Naive Bayes on Tennis Weather Dataset

Using the weather conditions provided, we have to determine whether or not to play tennis using *Naive Bayes Algorithm*- implemented from scratch.

Here, we use the Tennis Weather Dataset from Kaggle.

The completed Jupyter Notebook is available here.

Bayes Theorem

The underlying principle behind the Naive Bayes algorithm is the *Bayes Theorem*.

Bayes Theorem states that-

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

If X is the input variables and y is the output variable, we can rewrite the above equation as-

$$P(y|X) = \frac{P(X|y) * P(y)}{P(X)}$$

The "naive" part of the algorithm is that we make the naive assumption that the classes are conditionally independent.

That is, the effect of a predictor(x_1) on a given class(y) is independent of the values of other predictors($x_2, x_3 \ldots$).

We can therefore rewrite $P(X|y)$ as-

$$P(X|y) = P(x1|y) * P(x2|y) * \ldots * P(x(n)|y)$$

We can remove the denominator $P(X)$ -as it remains constant while solving for y- and introduce a proportionality.

$$P(y|X) \propto P(X|y) * P(y)$$

OR

$$P(y|X) \propto P(x1|y) * P(x2|y) * \ldots * P(x(n)|y) * P(y)$$

This is the basic idea behind the Naive Bayes algorithm.

Tennis Weather Dataset

The dataset consists of weather conditions such as *outlook*, *temp*, *humidity* and *windy* and contains a label *play* which tells whether tennis was played in the given conditions.

outlook	temp	humidity	windy	play
sunny	hot	high	false	no
sunny	hot	high	true	no
overcast	hot	high	false	yes
rainy	mild	high	false	yes
rainy	cool	normal	false	yes
rainy	cool	normal	true	no
overcast	cool	normal	true	yes
sunny	mild	high	false	no
sunny	cool	normal	false	yes
rainy	mild	normal	false	yes
sunny	mild	normal	true	yes
overcast	mild	high	true	yes
overcast	hot	normal	false	yes
rainy	mild	high	true	no

Dataset downloaded from Kaggle

There are 14 days in the dataset and based on the weather conditions, you have to decide whether or not to play tennis.

Using the dataset, we will first create a frequency table to get the values of $P(X|y)$ which we can use to solve for $P(y|X)$.

outlook	play = yes	play = no
sunny	2/9	3/5
overcast	4/9	
rain	3/9	2/5

temp	play = yes	play = no
hot	2/9	2/5
mild	4/9	2/5
cold	3/9	1/5

humidity	play = yes	play = no
high	3/9	4/5
normal	6/9	1/5

wind	play = yes	play = no
✓	3/9	3/5
✗	6/9	2/5

Frequency table

Making predictions using Naive Bayes

We will now use the Naive Bayes algorithm to find the probability of playing tennis given the weather conditions.

For example, to calculate the probability that you should play tennis for the following conditions:

- *outlook*- sunny
- *temp*- mild
- *humidity*- normal
- *windy*- False

We will calculate,

$$\begin{aligned} P(y = "yes"|X = [sunny,\ mild,\ normal,\ False]) &= P(outlook = "sunny"|y = "yes") \\ &* P(temp = "mild"|y = "yes") \\ &* P(humidity = "normal"|y = "yes") \\ &* P(windy = "False"|y = "yes") \\ &* P(y = "yes") \end{aligned}$$

$$= \frac{2}{9} * \frac{4}{9} * \frac{6}{9} * \frac{6}{9} * \frac{9}{14} = 0.282$$

Similarly,

$$P(y = "no"|X = [sunny,\ mild,\ normal,\ False]) = 0.0069$$

Since $P(y = "yes"|X) > P(y = "no"|X)$ the prediction would be to play tennis.

We obtain all these values from the frequency table.