Deep Learning
Lab 5: Regularization

DataLab, 2022

Department of Computer Science,
National Tsing Hua University, Taiwan

Regularization

techniques that improve the
generalizability of a trained model

Outline

- Scikit-learn
- Learning Theory
  - Error Curves and Model Complexity
  - Learning Curves and Sample Complexity
- Weight Decay
  - Ridge Regression
  - LASSO
- Validation
- Assignment

Outline

- Scikit-learn
- Learning Theory
  - Error Curves and Model Complexity
  - Learning Curves and Sample Complexity
- Weight Decay
  - Ridge Regression
  - LASSO
- Validation
- Assignment

Scikit-learn

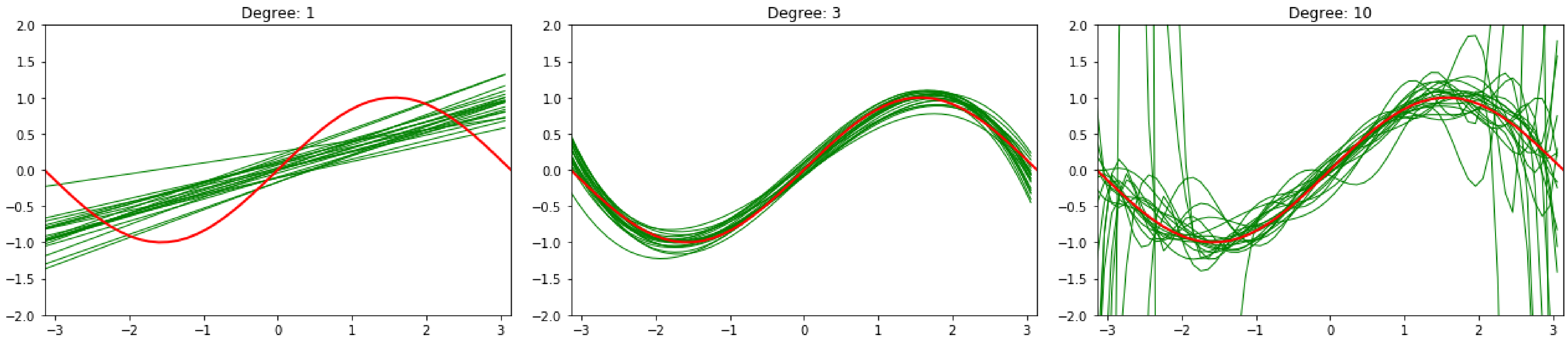
- Scikit-learn is a free software machine learning library for the Python programming language
- It features various classification, regression and clustering algorithms
  - including SVM (support vector machines), Random Forests, gradient boosting, k-means and DBSCAN, and is designed to interoperate with the Python numerical and scientific libraries NumPy and SciPy
- pip install scikit-learn / conda install scikit-learn

Outline

- Scikit-learn
- Learning Theory
  - Error Curves and Model Complexity
  - Learning Curves and Sample Complexity
- Weight Decay
  - Ridge Regression
  - LASSO
- Validation
- Assignment

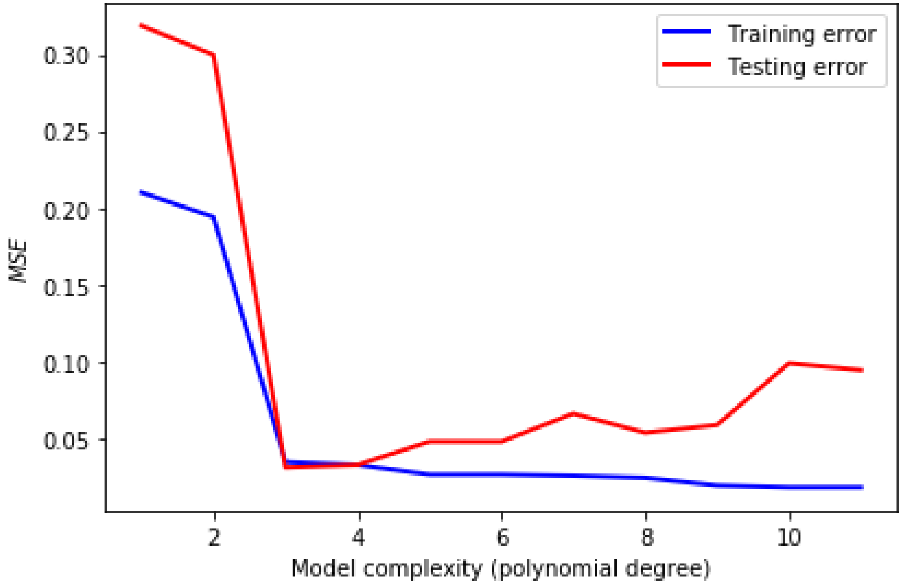
Learning Theory

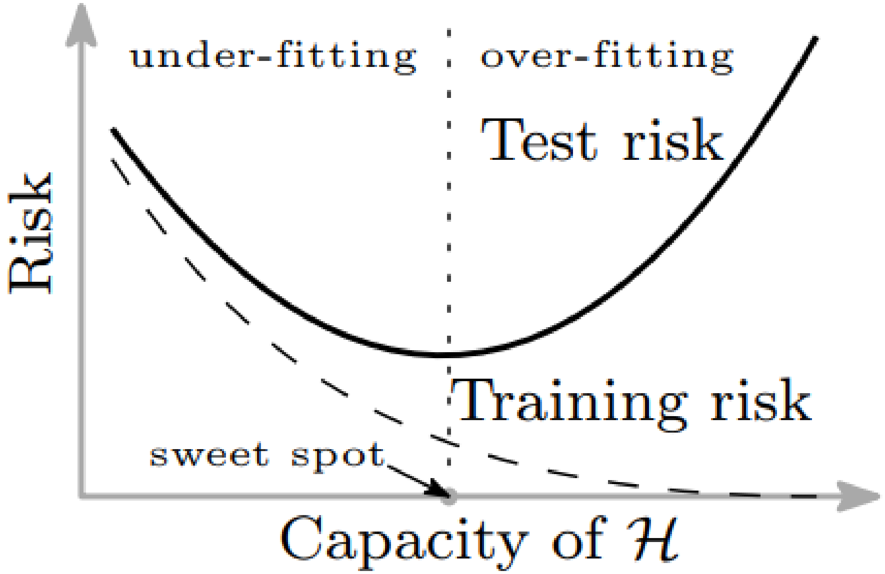
- Learning theory provides a means to understand the generalizability of the model
- **Model complexity** plays a crucial role
  - Too simple: high bias and underfitting
  - Too complex: high variance and overfitting



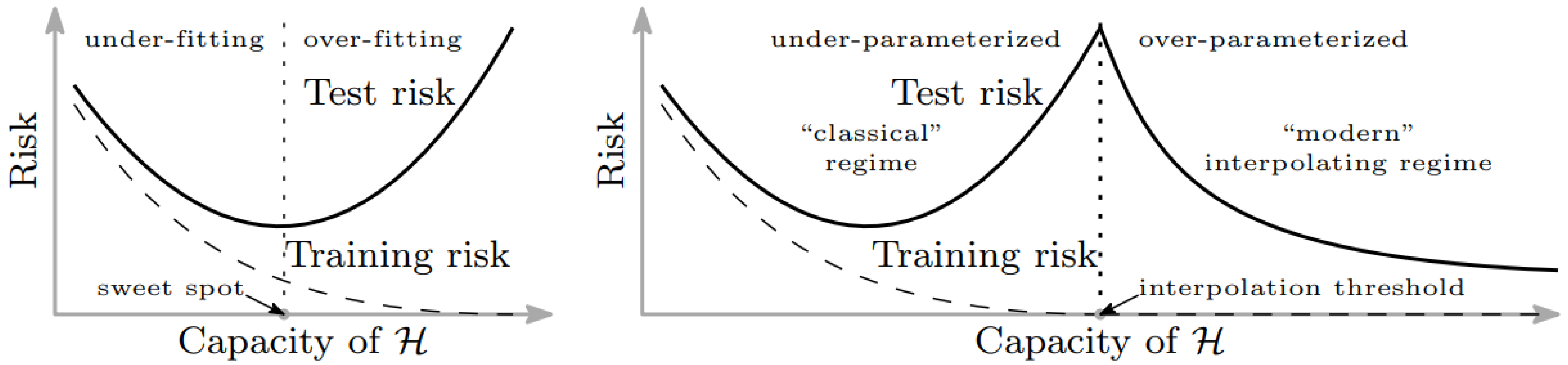
Error Curves and Model Complexity

- It is relatively hard to observe the figures showed in the last slide, since normally we will never know the data distribution of ground truth (red line in the last slide)
- Instead, we can get those information by observing the training and testing error curve





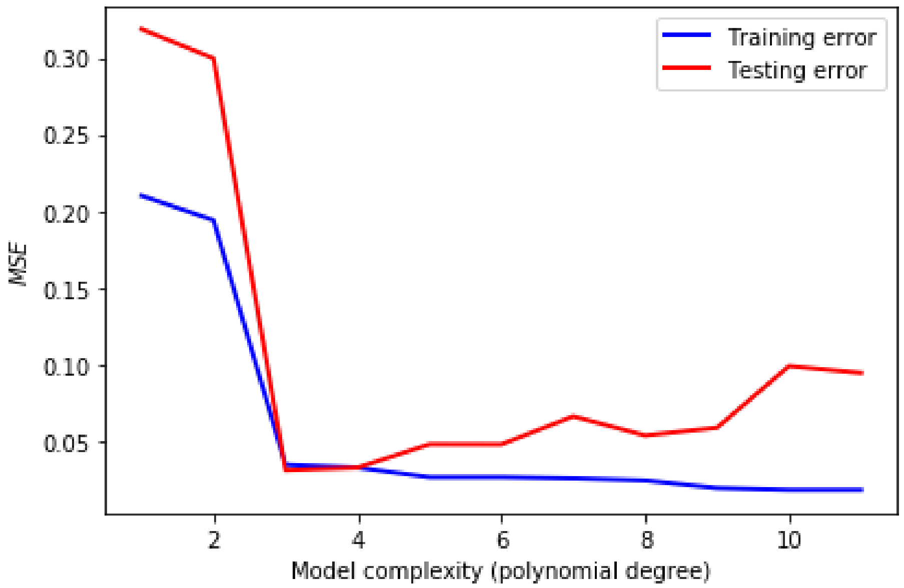
Double Descent Curves in **Modern** Machine Learning**

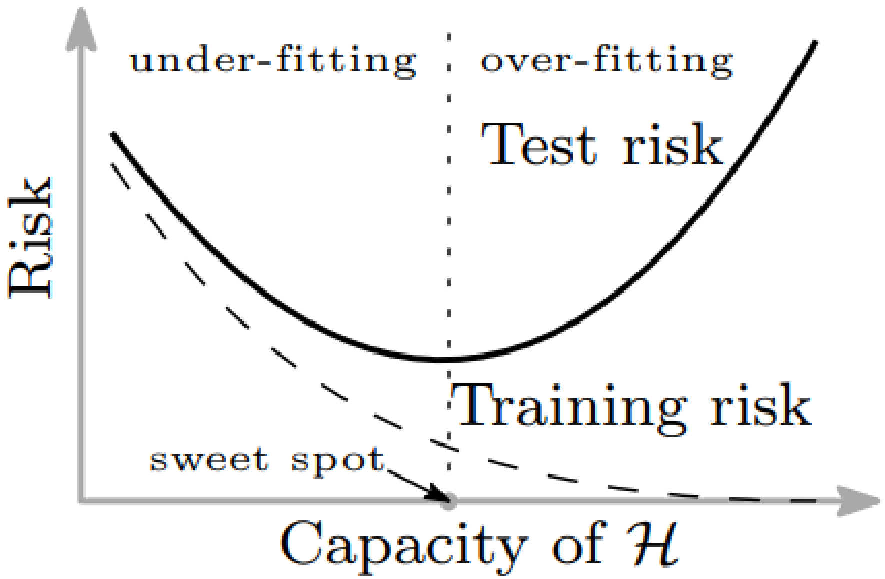


Reconciling modern machine learning practice and the bias-variance trade-of (PNAS'19)
Double-descent curves in neural networks: a new perspective using Gaussian processes (arXiv'21)

Error Curves and Model Complexity

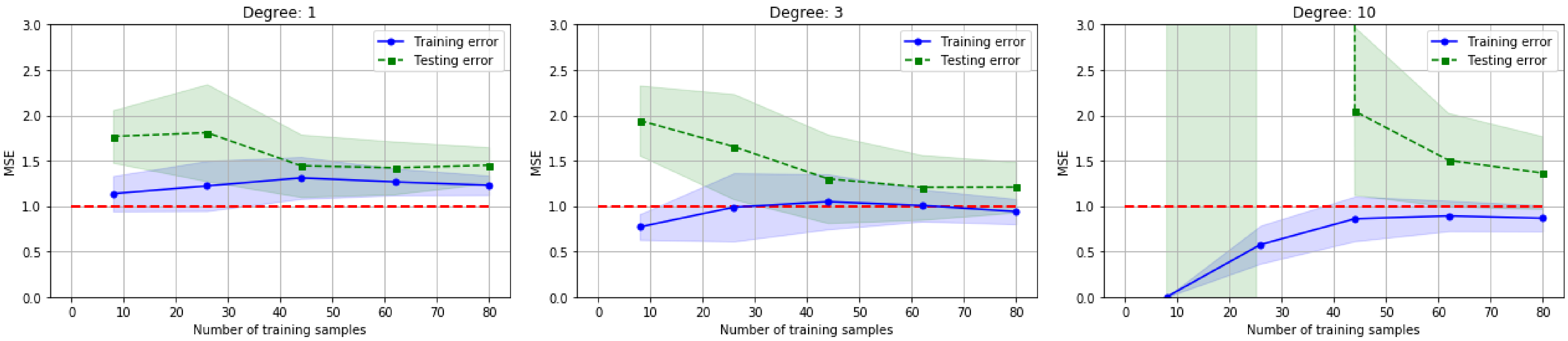
- Although the error curve visualizes the impact of model complexity, the bias-variance tradeoff holds **only when you have sufficient training examples**





Learning Curves and Sample Complexity

- The bounding methods of learning theory tell us that a model is likely to overfit regardless of it complexity **when the size of training set is small**. The **learning curves** are a useful tool for understanding how much training examples are sufficient



Outline

- Scikit-learn
- Learning Theory
    - Error Curves and Model Complexity
    - Learning Curves and Sample Complexity
- Weight Decay
    - Ridge Regression
    - LASSO
- Validation
- Assignment

Weight Decay

- A common regularization approach. The idea is to add a term in the cost function against complexity
  - Ridge Regression (L_2)

$$\arg\min_{\boldsymbol{w},b} \|\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w} - b\mathbf{1})\|^2 + \alpha\|\boldsymbol{w}\|^2$$

  - LASSO (L_1)

$$\arg\min_{\boldsymbol{w},b} \|\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w} - b\mathbf{1})\|^2 + \alpha\|\boldsymbol{w}\|_1$$

Ridge Regression

- A small value α drastically reduces the testing error. Nevertheless, it's not a good idea to increase α forever, since it will over-shrink the coefficients of w and result in underfitting

$$\arg\min_{w,b} \|\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w} - b\boldsymbol{1})\|^2 + \alpha\|\boldsymbol{w}\|^2$$

```python
from sklearn.linear_model import Ridge
from sklearn.metrics import mean_squared_error

poly = PolynomialFeatures(degree=3)
X_poly = poly.fit_transform(X_std)
X_train, X_test, y_train, y_test = train_test_split(
    X_poly, y, test_size=0.3, random_state=0)

for a in [0, 1, 10, 100, 1000]:
    lr_rg = Ridge(alpha=a)
    lr_rg.fit(X_train, y_train)

    y_train_pred = lr_rg.predict(X_train)
    y_test_pred = lr_rg.predict(X_test)

    print('\n[Alpha = %d]' % a )
    print('MSE train: %.2f, test: %.2f' % (
                    mean_squared_error(y_train, y_train_pred),
                    mean_squared_error(y_test, y_test_pred)))
```

```
[Alpha = 0]
MSE train: 0.00, test: 19958.68

[Alpha = 1]
MSE train: 0.73, test: 23.05

[Alpha = 10]
MSE train: 1.66, test: 16.83

[Alpha = 100]
MSE train: 3.60, test: 15.16

[Alpha = 1000]
MSE train: 8.81, test: 19.22
```

LASSO

- An alternative weight decay approach that can lead to sparse w is the LASSO. Depending on the value of α, certain weights can become zero much faster than others

$$\arg\min_{\boldsymbol{w},b} \|\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w} - b\mathbf{1})\|^2 + \alpha\|\boldsymbol{w}\|_1$$

```
from sklearn.linear_model import Lasso
from sklearn.metrics import mean_squared_error

for a in [0.001, 0.01, 0.1, 1, 10]:
    lr_rg = Lasso(alpha=a)
    lr_rg.fit(X_train, y_train)

    y_train_pred = lr_rg.predict(X_train)
    y_test_pred = lr_rg.predict(X_test)

    print('\n[Alpha = %.2f]' % a )
    print('MSE train: %.2f, test: %.2f' % (
                    mean_squared_error(y_train, y_train_pred),
                    mean_squared_error(y_test, y_test_pred)))
```

```
[Alpha = 0.0000]
MSE train: 0.55, test: 61.02

[Alpha = 0.0010]
MSE train: 0.64, test: 29.11

[Alpha = 0.0100]
MSE train: 1.52, test: 19.51

[Alpha = 0.1000]
MSE train: 4.34, test: 15.52

[Alpha = 1.0000]
MSE train: 14.33, test: 22.42

[Alpha = 10.0000]
MSE train: 55.79, test: 53.42
```

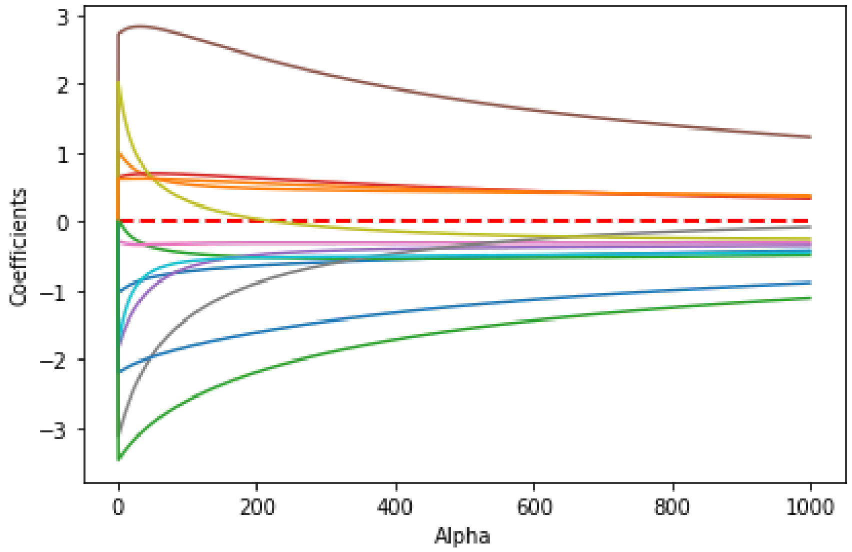
Ridge vs LASSO

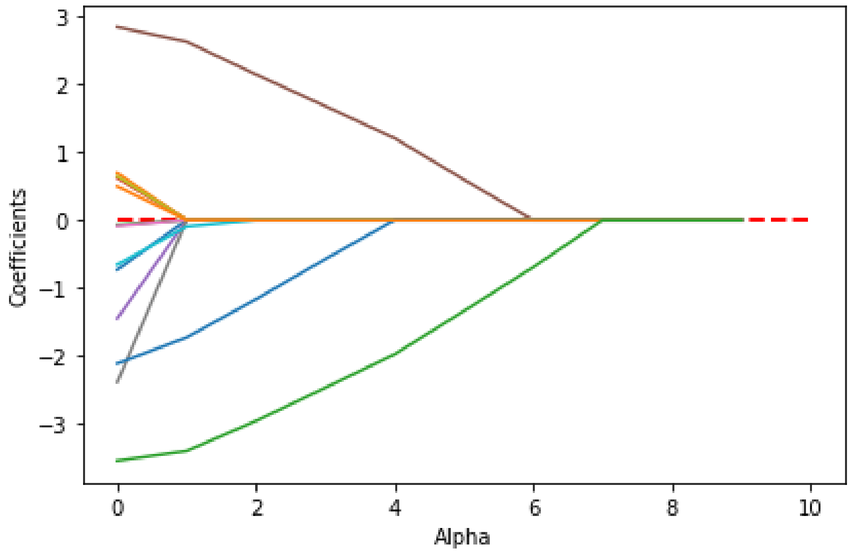
- Why is LASSO sparse?

Ridge: [0.5, 0.5, 0.5, 0.5]

Initial weights: [1, 0.5, 1, 0.5]

LASSO: [0.5, 0, 0.5, 0]



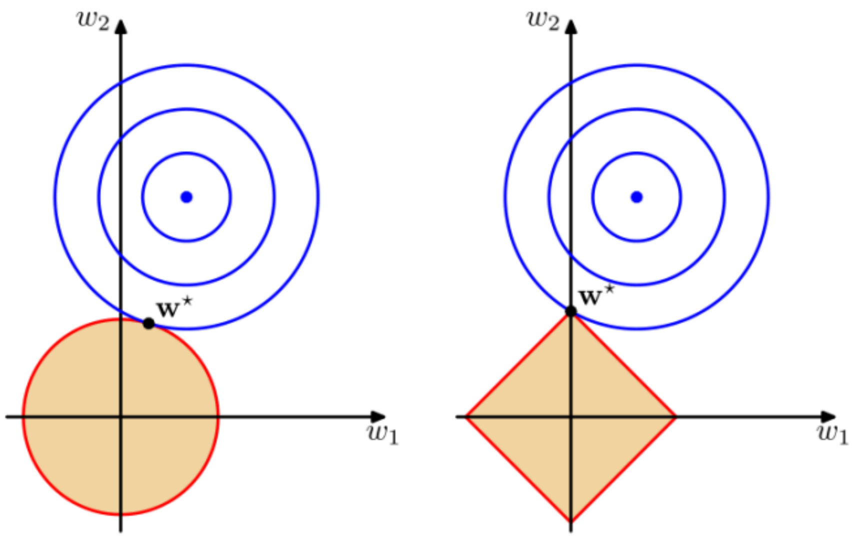


Ridge vs LASSO

- Why is LASSO sparse?

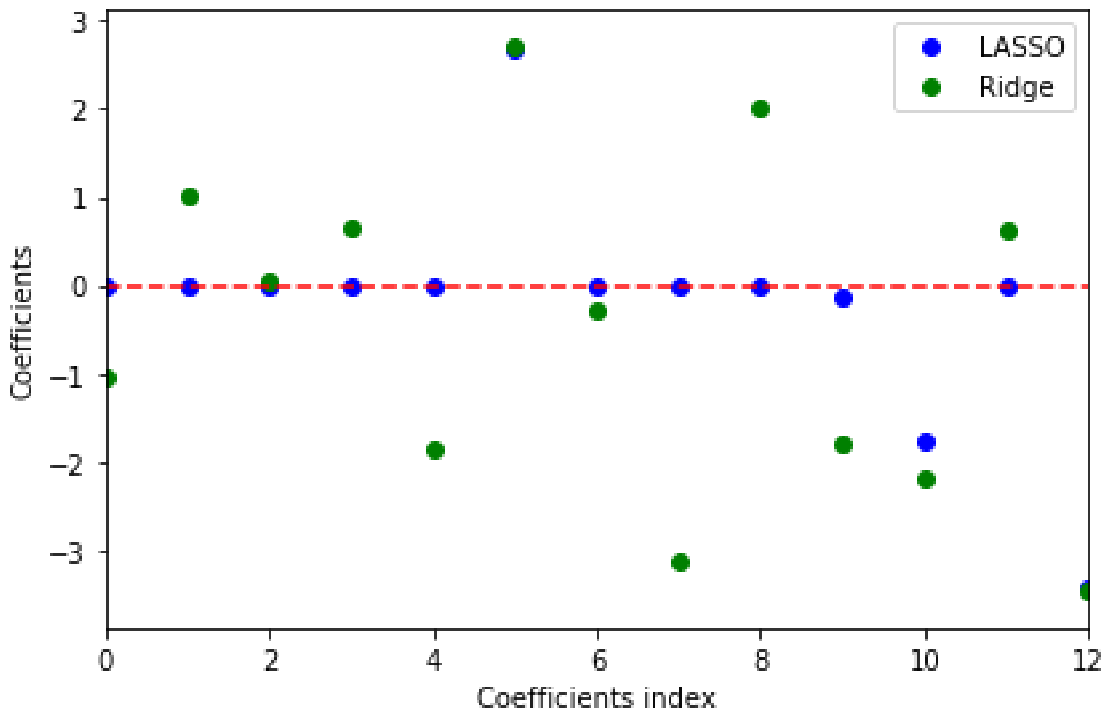
$$\arg\min_{\boldsymbol{w},b} \frac{1}{2N} \|\boldsymbol{y} - (\boldsymbol{X}\boldsymbol{w} - b\mathbf{1})\|^2 + \alpha \|\boldsymbol{w}\|_1$$

- The surface of the cost function is the sum of SSE (blue contours) and 1-norm (red contours)
- Optimal point locates on some axes



Ridge vs LASSO

- LASSO can also be treated as a supervised **feature selection** technique when choosing a suitable regularization strength α to make only part of coefficients become exactly zeros

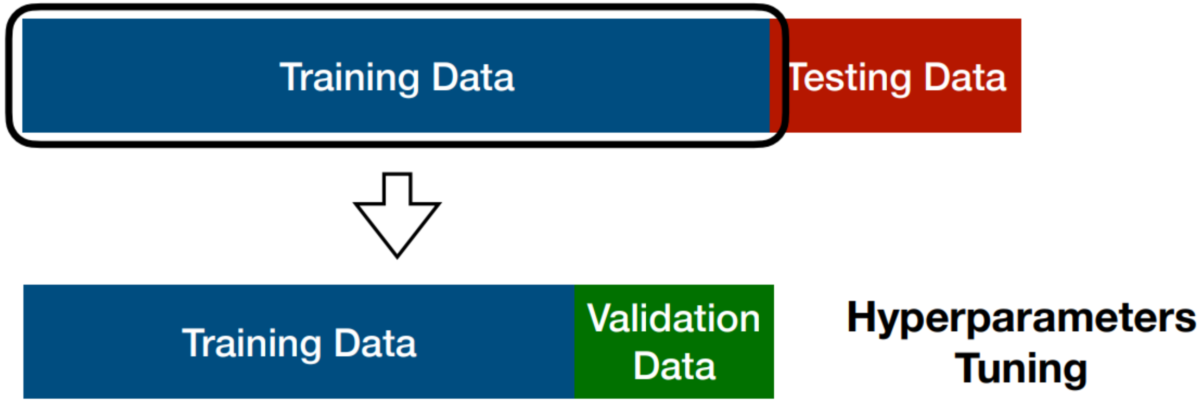


Outline

- Scikit-learn
- Learning Theory
  - Error Curves and Model Complexity
  - Learning Curves and Sample Complexity
- Weight Decay
  - Ridge Regression
  - LASSO
- Validation
- Assignment

Validation

- Another useful regularization technique that helps us decide the proper value of **hyperparameters**
- The idea is to split your data into the training, validation, and testing sets and then select the best value based on validation performance
- NOTE: It is important that we should never peep testing data during training



Validation

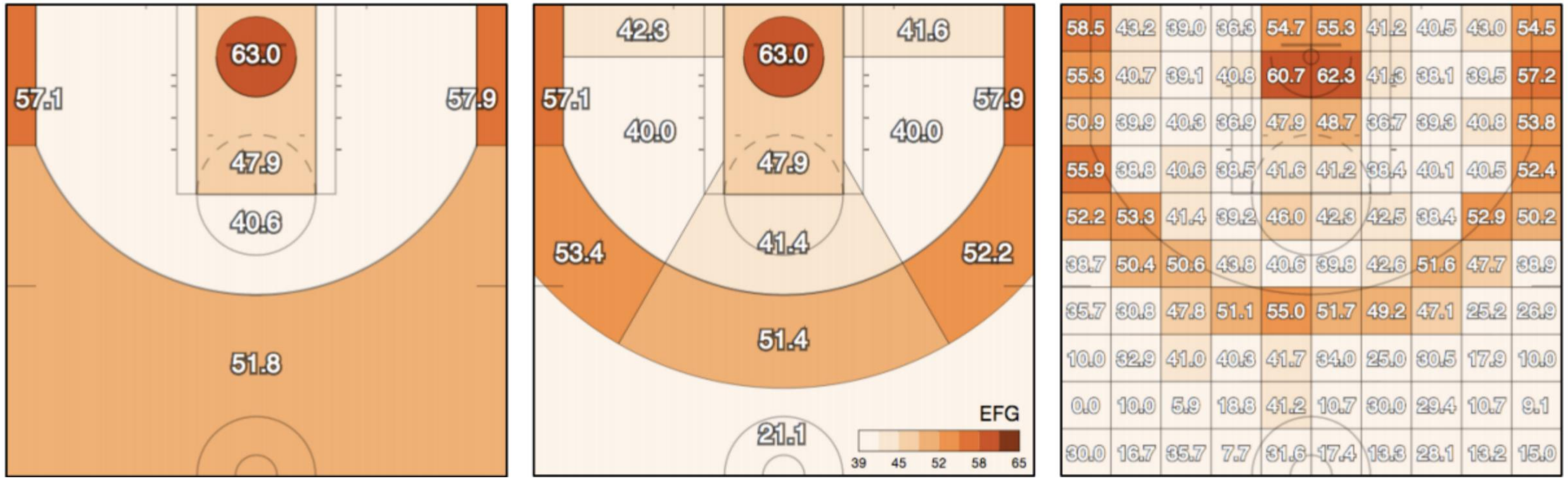
```
[Degree = 1]
MSE train: 25.00, valid: 21.43, test: 32.09

[Degree = 2]
MSE train: 9.68, valid: 14.24, test: 20.24

[Degree = 3]
MSE train: 3.38, valid: 17.74, test: 18.63

[Degree = 4]
MSE train: 1.72, valid: 16.67, test: 30.98

[Degree = 5]
MSE train: 0.97, valid: 59.73, test: 57.02

[Degree = 6]
MSE train: 0.60, valid: 1444.08, test: 33189.41
```

Outline

- Scikit-learn
- Learning Theory
  - Error Curves and Model Complexity
  - Learning Curves and Sample Complexity
- Weight Decay
  - Ridge Regression
  - LASSO
- Validation
- Assignment

Assignment

- In this assignment, we would like to predict the success of shots made by basketball players in the NBA



Assignment

- In this assignment, we would like to predict the success of shots made by basketball players in the NBA
  - **y_test** is hidden this time
  - Allow to use **any model** you have learned before to achieve the best accuracy
  - Select the best **3 features**, and show the accuracy with only those
- Hint
  - **Preprocess the data** to help your training
  - Since you don't have y_test this time, you may need to **split a validation set** for checking your performance
  - It is possible to use regression model as a classifier, for example RidgeClassifier

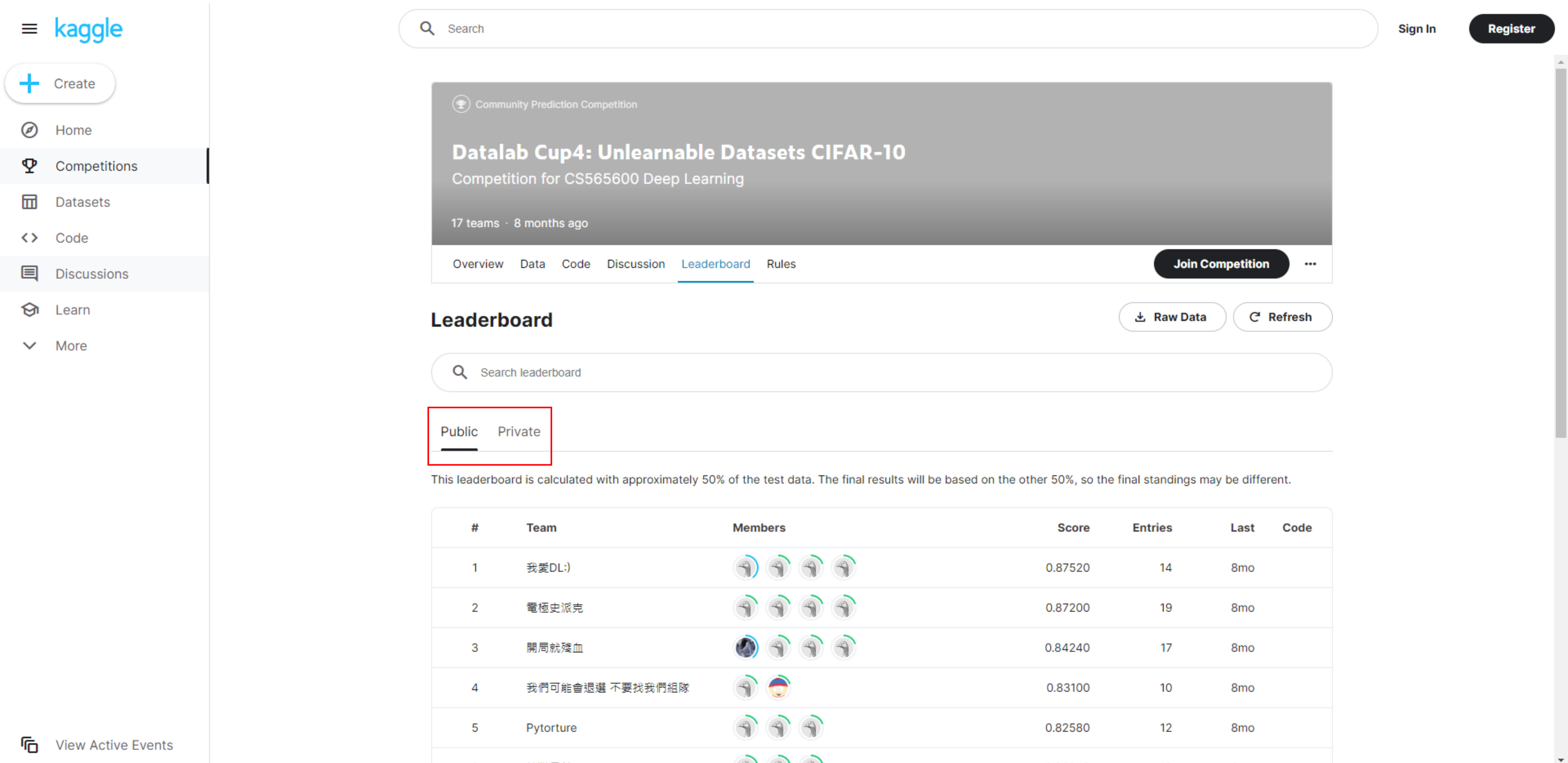
Assignment

- Submit to **eeclass** with your:
  - **ipynb** (Lab05_{student_id}.ipynb)
  - **Prediction** (Lab05_{student_id}_y_pred.csv)
- The notebook should contain
  - How you **evaluate** your model
  - **All models** you have tried and the result
  - Plot the **error curve** of your best model and tell if it is **over-fit or not**
  - The **top-3 features** you find and how you find it
  - A **brief report** what you do in this assignment
- Deadline: **2022-10-06(Thur) 23:59**

About Competition

- Students will group (2~4 people a group)
- This class suggests each group of students to **prepare a GPU card** to perform the necessary computing.
  - NO GPU CARD PROVIDED IN THE CLASS.
- Sign up for your group before the class on 10/4(Tue)

About Competition



1	▲ 1	電極史派克	0.87260	19	8mo
2	▲ 1	開局就殘血	0.84180	17	8mo
3	▲ 1	我們可能會退選 不要找我們組隊	0.84040	10	8mo
4	▲ 1	Pytorture	0.83080	12	8mo
5	▲ 2	百鬼組	0.82660	8	8mo
6	—	地獄兄弟	0.82260	13	8mo
		cifar_TA_80.csv	0.81780		
7	▼ 6	我愛DL:)	0.80660	14	8mo
8	—	kk	0.80360	7	8mo
9	▲ 1	Papaya	0.78640	8	8mo
10	▼ 1	Pytorch	0.78040	10	8mo
11	—	讓我活下來	0.74460	10	8mo
12	—	麻椅上訴	0.74080	10	8mo
13	—	歐🆒比	0.72400	12	8mo
14	▲ 1	哭啊	0.69880	11	8mo
15	▼ 1	Team123	0.68760	24	8mo
16	—	無敵的吧	0.68560	2	8mo
		cifar_TA_60.csv	0.63740		
17	—	深度度嚕嚕嚕	0.41960	3	8mo