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.
We also assume all the features contribute equally to the outcome.

Types of Naive Bayes Classifier

Multinomial Naive Bayes

It is mostly used in document classification where the feature values include the frequencies of words present in the document.

Bernoulli Naive Bayes

It involves Bernoulli distribution. The feature values involve binary values. It is used in spam filtration and to find whether a word is present in the document.
The decision rule is based on

$$P(x_i \mid y) = P(i \mid y)x_i + (1 - P(i \mid y))(1 - x_i) \tag{3}$$

Gaussian Naive Bayes

The feature values are continuous values and are assumed to be sampled from Gaussian distribution. We use maximum likelihood to estimate the parameters of the distribution with the given data. Using the probability distribution, conditional probability is given by

$$P(x_i \mid y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\Big(- \frac{(x_i - \mu_y)^2}{2\sigma_y^2}\Big) \tag{4}$$

Key points

- The assumptions made in the algorithm does not always hold true in real-life problems.
- It is robust to isolated noise points and irrelevant attributes.
- To prevent floating point underflow, we sum logs of probabilities rather than multiplying.
  Class with highest final un-normalized log probability score is still the most probable.
- Algorithm gives probability distribution over all classes.
- Since the algorithm is generative, we can handle missing data by ignoring the instance during probability estimation.

Questions

1. Naive Bayes comes under _____ learning.

   A. Inductive

   B. Transductive

   Solution:
   A. Inductive

2. Write the decision rule in terms of likelihood ratio for binary classification problem.
Solution:

$$\begin{aligned} g(x) &= \frac{P(C_1 \mid x)}{P(C_2 \mid x)} \\ &= \frac{P(x \mid C_1)P(C_1)}{P(x \mid C_2)P(C_2)} \end{aligned}$$

The decision rule is

$$\text{choose} \begin{cases} C_1 & \text{if } g(x) > 1 \\ C_2 & \text{otherwise} \end{cases}$$

3. Is the model log linear?
Solution:
Yes.

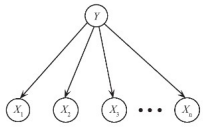
4. Which of the following distribution has pdf?

   A. Bernoulli distribution

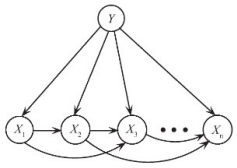
   B. Gaussian distribution

Solution:
B. Gaussian distribution

5. Explain Bayesian networks.
Solution:
Bayesian network is a probabilistic graphical model that represents a set of variables and their conditional dependencies via a directed acyclic graph (DAG).



(a) Naive Bayesian network



(b) Bayesian network with k dependencies