Naive Bayes

DataDNA

Introduction

- **<u>Naive Bayes</u>** is a Conditional Probability Based Algorithm
- The features provided to us are independent and do not affect each other
- Naive Bayes is a classification algorithm suitable for binary & multiclass classification.
- The calculation of the likelihood of different class values involves multiplying a lot of numbers together.
- When new data comes, the algorithm updates the probabilities of the model. This can be helpful if the data changes frequently
- The assumption taken for numerical variables is Normal Distribution

Features of Naive Bayes

- It’s a classification algorithm
- Purely based on Conditional Probability
- Quick calculation time
- Gives Probabilistic Outputs
- It requires lesser data compared to other algorithms

How Naive Bayes Works

Bayes' theorem is stated mathematically as the following equation:

Conditional Probability: Bayes' Theorem

$$P(A|B) = \frac{P(B|A)\,P(A)}{P(B)}$$

- Basically, we are trying to find probability of event A, given the event B is true. Event B is also termed as **evidence**.
- P(A) is the **priori** of A (the prior probability, i.e. Probability of event before evidence is seen). The evidence is an attribute value of an unknown instance(here, it is event B).
- P(A|B) is a posteriori probability , i.e. probability of event after evidence is seen.

Bayes Equation Derivation

1 → P(A/B) = P(AnB) / P(B)

2 → P(B/A) = P(BnA) / P(A)

1 can be written as → P(AnB) = P(A/B) * P(B) → **3**

2 can be written as → P(BnA) = P(B/A) * P(A) → **4**

Equating 3 and 4

P(A/B) * P(B) = P(B/A) * P(A)

Hence,

P(A/B) = (P(B/A) * P(A)) / P(B)

P(A/B) → Posterior Probability
P(B/A) → Likelihood Probability
P(A) → Class Prior Probability
P(B) → Predictor's Prior Probability

How Naive Bayes Works

Play	Counts	Probability	
Yes	9	2/3	0.642857
No	5	1/3	0.357143
Total	14	100%	100%

Outlook	YES	NO	P(YES)	P(NO)
Sunny	2	3	2/9	3/5
Overcast	4	0	4/9	0
Rainy	3	2	1/3	2/5
Total	9	5	100%	100%

Temperature	YES	NO	P(YES)	P(NO)
Hot	2	2	2/9	2/5
Mild	4	2	4/9	2/5
Cool	3	1	1/3	1/5
Total	9	5	100%	100%

Humidity	YES	NO	P(YES)	P(NO)
High	3	4	1/3	4/5
Normal	6	1	2/3	1/5
Total	9	5	100%	100%

Wind	YES	NO	P(YES)	P(NO)
FALSE	6	2	2/3	2/5
TRUE	3	3	1/3	3/5
Total	9	5	100%	100%

Model Evaluation Metrics

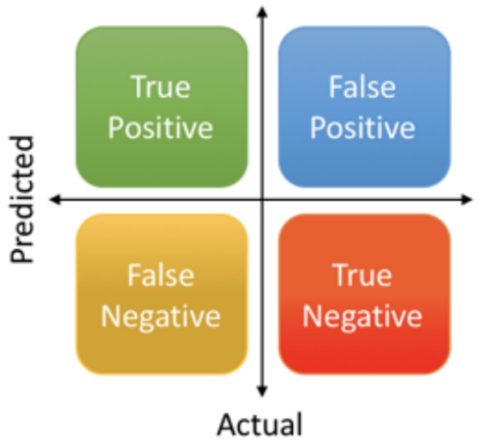
The various metrics used to evaluate
 the results of the prediction are :

1.Accuracy
2.Precision
3.Recall
4.Area Under ROC Curve

$$\textbf{Precision} = \frac{\text{True Positive}}{\text{Actual Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}}$$

$$\textbf{Recall} = \frac{\text{True Positive}}{\text{Predicted Results}} \quad \text{or} \quad \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}}$$

$$\textbf{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{Total}}$$



Naive Bayes-Cons

- The main limitation of Naive Bayes is the assumption of independent predictor features.
- If there is a new category in test dataset compared to Training, it assigns a probability of ZERO
- Data Scarcity leads to an issue
- Since continuous features are involved, the algorithm does binning on top of it so loss of information exists
- The assumption of Independent Variables does not practically exists