



< Previous	✓	✓	✓	✓		Next >
------------	---	---	---	---	--	--------

## Regularization Hyperparameters

Bookmark this page



## 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 motivation for performing model selection is to avoid **overfitting**, which we saw can happen when: there are too many predictors, because:

1. there are too many predictors, because:
  - the feature space has high dimensionality
  - the polynomial degree is too high
  - too many interaction terms are considered
2. the coefficients' values are too **extreme**

We've already seen ways to address the problem of choosing predictors and polynomial degree using greedy algorithms and cross-validation. But what about the second potential source of overfitting? How do we discourage extreme coefficient values in the model parameters?

## Regularization

What we want is low model error. We've been using mean squared error for our model's loss function:

$$\frac{1}{n} \sum_{i=1}^n |y_i - \beta^T x_i|^2$$

We also want to discourage extreme parameter values. We could also create a loss which is a function of the magnitudes of the model's parameters. We'll call this  $L_{reg}$ . We could do this in several ways. For example, we could sum the squares of the parameters or their absolute values.

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

Not that the summation index starts at 1. The model is not penalized for its  $\beta_0$  which can be interpreted as the intercept.

Now we can combine these two loss functions into a single loss function for our model using **regularization**.

$$L = \frac{1}{n} \sum_{i=1}^n |y_i - \beta^T x_i|^2 + \lambda L_{reg}$$

$\lambda$  is the **regularization parameter**. It controls the relative importance between model error and the regularization term.

$\lambda = 0$ : equivalent to regression model using no regularization.  $\lambda = \infty$ : yields a model where all  $\beta$ s are 0.

But how do we determine which value of  $\lambda$  to use? The answer is with cross-validation! We will try many different values of  $\lambda$  and pick the one that gives us the best cross-validation loss scores

## Regularization: LASSO Regression

**LASSO** regression: minimize  $L_{LASSO}$  with respect to  $\beta$ s.

$$L_{LASSO} = \frac{1}{n} \sum_{i=1}^n |y_i - \beta^T x_i|^2 + \lambda \sum_{j=1}^J |\beta_j|$$

Note that  $\sum_{j=1}^J |\beta_j|$  is the **L1** norm of the  $\beta$  vector.

There's no need to regularize the bias,  $\beta_0$ , since it is not connected to the predictors.

**Regularization: Ridge Regression**

**Ridge** regression: minimize  $L_{RIDGE}$  with respect to  $\beta$ s.

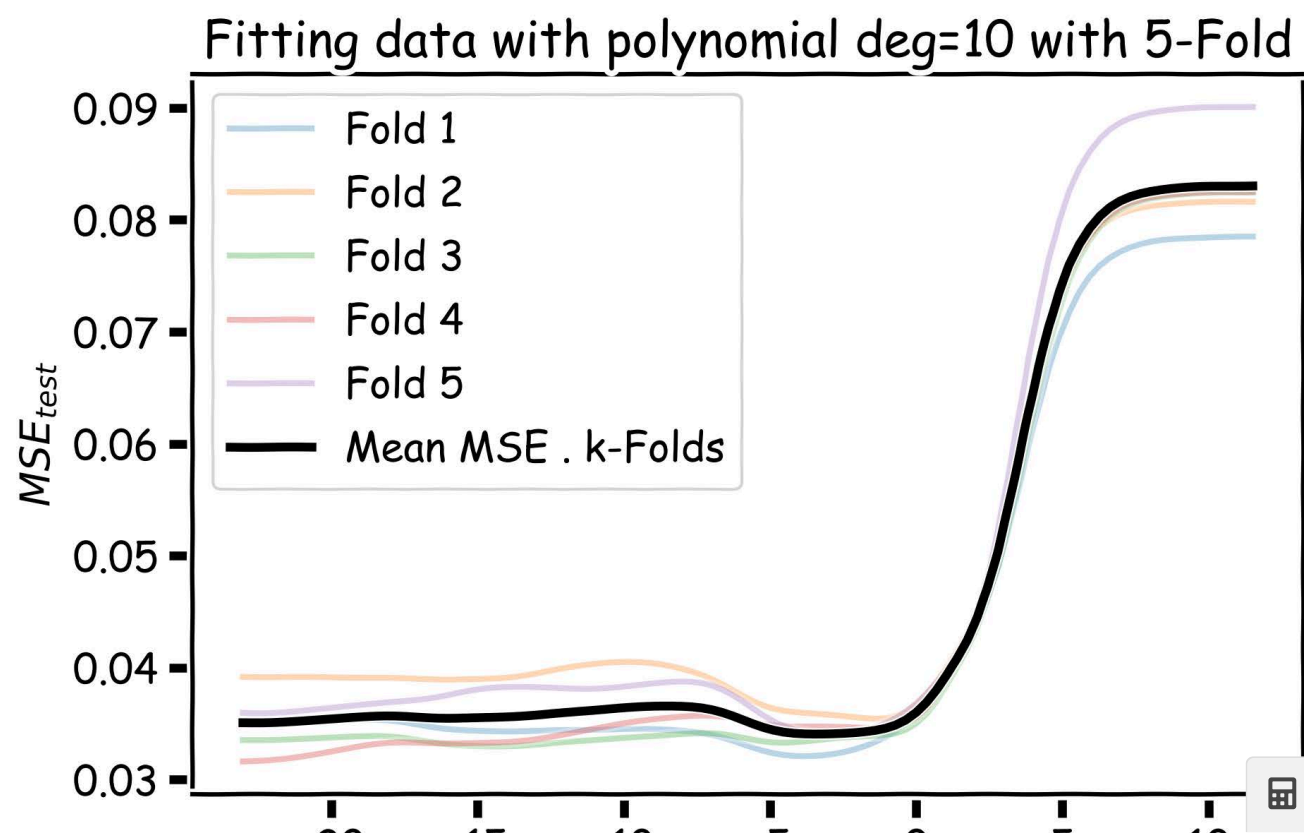
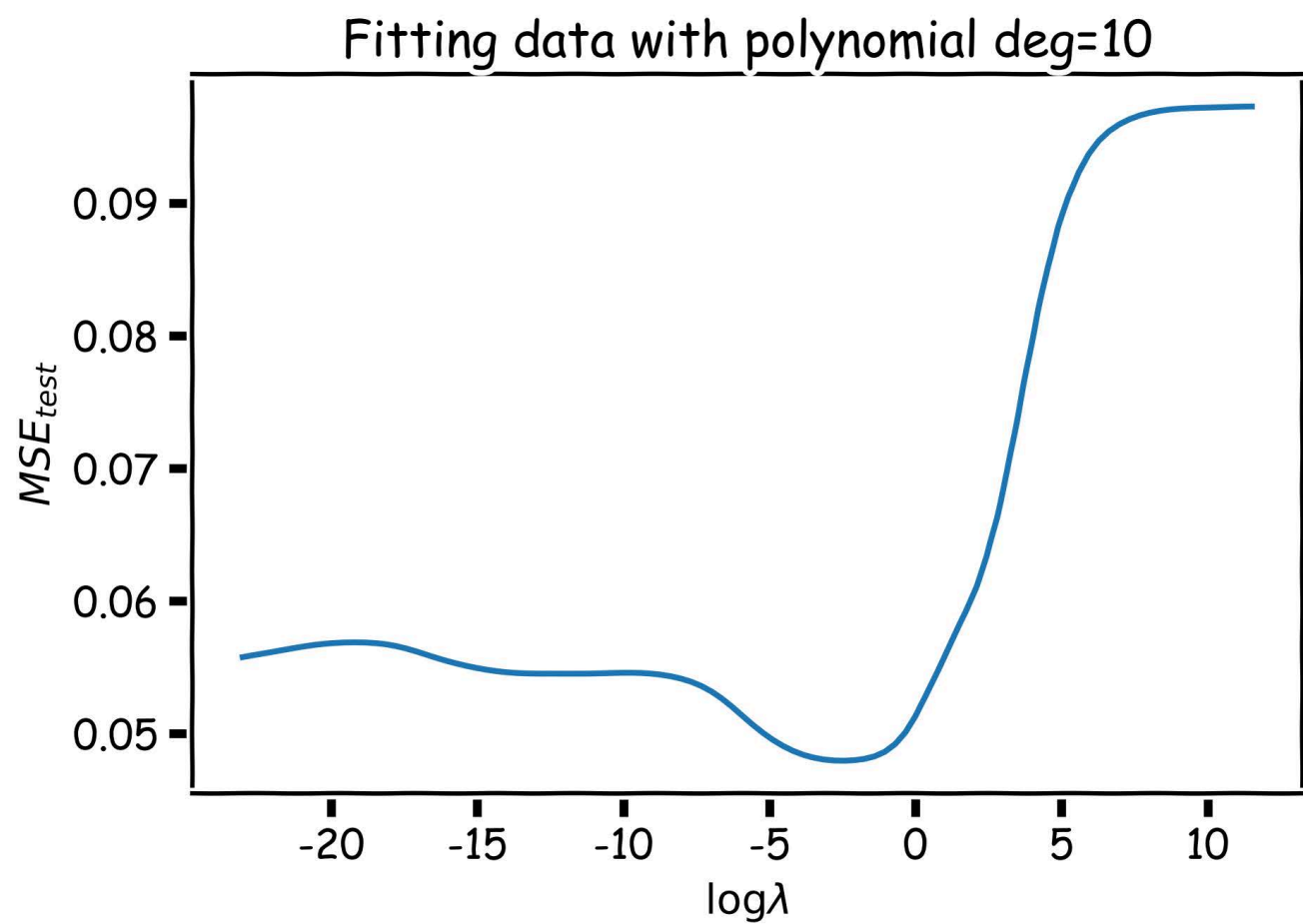
$$L_{RIDGE} = \frac{1}{n} \sum_{i=1}^n |y_i - \beta^T x_i|^2 + \lambda \sum_{j=1}^J \beta_j^2$$

Note that  $\sum_{j=1}^J \beta_j^2$  is the **L2** norm of the  $\beta$  vetor.

Again we do not regularize the bias,  $\beta_0$ .

**Ridge regularization with single validation set vs with cross-validation**

To emphasize the usefulness of cross-validation, compare these two plots demonstrating ridge regularization using a single validation set and using cross-validation. Note how by taking the average of the 5 folds we can get more reliable results than relying on just one single validation split.



-20

-15

-10

-5

0

5

10

$\log \lambda$

Discussion Board (External resource)

Haga clic en Aceptar para que su nombre de usuario y dirección de correo electrónico se envíen a una aplicación de terceros.

Aceptar

< Previous

Next >



edX

- About
- Affiliates
- edX for Business
- Open edX
- Careers
- News

Legal

- Terms of Service & Honor Code
- Privacy Policy
- Accessibility Policy
- Trademark Policy
- Sitemap
- Cookie Policy
- Your Privacy Choices

Connect

- Idea Hub
- Contact Us
- Help Center
- Security
- Media Kit

