_{j=1}^{J}|\beta_j|$$

Regularization: LASSO Regression

Lasso regression: minimize $\mathcal{L}_{LASSO}$ with respect to $\beta' s$

$$\mathcal{L}_{LASSO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda \sum_{j=1}^{J} |\beta_j|$$

Regularization: Ridge Regression

Ridge regression: minimize $\mathcal{L}_{RIDGE}$ with resp

Note that $\sum_{j=1}^{J} \beta_j^2$ is the $\boldsymbol{l_2}$ norm of the vector $\boldsymbol{\beta}$

$$\mathcal{L}_{LASSO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^{\top}\boldsymbol{x}_i\right|^2 + \lambda \sum_{j=1}^{J}\beta_j^2$$

Regularization: Ridge Regression

Ridge regression: minimize $\mathcal{L}_{RIDGE}$ with respect to $\beta's$

$$\mathcal{L}_{LASSO} = \frac{1}{n}\sum_{i=1}^{n}\left|y_i - \boldsymbol{\beta}^\top \boldsymbol{x}_i\right|^2 + \lambda \sum_{j=1}^{J}\beta_j^2$$

No need to regularize the bias, β_0, since is not connected to the predictors.

Ridge regularization with only **validation** : step by step

For ridge regression there exist an analytical solution for the coefficients:

$$\hat{\beta}_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$$

1. split data into $\{\{X,Y\}_{train}, \{X,Y\}_{validation}, \{X,Y\}_{test}\}$
2. for λ in $\{\lambda_{min}, \ldots \lambda_{max}\}$:
    1. determine the β that minimizes the L_{ridge}, $\beta_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$, using the train data.
    2. record $L_{MSE}(\lambda)$ using validation data.
3. select the λ that minimizes the MSE loss on the validation data,

$$\lambda_{ridge} = \mathrm{argmin}_{\lambda}\, L_{MSE}(\lambda)$$

4. Refit the model using both train and validation data, $\{\{X,Y\}_{train}, \{X,Y\}_{validation}\}$, now using λ_{ridge}, resulting to $\hat{\beta}_{ridge}(\lambda_{ridge})$
5. Report MSE or R^2 on $\{X,Y\}_{test}$ given the $\hat{\beta}_{ridge}(\lambda_{ridge})$

Ridge regularization with only **validation** : step by step

For ridge regression there exist an analytical solution for the coefficients:

$$\hat{\beta}_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$$

1. split data into $\{\{X, Y\}_{train}, \{X, Y\}_{validation}, \{X, Y\}_{test}\}$
2. for λ in $\{\lambda_{min}, \ldots \lambda_{max}\}$:
    1. determine the β that minimizes the L_{ridge}, $\beta_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$, using the train data.
    2. record $L_{MSE}(\lambda)$ using validation data.
3. select the λ that minimizes the MSE loss on the validation data,

$$\lambda_{ridge} = \mathrm{argmin}_{\lambda}\, L_{MSE}(\lambda)$$

4. Refit the model using both train and validation data, $\{\{X, Y\}_{train}, \{X, Y\}_{validation}\}$, now using λ_{ridge}, resulting to $\hat{\beta}_{ridge}(\lambda_{ridge})$
5. Report MSE or R^2 on $\{X, Y\}_{test}$ given the $\hat{\beta}_{ridge}(\lambda_{ridge})$

Ridge regularization with only **validation** : step by step

For ridge regression there exist an analytical solution for the coefficients:

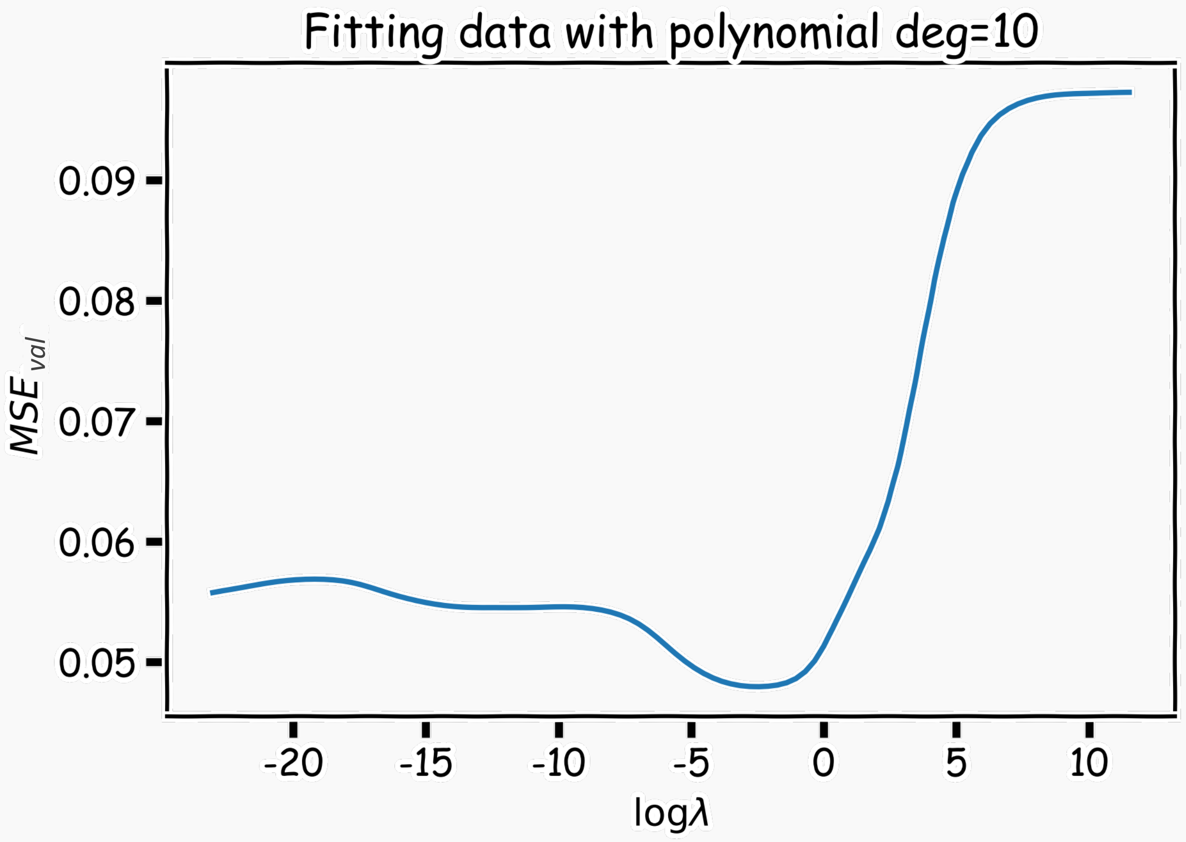
$$\hat{\beta}_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$$

1. split data into $\{\{X,Y\}_{train}, \{X,Y\}_{validation}, \{X,Y\}_{test}\}$
2. for λ in $\{\lambda_{min}, \dots \lambda_{max}\}$:
    1. determine the β that minimizes the L_{ridge}, $\beta_{Ridge}(\lambda) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$, using the train data.
    2. record $L_{MSE}(\lambda)$ using validation data.
3. select the λ that minimizes the *MSE* loss on the validation data,

$$\lambda_{ridge} = \mathrm{argmin}_{\lambda}\, L_{MSE}(\lambda)$$

4. Refit the model using both train and validation data, $\{\{X,Y\}_{train}, \{X,Y\}_{validation}\}$, now using λ_{ridge}, resulting to $\hat{\beta}_{ridge}(\lambda_{ridge})$
5. Report MSE or R^2 on $\{X,Y\}_{test}$ given the $\hat{\beta}_{ridge}(\lambda_{ridge})$

Ridge regularization with **validation** only



Lasso regularization with **validation** only: step by step

For Lasso regression, there is **no** analytical solution for the coefficients, so we use a **solver**.

1. split data into $\{\{X, Y\}_{train}, \{X, Y\}_{validation}, \{X, Y\}_{test}\}$
2. for λ in $\{\lambda_{min}, \ldots \lambda_{max}\}$:
   A. determine the β that minimizes the L_{lasso}, $\beta_{lasso}(\lambda)$, using the train data. **This is done using a solver.**
   B. record $L_{MSE}(\lambda)$ using the validation data.
3. select the λ that minimizes the ***MSE* loss** on the validation data,

$$\lambda_{lasso} = \text{argmin}_{\lambda}\, L_{MSE}(\lambda)$$

4. Refit the model using both train and validation data, $\{\{X, Y\}_{train}, \{X, Y\}_{validation}\}$, now using λ_{Lasso}, resulting to $\hat{\beta}_{lasso}(\lambda_{lasso})$
5. Report MSE or R^2 on $\{X, Y\}_{test}$ given the $\hat{\beta}_{lasso}(\lambda_{lasso})$

Lasso regularization with **validation** only: step by step

For Lasso regression, there is **no** analytical solution for the coefficients, so we use a **solver**.

1. split data into $\{\{X,Y\}_{train}, \{X,Y\}_{validation}, \{X,Y\}_{test}\}$
2. for λ in $\{\lambda_{min}, \ldots \lambda_{max}\}$:
   A. determine the β that minimizes the L_{lasso}, $\beta_{lasso}(\lambda)$, using the train data. **This is done using a solver.**

   B. record $L_{MSE}(\lambda)$ using the validation data.
3. select the λ that minimizes the ***MSE* loss** on the validation data,

$$\lambda_{Lasso} = \text{argmin}_{\lambda} L_{MSE}(\lambda)$$

4. Refit the model using both train and validation data, $\{\{X,Y\}_{train}, \{X,Y\}_{validation}\}$, now using λ_{Lasso}, resulting to $\hat{\beta}_{lasso}(\lambda_{lasso})$
5. Report MSE or R^2 on $\{X,Y\}_{test}$ given the $\beta_{lasso}(\lambda_{lasso})$

Ridge regularization with **CV**: step by step

	λ_1	λ_2	…	λ_n
k_1				
k_2				
…				
k_n				

1. `remove` $\{X, Y\}_{test}$ `from data`
2. `split the rest of data into K folds,` $\{\{X, Y\}_{train}^{-k}, \{X, Y\}_{val}^{k}\}$

Ridge regularization with **CV**: step by step

	λ_1	λ_2	…	λ_n
k_1				
k_2				
…				
k_n				

1. `remove` $\{X,Y\}_{test}$ `from data`
2. `split the rest of data into K folds,` $\{\{X,Y\}_{train}^{-k}, \{X,Y\}_{val}^{k}\}$
3. `for` k `in` $\{1,\dots,K\}$

    `for` λ in $\{\lambda_{0,} \dots, \lambda_n\}$:

Ridge regularization with **CV**: step by step

	λ_1	λ_2	…	λ_n
k_1	L_{11}			
k_2				
…				
k_n				

1. remove $\{X,Y\}_{test}$ from data
2. split the rest of data into K folds, $\{\{X,Y\}_{train}^{-k}, \{X,Y\}_{val}^{k}\}$
3. for k in $\{1,\ldots,K\}$

    for λ in $\{\lambda_0, \ldots, \lambda_n\}$:

    A. determine the β that minimizes the L_{ridge}, $\beta_{ridge}(\lambda,k) = \left(X^TX + \lambda I\right)^{-1} X^T Y$, using the train data of the fold, $\{X,Y\}_{train}^{-k}$.

    B. record $L_{MSE}(\lambda,k)$ using the validation data of the fold $\{X,Y\}_{val}^{k}$

Ridge regularization with **CV**: step by step

	λ_1	λ_2	…	λ_n
k_1	L_{11}	L_{12}	..	…
k_2	L_{21}	…	..	…
…	..	…	..	…
k_n	…	…	…	…

1. remove $\{X,Y\}_{test}$ from data
2. split the rest of data into K folds, $\{\{X,Y\}_{train}^{-k}, \{X,Y\}_{val}^{k}\}$
3. for k in $\{1,\ldots,K\}$

   for λ in $\{\lambda_0,\ldots,\lambda_n\}$:

   A. determine the β that minimizes the L_{ridge}, $\beta_{ridge}(\lambda,k) = \left(X^TX+\lambda I\right)^{-1}X^TY$, using the train data of the fold, $\{X,Y\}_{train}^{-k}$.

   B. record $L_{MSE}(\lambda,k)$ using the validation data of the fold $\{X,Y\}_{val}^{k}$

Ridge regularization with **CV**: step by step

	λ_1	λ_2	…	λ_n
k_1	L_{11}	L_{12}	..	…
k_2	L_{21}	…	..	…
…	..	…	..	…
k_n	…	…	…	…

1. remove $\{X,Y\}_{test}$ from data
2. split the rest of data into K folds, $\{\{X,Y\}_{train}^{-k}, \{X,Y\}_{val}^{k}\}$
3. for k in $\{1,\ldots,K\}$

   for λ in $\{\lambda_{0,} \ldots, \lambda_n\}$:

   A. determine the β that minimizes the L_{ridge}, $\beta_{ridge}(\lambda,k) = \left(X^T X + \lambda I\right)^{-1} X^T Y$, using the train data of the fold, $\{X,Y\}_{train}^{-k}$.

   B. record $L_{MSE}(\lambda,k)$ using the validation data of the fold $\{X,Y\}_{val}^{k}$

At this point we have a 2-D matrix, rows are for different k, and columns are for different λ values.

Ridge regularization with **CV**: step by step

	λ_1	λ_2	...	λ_n
k_1	L_{11}	L_{12}	..	...
k_2	L_{21}	...	..	...
...	..	...	..	...
k_n	...	...	...	...
E[]	$\bar{L}_1$	$\bar{L}_2$	...	$\bar{L}_n$

1. remove $\{X,Y\}_{test}$ from data
2. split the rest of data into K folds, $\{\{X,Y\}_{train}^{-k}, \{X,Y\}_{val}^{k}\}$
3. for k in $\{1,\dots,K\}$

    for λ in $\{\lambda_{0,} \dots, \lambda_n\}$:

    A. determine the β that minimizes the L_{ridge}, $\beta_{ridge}(\lambda,k) = \left(X^TX + \lambda I\right)^{-1} X^T Y$, using the train data of the fold, $\{X,Y\}_{train}^{-k}$.

    B. record $L_{MSE}(\lambda,k)$ using the validation data of the fold $\{X,Y\}_{val}^{k}$

At this point we have a 2-D matrix, rows are for different k, and columns are for different λ values.

4. Average the $L_{MSE}(\lambda,k)$ for each λ, $\bar{L}_{MSE}(\lambda)$.

Ridge regularization with **CV**: step by step

	λ_1	λ_2	…	λ_n
k_1	L_{11}	L_{12}	..	…
k_2	L_{21}	…	..	…
…	..	…	..	…
k_n	…	…	…	…
E[]	$\bar{L}_1$	$\bar{L}_2$	…	$\bar{L}_n$

1. remove $\{X,Y\}_{test}$ from data
2. split the rest of data into K folds, $\{\{X,Y\}_{train}^{-k}, \{X,Y\}_{val}^{k}\}$
3. for k in $\{1, \ldots, K\}$

   for λ in $\{\lambda_0, \ldots, \lambda_n\}$:

   A. determine the β that minimizes the L_{ridge}, $\beta_{ridge}(\lambda, k) = \left(X^T X + \lambda I\right)^{-1} X^T Y$, using the train data of the fold, $\{X,Y\}_{train}^{-k}$.

   B. record $L_{MSE}(\lambda, k)$ using the validation data of the fold $\{X,Y\}_{val}^{k}$

At this point we have a 2-D matrix, rows are for different k, and columns are for different λ values.

1. Average the $L_{MSE}(\lambda, k)$ for each λ, $\bar{L}_{MSE}(\lambda)$.
2. Find the λ that minimizes the $\bar{L}_{MSE}(\lambda)$, resulting to λ_{ridge}.

Ridge regularization with **CV**: step by step

1. remove $\{X,Y\}_{test}$ from data
2. split the rest of data into K folds, $\{\{X,Y\}_{train}^{-k}, \{X,Y\}_{val}^{k}\}$
3. for k in $\{1,\dots,K\}$

   for λ in $\{\lambda_{0,} \dots, \lambda_n\}$:

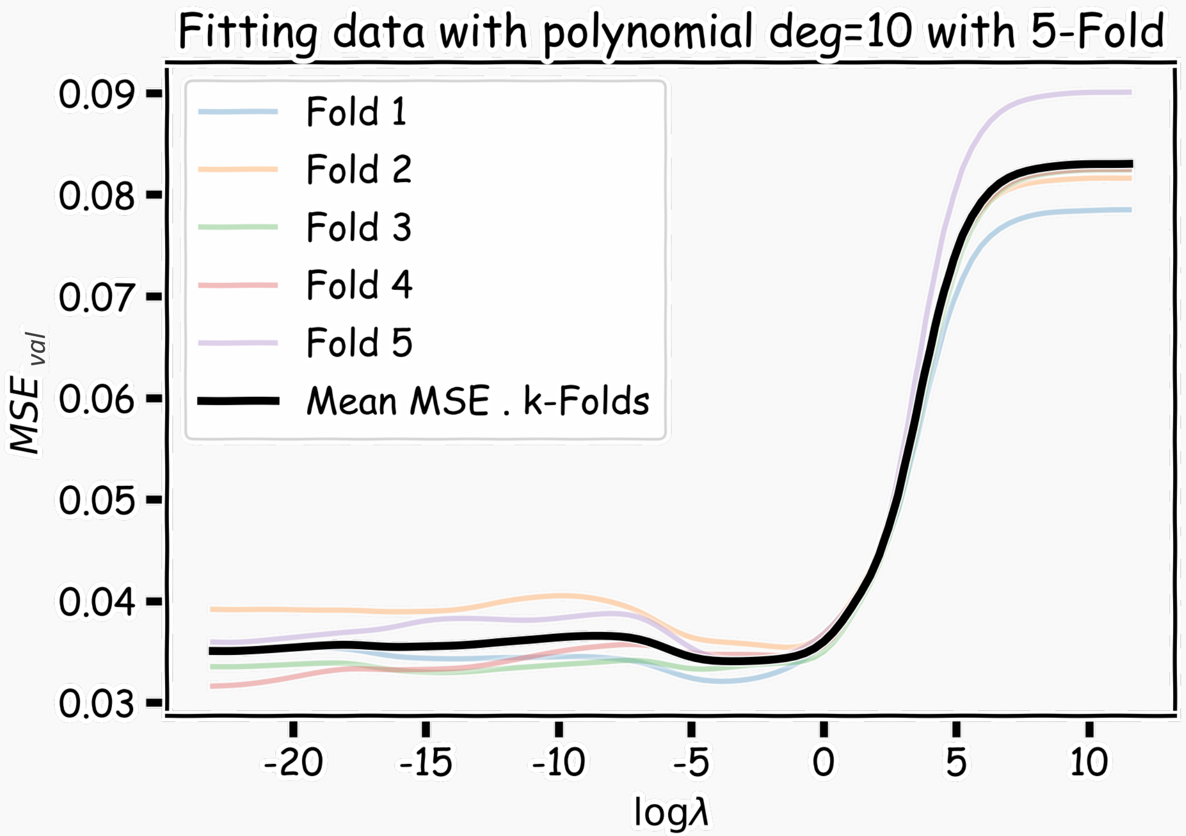
   A. determine the β that minimizes the L_{ridge}, $\beta_{ridge}(\lambda,k) = \left(\mathrm{X}^{\mathrm{T}}\mathrm{X} + \lambda I\right)^{-1} X^T Y$, using the train data of the fold, $\{X,Y\}_{train}^{-k}$.

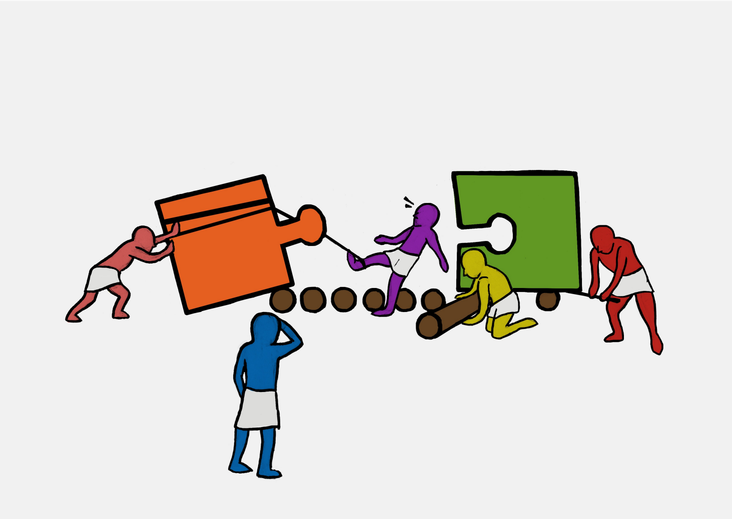
   B. record $L_{MSE}(\lambda,k)$ using the validation data of the fold $\{X,Y\}_{val}^{k}$

At this point we have a 2-D matrix, rows are for different k, and columns are for different λ values.

4. Average the $L_{MSE}(\lambda,k)$ for each λ, $\bar{L}_{MSE}(\lambda)$.
5. Find the λ that minimizes the $\bar{L}_{MSE}(\lambda)$, resulting to λ_{ridge}.
6. Refit the model using the full training data, $\{\{X,Y\}_{train}, \{X,Y\}_{val}\}$, resulting to $\hat{\beta}_{ridge}(\lambda_{ridge})$
7. report MSE or R^2 on $\{X,Y\}_{test}$ given the $\hat{\beta}_{ridge}(\lambda_{ridge})$

Ridge regularization with **cross-validation** only: step by step





👩‍🏫 Exercise: Simple Lasso and Ridge Regularization

The aim of this exercise is to understand **Lasso and Ridge regularization.**

For this we will plot the predictor vs coefficient as a horizontal bar chart. The graph will look similar to the one given below.