Adding Naive Bayes to the Study Guide

Naive Bayes is a simple yet powerful classification algorithm based on Bayes' Theorem, often used in machine learning and data classification tasks.

What is Naive Bayes?

Naive Bayes is a probabilistic algorithm that applies Bayes' Theorem under the assumption that the features are independent of each other. This assumption simplifies computations and is called the "naive" assumption.

Formula for Naive Bayes

Given a set of features $X = \{x_1, x_2, \ldots, x_n\}$, the probability of a class C given the features is:

$$P(C|X) = \frac{P(X|C) \cdot P(C)}{P(X)}.$$

Since $P(X)$ is constant for all classes, we focus on the numerator:

$$P(C|X) \propto P(X|C) \cdot P(C).$$

For independent features:

$$P(X|C) = P(x_1|C) \cdot P(x_2|C) \cdot \ldots \cdot P(x_n|C).$$

Steps in Naive Bayes Classification

1. **Calculate Prior Probability**:

$$P(C) = \frac{\text{Number of samples in class } C}{\text{Total number of samples}}.$$

2. **Calculate Likelihood**: For each feature x_i:

$$P(x_i|C) = \frac{\text{Number of samples where } x_i \text{ and } C \text{ occur}}{\text{Number of samples in class } C}.$$

3. **Compute Posterior Probability**: Using:

$$P(C|X) \propto P(C) \cdot P(x_1|C) \cdot P(x_2|C) \cdot \ldots \cdot P(x_n|C).$$

4. **Classify**: Assign X to the class with the highest posterior probability.

Step-by-Step Example

Scenario:

A weather dataset is used to predict if we should play tennis. The dataset includes the following features:

- **Outlook**: {Sunny, Overcast, Rainy}.
- **Temperature**: {Hot, Mild, Cool}.
- **Humidity**: {High, Normal}.
- **Wind**: {Weak, Strong}.
- **PlayTennis**: {Yes, No}.

Dataset:

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Weak	No

Outlook	Temperature	Humidity	Wind	PlayTennis
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rainy	Mild	High	Weak	Yes
Rainy	Cool	Normal	Weak	Yes
Rainy	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rainy	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rainy	Mild	High	Strong	No

Question:

Predict if we should play tennis ($PlayTennis = Yes/No$) for the following conditions:

- Outlook = Sunny,
- Temperature = Cool,
- Humidity = High,
- Wind = Strong.

Step 1: Calculate Prior Probabilities

$$P(\text{Yes}) = \frac{\text{Number of Yes}}{\text{Total samples}} = \frac{9}{14} = 0.643.$$

$$P(\text{No}) = \frac{\text{Number of No}}{\text{Total samples}} = \frac{5}{14} = 0.357.$$

Step 2: Calculate Likelihood for Each Feature

1. **Outlook**:
   - $P(\text{Sunny}|\text{Yes}) = \frac{2}{9} = 0.222.$
   - $P(\text{Sunny}|\text{No}) = \frac{3}{5} = 0.6.$
2. **Temperature**:
   - $P(\text{Cool}|\text{Yes}) = \frac{3}{9} = 0.333.$
   - $P(\text{Cool}|\text{No}) = \frac{1}{5} = 0.2.$
3. **Humidity**:
   - $P(\text{High}|\text{Yes}) = \frac{3}{9} = 0.333.$
   - $P(\text{High}|\text{No}) = \frac{4}{5} = 0.8.$
4. **Wind**:
   - $P(\text{Strong}|\text{Yes}) = \frac{3}{9} = 0.333.$
   - $P(\text{Strong}|\text{No}) = \frac{3}{5} = 0.6.$

Step 3: Compute Posterior Probabilities

1. For $PlayTennis = Yes$:

$$P(\text{Yes}|\text{X}) \propto P(\text{Yes}) \cdot P(\text{Sunny}|\text{Yes}) \cdot P(\text{Cool}|\text{Yes}) \cdot P(\text{High}|\text{Yes}) \cdot P(\text{Strong}|\text{Yes})$$

$$P(\text{Yes}|\text{X}) \propto 0.643 \cdot 0.222 \cdot 0.333 \cdot 0.333 \cdot 0.333 = 0.0157.$$

2. For $PlayTennis = No$:

$$P(\text{No}|\text{X}) \propto P(\text{No}) \cdot P(\text{Sunny}|\text{No}) \cdot P(\text{Cool}|\text{No}) \cdot P(\text{High}|\text{No}) \cdot P(\text{Strong}|\text{No})$$

$$P(\text{No}|\text{X}) \propto 0.357 \cdot 0.6 \cdot 0.2 \cdot 0.8 \cdot 0.6 = 0.0206.$$

Step 4: Classify

- Compare $P(\text{Yes}|\text{X}) = 0.0157$ and $P(\text{No}|\text{X}) = 0.0206$.
- Since $P(\text{No}|\text{X}) > P(\text{Yes}|\text{X})$, the prediction is $PlayTennis = No$.

Advantages of Naive Bayes

1. Easy to implement and computationally efficient.
2. Works well with large datasets.
3. Performs well in text classification tasks (e.g., spam detection).

Disadvantages of Naive Bayes

1. Assumes feature independence, which is rarely true in real-world scenarios.
2. Struggles with small datasets or when probabilities are very low (e.g., zero-frequency problem).