Weekly Report 2- Naive Bayes

Ganji Varshitha
AI20BTECH11009

Introduction

Naive Bayes is a supervised learning algorithm which predicts the most probable class. It is very useful in text classification. Some of the applications include spam filtration, sentiment analysis, etc.

Why the name Naive Bayes?

It is called Bayes because it estimates the bayesian probability of a class.
Let us assume a data point x has n features $A_1, A_2, \cdots A_n$. The posterior probability $P(C|A_1, A_2, \cdots A_n)$ where C denotes class is given by bayes theorem:

$$P(C|A_1, A_2, \cdots A_n) = \frac{P(A_1, A_2, \cdots A_n|C)P(C)}{P(A_1, A_2, \cdots A_n)} \tag{1}$$

It is naive because we assume independence of features A_i when class is given.
This gives likelihood $P(A_1, A_2, \cdots A_n|C) = P(A_1|C)P(A_2|C)\cdots P(A_n|C)$.

Algorithm

From the above theorem, we know that we should choose a class which maximises posterior probability i.e $P(C|A_1, A_2, \cdots A_n)$.
Since $P(A_1, A_2, \cdots A_n)$ is same for all values of C, we can omit the marginal probability.

$$\hat{C} = \arg\max_C (P(A_1|C)P(A_2|C)\cdots P(A_n|C))P(C) \tag{2}$$

This is called Maximum A Posteriori estimation.

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

$$p(y|X, \mathbf{w}, \beta) = \prod_{n=1}^{N} \mathcal{N}(y_n|\mathbf{w}^\top x_n, \beta^{-1}) \tag{3}$$

   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.

$$\arg\max_{\mathbf{w}} L = \arg\min_{\mathbf{w}} E \tag{4}$$

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 P($Y|X$). This is the basic definition of discriminative model where we assume functional form of 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