Regularization

Chris Cornwell

April 21, 2025

Outline

Outline

Introduction

- **Regularization**, or *weight* regularization, is a technique to prevent a hypothesis class from having high Variance, despite having a potentially large number of parameters.

Introduction

- **Regularization**, or *weight* regularization, is a technique to prevent a hypothesis class from having high Variance, despite having a potentially large number of parameters.
- Broadly speaking, it involves adding a penalty on the loss function that will make the loss larger if the absolute value of model parameters are large.

Introduction

- **Regularization**, or *weight* regularization, is a technique to prevent a hypothesis class from having high Variance, despite having a potentially large number of parameters.
- Broadly speaking, it involves adding a penalty on the loss function that will make the loss larger if the absolute value of model parameters are large.
- This causes updates during training to avoid increasing the size of parameters unless that objective is outweighed by a significant decrease in prediction error.

Introduction

- **Regularization**, or *weight* regularization, is a technique to prevent a hypothesis class from having high Variance, despite having a potentially large number of parameters.
- Broadly speaking, it involves adding a penalty on the loss function that will make the loss larger if the absolute value of model parameters are large.
- This causes updates during training to avoid increasing the size of parameters unless that objective is outweighed by a significant decrease in prediction error.

A motivation for why to use regularization to balance high Variance comes from polynomial fitting.

Fitting a Polynomial

A polynomial function will be fit to the data depicted below. The blue points are part of the training set (32 points) and the reddish-orange points are in the test set (8 points).

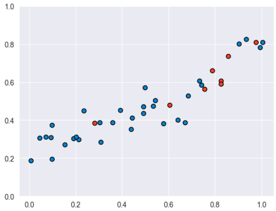

Figure: Data for polynomial fit, training set in blue

Fitting a Polynomial

A polynomial function will be fit to the data depicted below. The blue points are part of the training set (32 points) and the reddish-orange points are in the test set (8 points).
Using regression to fit a degree 18 polynomial to the data gives the curve depicted. The curve requires many local maxima and minima (in a small interval) to pass close to the training data. This requires some of the coefficients to have large absolute value.[2]

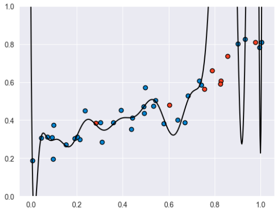

Figure: Degree 18 polynomial fit, no regularization

Several coefficients of the polynomial are in the millions. $\text{MSE}_{test} \approx 3.5$.

[2]The derivative needs to be relatively large and change sign quickly, over small changes in x. With many coefficients, they must be large in absolute value.

Regularization in Regression to Fit a Polynomial

One option for regularization while fitting a polynomial to the data:

- use gradient descent on the Mean Squared Error, but add a term of the form $\lambda|\mathbf{w}|^2$, for some constant λ.

Regularization in Regression to Fit a Polynomial

One option for regularization while fitting a polynomial to the data:

- use gradient descent on the Mean Squared Error, but add a term of the form $\lambda|\mathbf{w}|^2$, for some constant λ.

That is, with $f_{(\mathbf{w},b)}(x)$ being a degree 18 polynomial (non-constant coefficients from $\mathbf{w}$), have the loss function be

$$\mathcal{L}_{\mathcal{S}}(\mathbf{w}, b) = \lambda|\mathbf{w}|^2 + \frac{1}{n}\sum_{i=1}^{n}(f_{\mathbf{w},b}(x_i) - y_i)^2.$$

Regularization in Regression to Fit a Polynomial

One option for regularization while fitting a polynomial to the data:

- use gradient descent on the Mean Squared Error, but add a term of the form $\lambda|\mathbf{w}|^2$, for some constant λ.

That is, with $f_{(\mathbf{w},b)}(x)$ being a degree 18 polynomial (non-constant coefficients from $\mathbf{w}$), have the loss function be

$$\mathcal{L}_{\mathcal{S}}(\mathbf{w}, b) = \lambda|\mathbf{w}|^2 + \frac{1}{n}\sum_{i=1}^{n}(f_{\mathbf{w},b}(x_i) - y_i)^2.$$

Then, for $1 \le j \le d$ ($d =$ degree of the polynomial), the partial derivative $\frac{\partial}{\partial w_j}\mathcal{L}_{\mathcal{S}}$ is the same as in the non-regularized case except for one added term, $2\lambda w_j$.

Regularization in Regression to Fit a Polynomial

One option for regularization while fitting a polynomial to the data:

- use gradient descent on the Mean Squared Error, but add a term of the form $\lambda|\mathbf{w}|^2$, for some constant λ.

That is, with $f_{(\mathbf{w},b)}(x)$ being a degree 18 polynomial (non-constant coefficients from $\mathbf{w}$), have the loss function be

$$\mathcal{L}_{\mathcal{S}}(\mathbf{w}, b) = \lambda|\mathbf{w}|^2 + \frac{1}{n}\sum_{i=1}^{n}(f_{\mathbf{w},b}(x_i) - y_i)^2.$$

Then, for $1 \leq j \leq d$ ($d =$ degree of the polynomial), the partial derivative $\frac{\partial}{\partial w_j}\mathcal{L}_{\mathcal{S}}$ is the same as in the non-regularized case except for one added term, $2\lambda w_j$. This is called *Ridge regression*. There is also a closed form solution to minimize this loss function (in homework).

Regularization in Regression to Fit a Polynomial

One option for regularization while fitting a polynomial to the data:

- use gradient descent on the Mean Squared Error, but add a term of the form $\lambda|\mathbf{w}|^2$, for some constant λ.

That is, with $f_{(\mathbf{w},b)}(x)$ being a degree 18 polynomial (non-constant coefficients from **w**), have the loss function be

$$\mathcal{L}_{\mathcal{S}}(\mathbf{w}, b) = \lambda|\mathbf{w}|^2 + \frac{1}{n}\sum_{i=1}^{n}(f_{\mathbf{w},b}(x_i) - y_i)^2.$$

Then, for $1 \leq j \leq d$ ($d =$ degree of the polynomial), the partial derivative $\frac{\partial}{\partial w_j}\mathcal{L}_{\mathcal{S}}$ is the same as in the non-regularized case except for one added term, $2\lambda w_j$.
This is called *Ridge regression*. There is also a closed form solution to minimize this loss function (in homework).
In more general context, the adding of a penalty term like this is called either L_2 regularization or Tikhonov regularization.

Regularization in Regression to Fit a Polynomial

Implementing L_2 regularization, a degree 18 polynomial that is fit to this data gives the curve depicted below.

Regularization in Regression to Fit a Polynomial

Implementing L_2 regularization, a degree 18 polynomial that is fit to this data gives the curve depicted below.

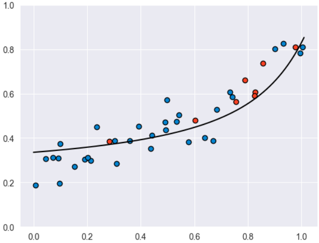

Figure: Degree 18 polynomial fit, with regularization

Regularization in Regression to Fit a Polynomial

Implementing L_2 regularization, a degree 18 polynomial that is fit to this data gives the curve depicted below.

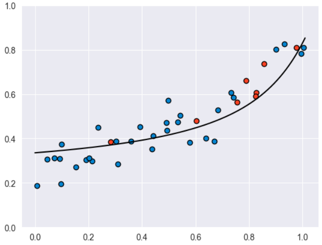

Figure: Degree 18 polynomial fit, with regularization

All coefficients in the newly fit degree 18 polynomial have absolute value that is less than 0.35. The MSE on the test data is about 0.004.

A different penalty term for regularization

Instead of adding a constant λ times the sum of squares $w_1^2 + w_2^2 + \ldots + w_d^2 = |\mathbf{w}|^2$ as a term in the loss, one can use a different penalty term.

A different penalty term for regularization

Instead of adding a constant λ times the sum of squares $w_1^2 + w_2^2 + \ldots + w_d^2 = |\mathbf{w}|^2$ as a term in the loss, one can use a different penalty term.