Weekly Report 1 - Linear Regression

Ganji Varshitha
AI20BTECH11009

Introduction

In Regression, we assume the target(i.e dependent) variable is a continuous function of input (independent) variable. When the function is linear in its parameters, it is called Linear Regression.

Training

Let $\mathbf{x}$ is a D dimensional vector

$$\hat{f}(\mathbf{x}) = y(\mathbf{x}, \mathbf{w}) = w_0 + w_1 x_1 + ... + w_D x_D \tag{1}$$

We need to estimate $\mathbf{w}$ inorder to minimise the prediction loss over entire population which is given by

$$\min_{\mathbf{w}} \mathbb{E}_{\mathbf{x},y}\left[L(\hat{f}(\mathbf{x}), y)\right] \tag{2}$$

If the cost function is mean-squared error ,

$$\mathbf{w}^* = \arg\min_{\mathbf{w}} ||X.\mathbf{w} - y||^2 \tag{3}$$

Ordinary Least Squares Method
Equating the derivative of $||X\mathbf{w} - y||^2$ to 0, we get

$$X^\top X\mathbf{w} = X^\top y \tag{4}$$

If $X^\top X$ is a positive definite matrix, we get optimal $\mathbf{w}$ to be

$$\mathbf{w}^* = (X^\top X)^{-1} X^\top y \tag{5}$$

Gradient descent Approach
We need to initialise weights and declare our hyperparameters which are number of iterations and learning rate.
For MSE error, we compute the gradient of cost function w.r.t the weights and update the weights in order to achieve minimum gradient.

$$\mathbf{w}^{(i+1)} = \mathbf{w}^i - \alpha \frac{\partial ||X.\mathbf{w}^i - y||^2}{\partial \mathbf{w}^i} \tag{6}$$

$$\mathbf{w}^{(i+1)} = \mathbf{w}^i - \alpha(-X(y - X\mathbf{w})) \tag{7}$$

We will loop till the minimum error threshold is reached or the gradient does not change on further iterations.

Key points

- The algorithm works if the dependent variables are linearly related with independent variables.
- It assumes the error terms to have constant variance and no correlation between one another.
- Since we need to compute the gradients, there is need to normalize the training data so that it does not take longer time.
- Learning rate decides how fast the algorithm converges.
- The model tends to overfit if there are many features in the dataset. Besides, there is a high chance correlation between some input variables.

Questions

1. Bias and variance in linear regression model are __________ related.
   A. directly B. inversely
   Solution:
   B. inversely

2. Explain the geometric and probabilistic interpretation of the model.
   Solution:
   Geometric Interpretation:
   The least-squares regression function is obtained by finding the orthogonal projection of the output vector y onto the subspace spanned by $x_1, x_2, x_3, \cdots, x_d$.
   Probabilistic Interpretation:
   Let us assume the target variable be given by the deterministic function with added gaussian noise. This gives

   $$p(y|X, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}^\top x_n, \beta^{-1}) \tag{8}$$

   where β is inverse variance of zero mean Gaussian random variable. We estimate the probability model using maximum likelihood which is same as minimising the least squares error.

   $$\arg\max_{\mathbf{w}} L = \arg\min_{\mathbf{w}} E \tag{9}$$

3. Do the missing values in the data affect the model?
   A. Yes B. No
   Solution:
   B. No

4. Is the model generative or discriminative? Explain why.
   Solution:
   Linear regression is a discriminative model. As seen in the probabilistic interpretation of the model above, we learn the parameters that maximises the conditional probability $\mathrm{P}(Y|X)$. This is the basic definition of discriminative model where we assume functional form of $\mathrm{P}(Y|X)$ and estimate its parameters.

5. When to prefer gradient descent approach to ordinary least squares method?
   Solution:
   Time complexity of least squares method is $\mathcal{O}(n^3)$ whereas time complexity of gradient descent is $\mathcal{O}(n)$. Gradient descent is preferred to ordinary least squares method when n i.e number of input features is greater than 10,000.

6. Which error function is sensitive to outliers?

   A. MSE(Mean squared error)

   B. MAE(Mean absolute error)

   Solution:
   A. MSE