Name: Jay Joshi

Course: Applied Machine Learning

Student id: 200440993.

1 Centering and Ridge Regression

Assuming $\bar{x}=0$, so the input data has been centered. Show that optimizer of the following function is

$$J(\omega,\omega_0) = (y - HW - W_0 1)^T (y - HW - w_0 1) + \lambda w^T w$$

$$\hat{w}_0 = \bar{y}$$

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

solution: Since the input data is in normalized form, i.e $\bar{x}=0$ minimizing the following function

$$J(\omega,\omega_0) = (y - HW - W_0 1)^T (y - HW - w_0 1) + \lambda w^T w$$

is equivalent to minimizing

$$J(\omega, w_0) = \frac{1}{N} \sum_{i=1}^{N} (y_i - HW - W_0 I)^T (y_i - HW - W_0 I) + \lambda w^T w \quad \text{——— (1)}$$

Now, for finding optimal values of $\hat{w}_0$ & $\hat{w}$ we need to differentiate equation ——— (1) w.r.t to $\hat{w}_0$ and $\hat{w}$ respectively. and equate both the equations with zero.

Hence differentiating equation —(1) w.r.t w_0

$$\frac{\partial J(w, w_0)}{w_0} = -2\left(\frac{1}{N}\sum_{i=1}^{N} y_i - \hat{w}_0 I\right)$$

Now equating above equation with zero we get

$$-2\left(\frac{1}{N}\sum_{i=1}^{N} y_i - \hat{w}_0 I\right) = 0$$

$$= \frac{1}{N}\sum_{i=1}^{N} y_i = \hat{w}_0 \quad \text{Since } \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y} \text{ (mean)}$$

$$= \boxed{\hat{w}_0 = \bar{y}} \qquad \text{———— (2)}$$

Now differentiating equation —(1) w.r.t w

$$\frac{\partial J(w, w_0)}{w} = -2H^T(y - H\hat{w}) + 2\lambda I\hat{w} = 0$$

$$\therefore \quad -2H^T(y - H\hat{w}) + 2\lambda I\hat{w} = 0$$

$$\therefore \quad -2H^T y + 2H^T H\hat{w} + 2\lambda I\hat{w} = 0$$

$$\therefore \quad 2H^T H\hat{w} + 2\lambda I\hat{w} - 2H^T y = 0$$

$$\therefore \quad \hat{w}(H^T H + \lambda I) - H^T y = 0 \qquad \text{———— (3)}$$

$$\therefore \quad \boxed{\hat{w} = (H^T H + \lambda I)^{-1} H^T y}$$

From equation (2) and (3) we get

$$\hat{w}_0 = \bar{y} \quad \& \quad \hat{w} = (H^T H + \lambda I)^{-1} H^T y.$$

Q 2

Solution: The graph shows typical behaviour of training and test sample on error vs model complexity. You should always think of data as amalgation of information and noise

Data = Information + noise.

From the graph it is concluded that

→ Training error decreases as we increase model complexity. However with too much fitting, the model will start remembering data points, in other words it's start capturing noise. Hence the capability of model to "generalize" decreases along with increase in model complexity. Hence in this case we will have large prediction error on test sample i.e. variance will be high.

→ On the opposite hand, if the model complexity is low, the ability of the model to capture real world information is low, which resuts in large value of bias, hence model will not generalize well. This is called underfitting.

→ However, we are interested in models with low training and test error. Hence, we are looking for that minimum point on function of training and test error, so the ability of model for generalization will be more. As soon as complexity of model increases from that critical point the model will start capturing noise then information which will result in high variance.

Q3

Solution:

Fig a: Once the test error reaches to some constant value over increasing size of training sample, the sources which contributes to error are noise, bias and variance

In Fig a since the model is simple than the truth degree, initially with low number of training samples, the ability of model to capture information is less due to high bias, but has the number of training samples increases and test error reaches to some plateau, the test error is above noise level, hence due to structural difference model has higher error level than noise, bias is the contributor to some extent for this level of test error.

In Fig b: Since there is no structural difference between true and model degree, as we more in positive X direction on (x-axis) only noise is contributing to train and test error, once model has reached to some constant value. Bias and Variance i.e structural error is zero basically.

Scanned with Car

In fig c:

Since complexity of model is slightly more than true degree, during the initial (low set of training samples), the high test error on low number of samples is due to variance, but since model is exposed to all the possible set of examples, the only source of error is irreducible noise.

In fig d: Since complexity of model is too high than true degree, test error on initial set of examples is due to model overfitting i.e high variance than compared to model in fig c. But as the model learns all the possible examples i.e as training set size increases model starts generalizing and the source of error is irreducible noise after some critical constant value. i.e. size = 180.

It is important to note that error due to noise is irreducible in all above figures, hence even if model gets generalized, there is some amount of noise that contributes to train and test error.

Q4

Solution: The intercept w_0 in L1 and L2 regularization does not affects the complexity of the model, w_0 only affects the height of the function, it is not related to overfitting. Hence it should not be penalized.

→ If the input features are normalized i.e mean = 0, standard deviation = 1, then the cost function in L1 and L2 is updated to

L2 / Ridge Regression

$$Cost(w, w_0) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + w^T H(x_i)))^2 + \lambda w^T w$$

L1 / Lasso Regression

$$cost(w, w_0) = \frac{1}{N} \sum_{i=1}^{N} (y_i - (w_0 + w H(x_i)))^2 + \lambda \|w\|_1$$

In both the above updated cost function, the value of w_0 is not affected by the value of λ. Here optimal value of $\hat{w}_0 = \bar{y}$ (y-intercept) cost functions respectively. Hence by using above two cost functions respectively in L2 and L1 the intercept w_0 remains unaffected.

List of References :-

→ Machine Learning : A Probabilistic Perspective by kevin P Murphy

→ Elements of Statistical Learning by Hastie et al.