

Penalized regression

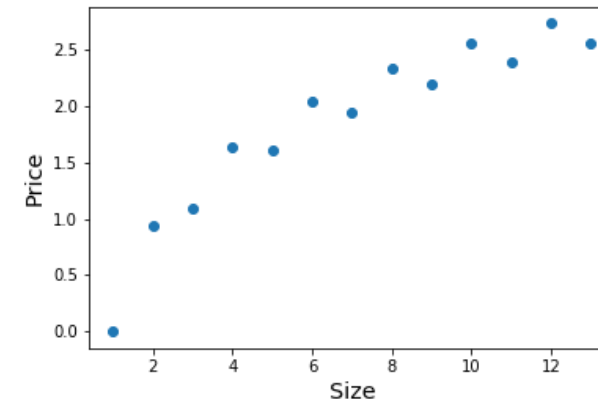
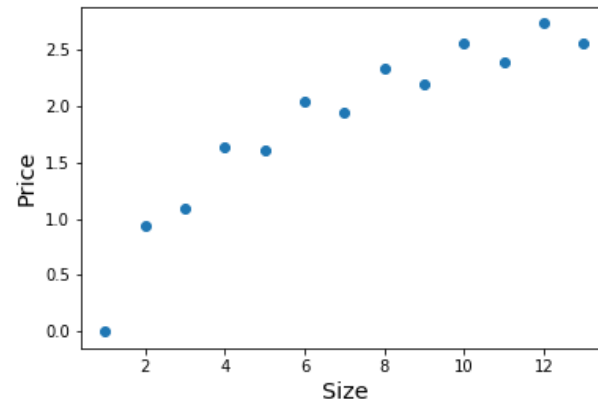
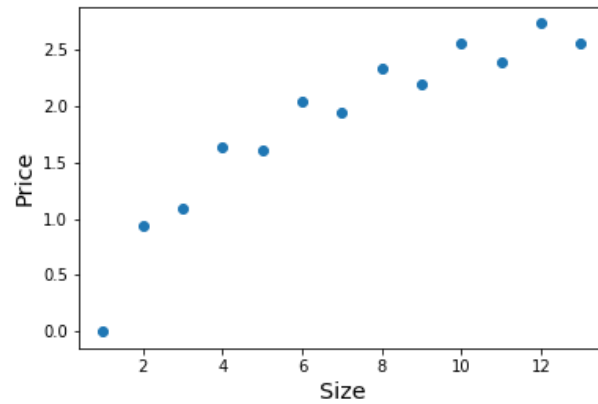
I/ Problems of overfitting and underfitting

- ▶ I/ Problems of overfitting and underfitting
- ▶ II/ Regularization techniques

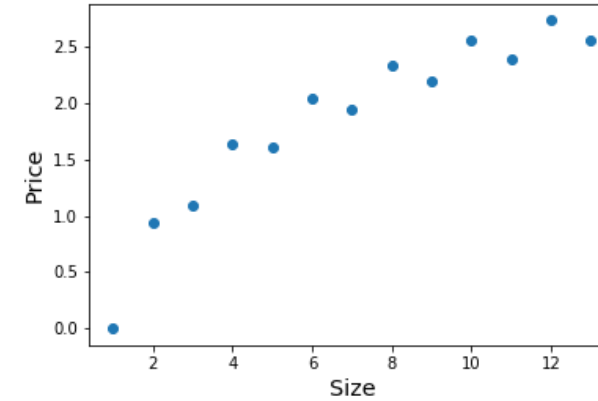
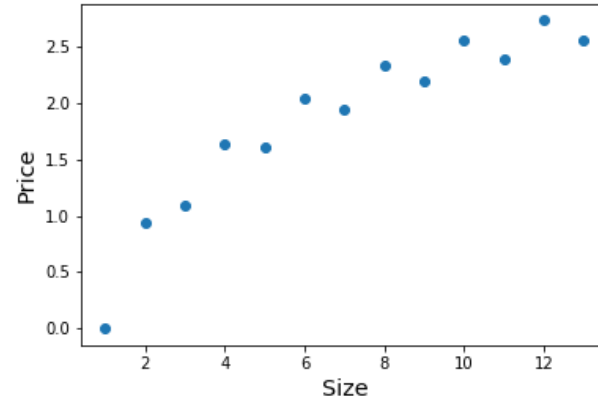
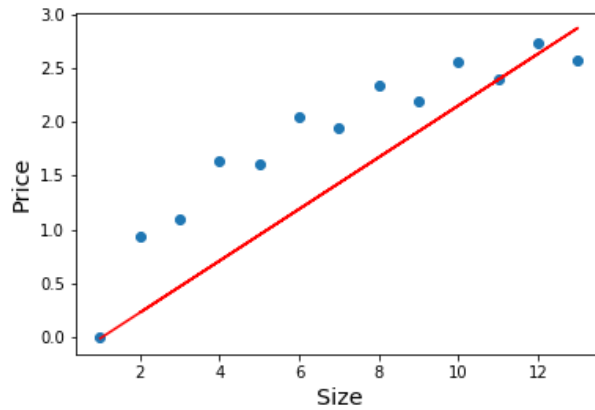
1. Overfitting & Underfitting

- ▶ When you use an algorithm of machine learning it is to create a model from training examples. But the aim is to apply your model on new data that your model have never seen.
- ▶ Your model is overfitted when your model perform well on your dataset, but have trouble to predict new data.
- ▶ Your model is underfitted when your model do not understand well your problem and have trouble to perform well on your training set.

1. For regression

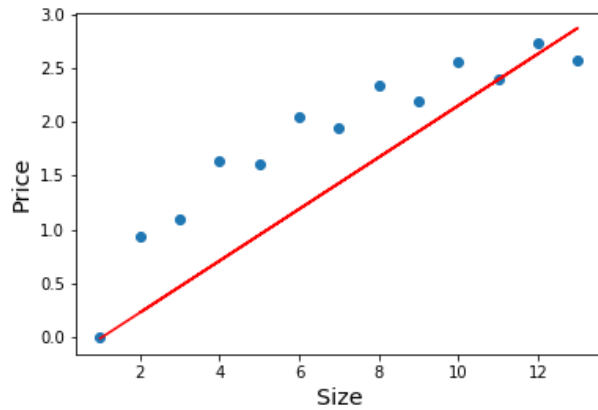


1. For regression

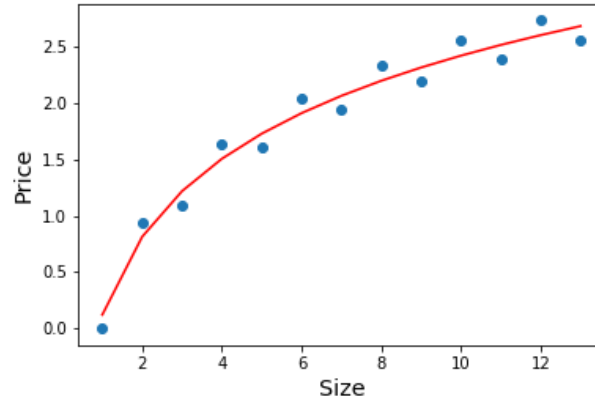


Underfitting or high bias

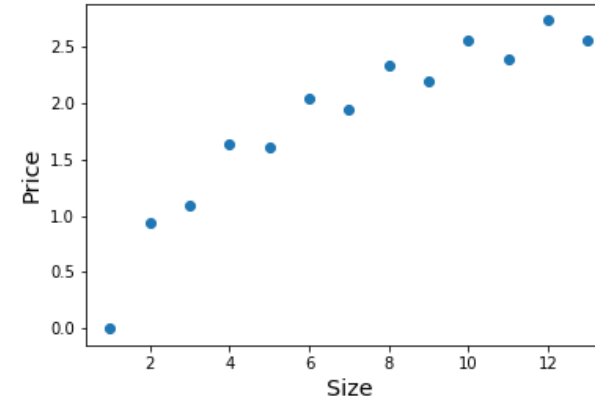
1. For regression



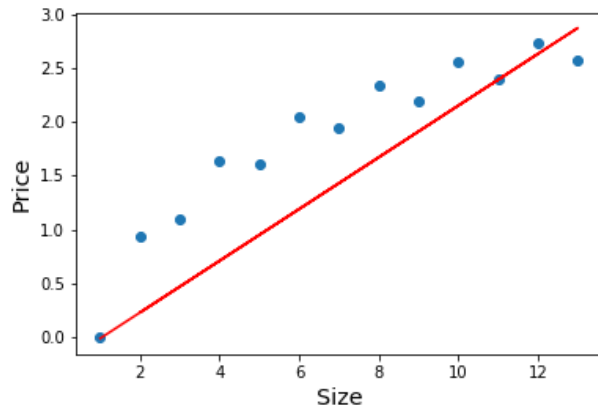
Underfitting or high bias



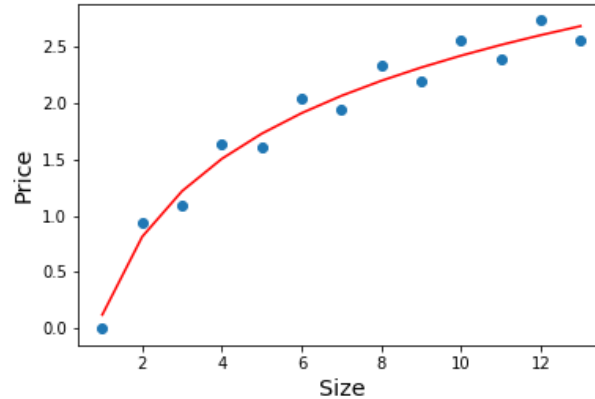
Good fitting



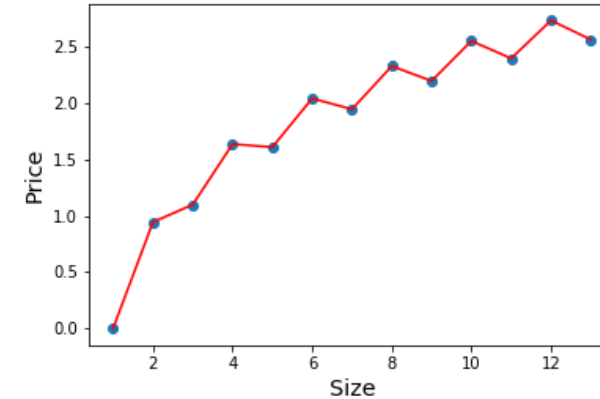
1. For regression



Underfitting or high bias



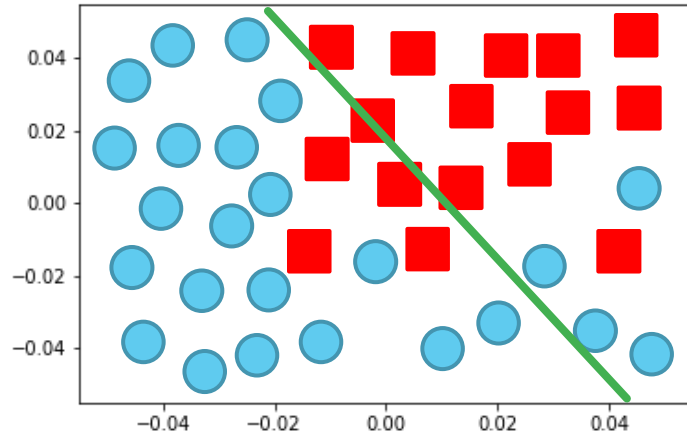
Good fitting



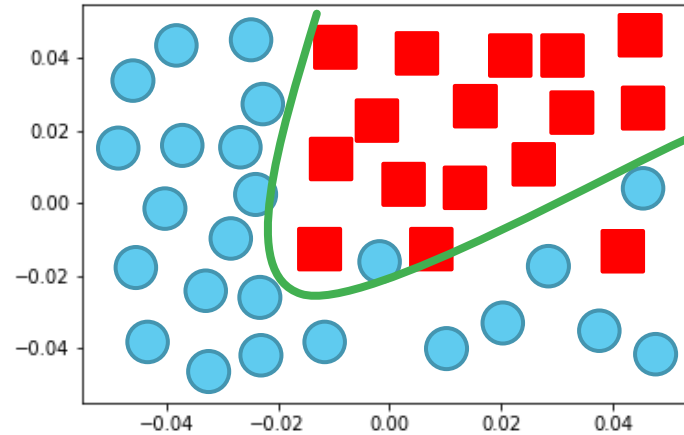
Overfitting Or High variance

Overfitting : If we have too many features, the learned hypothesis may fit the training set very well $J(W) = \frac{1}{2m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 \approx 0$, but fail to generalize to new examples. In our case fail to predict price for new houses.

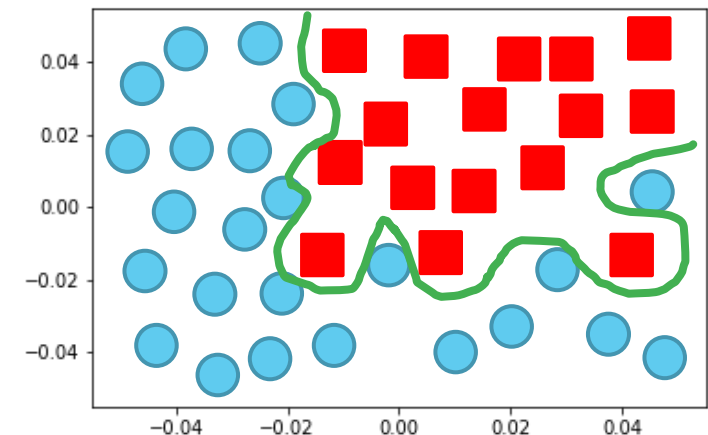
1. For classification



Underfitting or high bias



Good fitting



Overfitting Or High variance

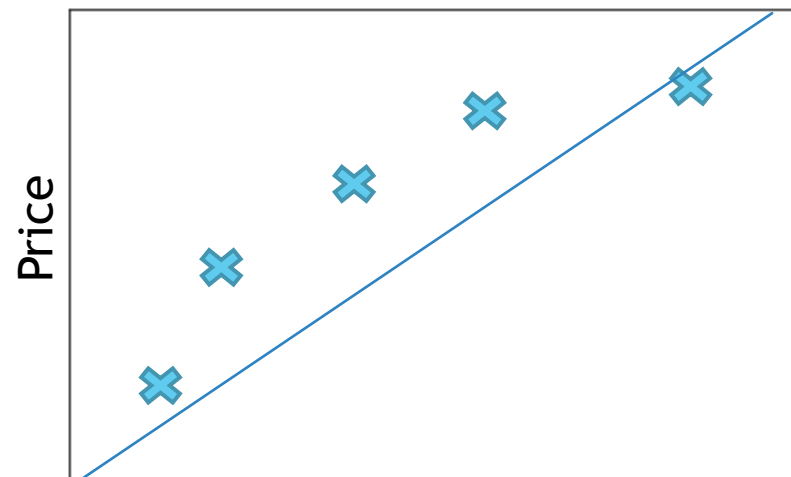
II/ Regularization techniques

- ▶ I/ Problems of overfitting and underfitting
- ▶ II/ Regularization techniques
 1. *L2 regularization*
 2. L1 regularization
 3. *Elastic-net*

3.1 Regularization

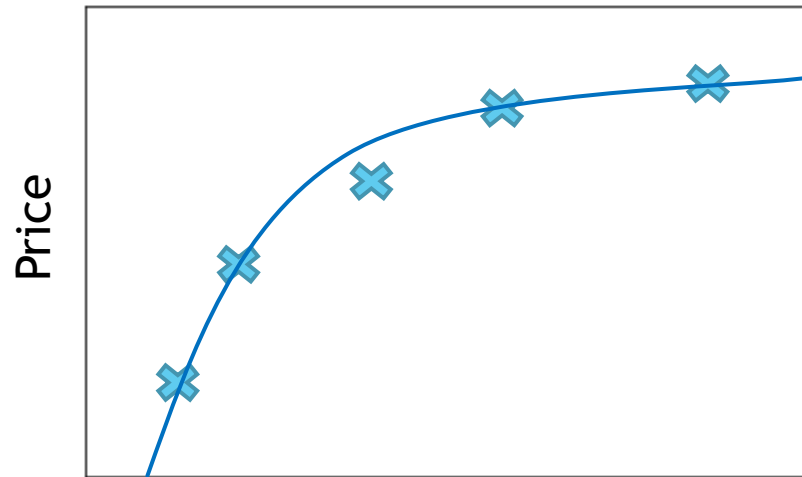
A central problem in machine learning is how to make an algorithm that will perform well not just on the training data, but also on new inputs. Many strategies used in machine learning are explicitly designed to reduce the test error, possibly at the expense of increased training error. These strategies are known collectively as regularization.

3.1 Intuition L2 regularization



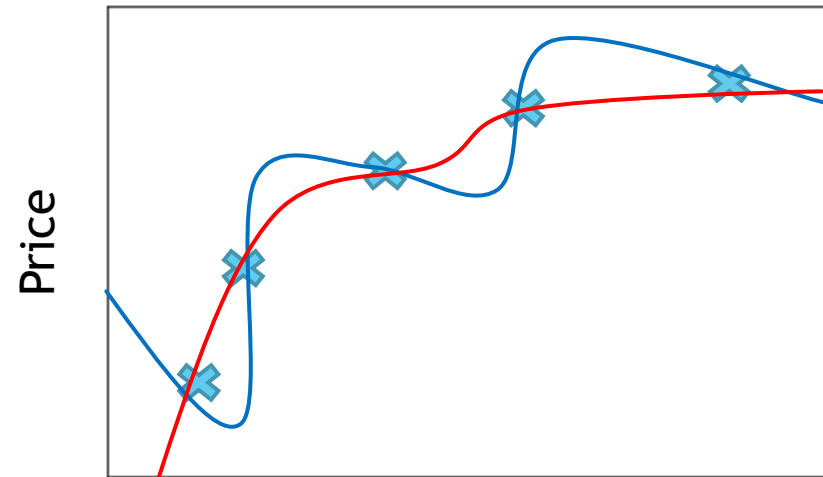
Size of house

$$w_0 + w_1 \cdot x$$



Size of house

$$w_0 + w_1 \cdot x + w_2 \cdot x^2$$



Size of house

$$w_0 + w_1 \cdot x + w_2 \cdot x^2 + w_3 \cdot x^3 + w_4 \cdot x^4$$

Suppose we penalize and make w_3, w_4 really small.

$$\min_w \frac{1}{2m} \sum_{i=1}^m (h(x^i) - y^i)^2 + 1000 \cdot w_3 + 1000 \cdot w_4$$

3.1 Cost function for ridge regression

- ▶ Model with “simpler” hypothesis with small values for parameters w_0, w_1, \dots, w_n are less prone to overfitting.

Regularization parameter

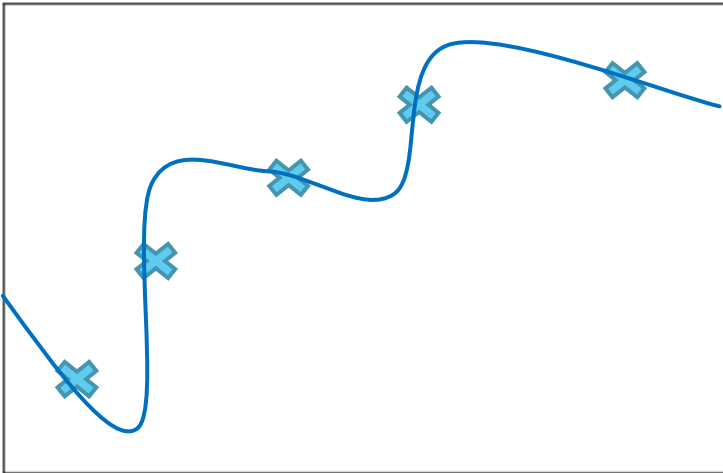
$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n w_j^2 \right]$$

Regularization term

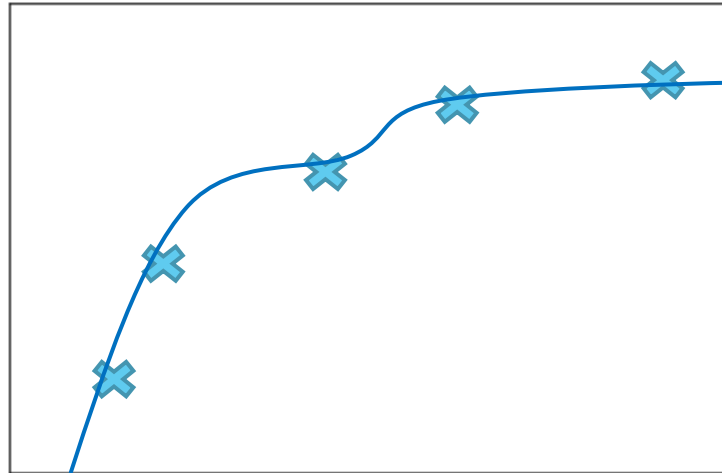
- ▶ With this regularization all of our parameters w_1, w_2, \dots, w_n will have a smaller magnitude and each one can help to predict the output.

3.1 Impact of lambda on the cost function

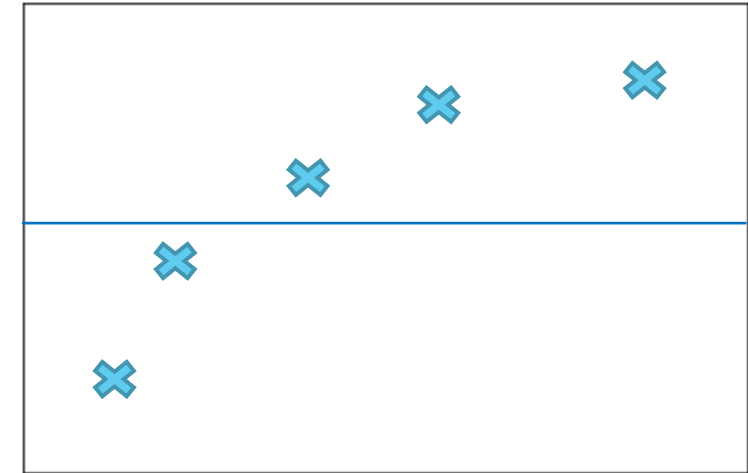
$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n w_j^2 \right]$$



λ equal to 0 -> overfitting



Perfect value for λ -> good fitting



λ too large -> underfitting

3.1 Gradient descent for ridge regression

Repeat until convergence

{

$$w_0 := w_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$w_j := w_j - \alpha \left[\frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_j^{(i)} - \frac{\lambda}{m} w_j \right]$$

}



$$w_j := w_j \underbrace{\left(1 - \alpha \frac{\lambda}{m} \right)}_{<1} - \alpha \frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_j^{(i)}$$

3.1 Cost function for ridge classification

- ▶ Model with “simpler” hypothesis with small values for parameters w_0, w_1, \dots, w_n are less prone to overfitting.

Regularization parameter

$$J(w) = - \left[\frac{1}{m} \sum_{i=1}^m y^{(i)} \log(h(x^{(i)})) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \right] + \boxed{\frac{\lambda}{2m} \sum_{j=1}^n w_j^2}$$

Regularization term

- ▶ With this regularization all of our parameters w_1, w_2, \dots, w_n will have a smaller magnitude and each one can help to predict the output.

3.1 Gradient descent for ridge classification

Repeat until convergence

{

$$w_0 := w_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$w_j := w_j - \alpha \left[\frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_j^{(i)} - \frac{\lambda}{m} w_j \right]$$

}

II/ Regularization techniques

- ▶ I/ Problems of overfitting and underfitting
- ▶ II/ Regularization techniques
 1. L2 regularization
 2. **L1 regularization**
 3. Elastic-net

3.1 Cost function for lasso regression

- ▶ Model with “simpler” hypothesis with small values for parameters w_0, w_1, \dots, w_n are less prone to overfitting.

Regularization parameter

$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda \sum_{j=1}^n |w_j| \right]$$

Regularization term

- ▶ With this regularization all of our parameters w_1, w_2, \dots, w_n will have a smaller magnitude and each one can help to predict the output.

3.1 Gradient descent for lasso regression

Repeat until convergence

{

$$w_0 := w_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$w_j := w_j - \alpha \left[\frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_j^{(i)} - \frac{\lambda}{2m} \text{sign}(w_j) \right]$$

}

3.1 Cost function for lasso classification

- ▶ Model with “simpler” hypothesis with small values for parameters w_0, w_1, \dots, w_n are less prone to overfitting.

$$J(w) = - \left[\frac{1}{m} \sum_{i=1}^m y^{(i)} \log(h(x^{(i)})) + (1 - y^{(i)}) \log(1 - h(x^{(i)})) \right] + \frac{\lambda}{2m} \sum_{j=1}^n |w_j|$$

Regularization parameter

Regularization term

3.1 Gradient descent for lasso classification

Repeat until convergence

{

$$w_0 := w_0 - \alpha \frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_0^{(i)}$$

$$w_j := w_j - \alpha \left[\frac{1}{m} \sum_{i=1}^m (h(x^{(i)}) - y^{(i)}) x_j^{(i)} - \frac{\lambda}{2m} \text{sign}(w_j) \right]$$

}

L1 and L2 regularization

- ▶ L1 and L2 regularization are very similar, they are both used to beat overfitting.
- ▶ Big difference between L1 and L2 regularization is that L2 can only shrink the slope asymptotically close to 0 while L1 can shrink the slope all the way to 0.
- ▶ If you have a lot of useless variables, L1 regularization can exclude them from the equation, it is little better than L2 regularization at reducing the variance. In contrast, L2 regularization tends to do little better when most variables are useful.

II/ Regularization techniques

- ▶ I/ Problems of overfitting and underfitting
- ▶ II/ Regularization techniques
 1. L2 regularization
 2. L1 regularization
 3. Elastic-net

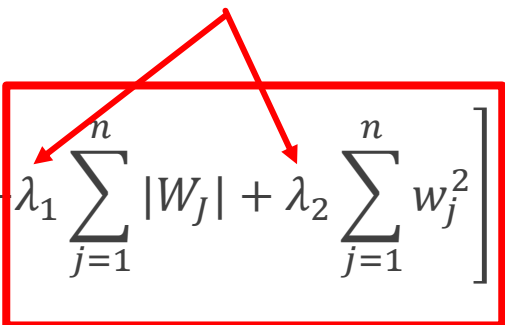
Cost function for Elasticnet

- ▶ Model with “simpler” hypothesis and homogeneous values for parameters w_0, w_1, \dots, w_n are less prone to overfitting.

Regularization parameter

$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda_1 \sum_{j=1}^n |w_j| + \lambda_2 \sum_{j=1}^n w_j^2 \right]$$

Regularization term



- ▶ The l1 regularization will select the most interesting parameters to make the model simpler and the l2 regularization will homogenize the values of these parameters w_1, w_2, \dots, w_n .

Cost function for Elasticnet

$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda_1 \sum_{j=1}^n |W_j| + \lambda_2 \sum_{j=1}^n w_j^2 \right]$$

$$J(w) = \frac{1}{2m} \left[\sum_{i=1}^m (h(x^{(i)}) - y^{(i)})^2 + \lambda \left[\alpha \sum_{j=1}^n |W_j| + \frac{1-\alpha}{2} \sum_{j=1}^n w_j^2 \right] \right]$$

α the mixing parameter or the l1_ratio between ridge ($\alpha = 0$) and lasso ($\alpha = 1$)

λ regularization parameter



Conclusions



1. L1 and L2 regularization

- ▶ L1 and L2 regularization are very similar, they are both used to beat overfitting.
- ▶ Big difference between L1 and L2 regularization is that L2 can only shrink the slope asymptotically close to 0 while L1 can shrink the slope all the way to 0.
- ▶ If you have a lot of useless variables, L1 regularization can exclude them from the equation, it is little better than L2 regularization at reducing the variance. In contrast, L2 regularization tends to do little better when most variables are useful.



2. Overfitting

- ▶ Add regularization (L1 or L2)
- ▶ Collect more data
- ▶ Reduce the number of features



3. Underfitting

- ▶ Change the type of model
- ▶ Create or collect new features