Brain
Code
Camp

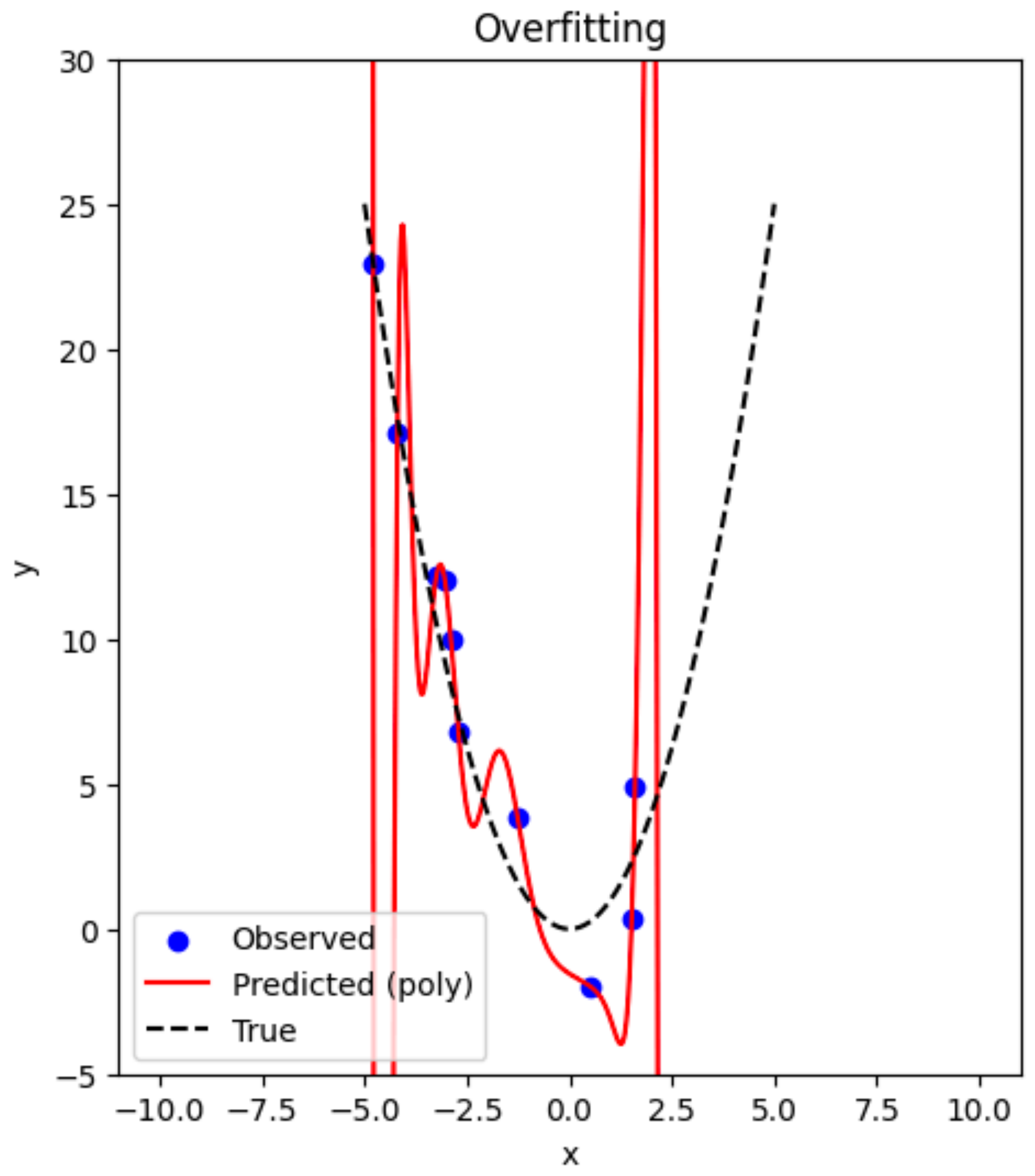
Polynomial Regression, Overfitting and Regularization

Itthi Chatnuntawech

Brain Code Camp

Overfitting

$$\hat{y} = \hat{w}_0 + \hat{w}_1 x + \hat{w}_2 x^2 + \ldots + \hat{w}_p x^p$$



training data

$$\min_{\hat{w}_0,\ldots,\hat{w}_p} MSE(Y, \hat{Y}) = \min_{\hat{w}_0,\ldots,\hat{w}_p} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - (\hat{w}_0 + \hat{w}_1 x_i + \ldots + \hat{w}_p x_i^p) \right)^2$$

What if our model "memorizes" the training data?

สร้างคน ข้ามพรมแดน

Regularization

training data

$$\min_{\hat{w}_0,\ldots,\hat{w}_p} MSE(Y, \hat{Y})$$

$$\min_{\hat{\mathbf{w}}} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2$$

Regularization

training data regularization term

$$\min_{\hat{w}_0,\ldots,\hat{w}_p} MSE(Y, \hat{Y}) \quad + \lambda R(\hat{w}_0, \hat{w}_1, \ldots, \hat{w}_p)$$

regularization parameter

$$\min_{\hat{\mathbf{w}}} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 + \lambda R(\hat{\mathbf{w}})$$

L2 regularization/ Tikhonov regularization

$$\min_{\hat{w}_0,\ldots,\hat{w}_p} MSE(Y, \hat{Y}) \quad + \lambda(\hat{w}_0^2 + \hat{w}_1^2 + \ldots + \hat{w}_p^2)$$

$$\min_{\hat{\mathbf{w}}} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 + \lambda\|\hat{\mathbf{w}}\|_2^2$$

Ridge regression

L1 regularization

$$\min_{\hat{w}_0,\ldots,\hat{w}_p} MSE(Y, \hat{Y}) \quad + \lambda(|\hat{w}_0| + |\hat{w}_1| + \ldots + |\hat{w}_p|)$$

$$\min_{\hat{\mathbf{w}}} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 + \lambda\|\hat{\mathbf{w}}\|_1$$

Least Absolute Shrinkage and Selection Operator (LASSO)

L2 Regularization

training data regularization term

$$\min_{\hat{w}_0,\ldots,\hat{w}_p} MSE(Y, \hat{Y}) + \lambda(\hat{w}_0^2 + \hat{w}_1^2 + \ldots + \hat{w}_p^2)$$

regularization parameter

Ensures that what we predict, $\hat{Y}$, matches what we have collected, Y

Ensures that the parameters do not become too large

$\lambda = 0$ $\min_{\hat{w}_0,\ldots,\hat{w}_p} MSE(Y, \hat{Y}) + 0$ no regularization

λ large $\min_{\hat{w}_0,\ldots,\hat{w}_p}$ small $+ \lambda(\hat{w}_0^2 + \hat{w}_1^2 + \ldots + \hat{w}_p^2)$ Do not care about the training data

A good λ cares about the training data, while also pays attention to the regularization term

Linear regression

sklearn.linear_model.LinearRegression

class sklearn.linear_model.**LinearRegression**(**, fit_intercept=True, copy_X=True, n_jobs=None, positive=False*) [source]

Linear regression with L2 regularization/ Tikhonov regularization

sklearn.linear_model.Ridge

class sklearn.linear_model.**Ridge**(*alpha=1.0, *, fit_intercept=True, copy_X=True, max_iter=None, tol=0.0001, solver='auto', positive=False, random_state=None*) [source]

Linear regression with L1 regularization

sklearn.linear_model.Lasso

class sklearn.linear_model.**Lasso**(*alpha=1.0, *, fit_intercept=True, precompute=False, copy_X=True, max_iter=1000, tol=0.0001, warm_start=False, positive=False, random_state=None, selection='cyclic'*) ¶ [source]

Brain Code Camp

Optional: Solution to Ridge Regression

$$\min_{\hat{\mathbf{w}}} \|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 + \lambda\|\hat{\mathbf{w}}\|_2^2$$

Compute the gradient of the loss function with respect to $\hat{\mathbf{w}}$ and set it to 0

$$\nabla_{\hat{\mathbf{w}}}(\|\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}\|_2^2 + \lambda\|\hat{\mathbf{w}}\|_2^2) = -2\mathbf{X}^T(\mathbf{y} - \mathbf{X}\hat{\mathbf{w}}) + 2\lambda\hat{\mathbf{w}} = 0$$

$$-\mathbf{X}^T\mathbf{y} + \mathbf{X}^T\mathbf{X}\hat{\mathbf{w}} + \lambda\hat{\mathbf{w}} = 0$$

$$\mathbf{X}^T\mathbf{X}\hat{\mathbf{w}} + \lambda\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{y}$$

$$(\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})\hat{\mathbf{w}} = \mathbf{X}^T\mathbf{y}$$

$$\hat{\mathbf{w}} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y}$$