,

FADML Scribe

Pabbaraju Harsha

(24BM6JP38)

Indian Institute of Technology Kharagpur

Bayesian Learning and Naive Bayes Classification

1 Introduction

Bayesian Learning is a fundamental approach in probabilistic modeling that allows us to determine the probability of an outcome given a set of observed variables. This method is widely used in machine learning, decision theory, and artificial intelligence due to its ability to incorporate prior knowledge and update beliefs as new data become available.

The core objective of Bayesian Learning is to compute the posterior probability:

$$P(Y|X_1, X_2, \ldots, X_n)$$

where Y is the target variable (e.g. pass/fail, spam/non-spam), and $X_1, X_2, ..., X_n$ represents the observed features. The ability to make probabilistic inferences using Bayes' Theorem makes this approach particularly powerful in scenarios where data are limited or uncertain.

1.1 Why Bayesian Learning?

To compute $P(Y|X_1, ..., X_n)$, consider the **joint probability distribution** of all variables. This distribution captures the dependencies between all the features and the target variable.

Consider a dataset with n binary categorical attributes such as **Gender**, **Work Hours**, **Economic Status (Poor/Rich)**, etc. Each of the attributes has only 2 values. Then the probability distribution can be represented as:

Gender	**Work Hours**	**...**	**...**	**Economic Status**	**Probability**
Male	more than 40	....	....	Rich (R)	0.1
Female	less than 40	....	....	Poor (P)	0.15

If we have n binary features, the number of combinations of off-values needed to fill the table is $2^n - 1$ values, which is computationally expensive.

1.2 Issues with the Direct Estimation

- **Data Sparsity:** Many feature combinations may not appear in the training data, making the probability estimation unreliable.
- **Exponential Growth of Parameters:** Without simplifying assumptions, estimating $2^n - 1$ probabilities becomes infeasible for large n. For example, with 10 binary features, we need to have 1023 probability values.

To overcome these challenges, we turn to **Bayesian parameter estimation techniques** that use prior knowledge to make probability estimation feasible.

2 Bayesian Inference and Conditional Independence

Now, let us see if we can reduce the values needed for a n-binary feature dataset using Bayes' rule.

$$P(Y|X_1, ..., X_n) = \frac{P(Y)P(X_1, ..., X_n|Y)}{P(X_1, ..., X_n)}$$

$$P(Y|X_1, ..., X_n) = \frac{P(Y)P(X_1, ..., X_n|Y)}{P(Y=1)P(X_1, ..., X_n|Y=1) + P(Y=0)P(X_1, ..., X_n|Y=0)}$$

So, now we need $2^n - 1$ estimations for each for $P(Y = 1)$, $P(X_1, ..., X_n|Y = 1)$, and $P(Y = 0)$, $P(X_1, ..., X_n|Y = 0)$. Then, two more estimations are combined, leading to a total estimations of $2(2^n - 1) + 2$ **estimations.**

Interestingly, applying Bayes' rule initially increases the number of required estimations rather than reducing them.

To simplify computations, we assume **conditional independence** between features given Y:

$$P(X_1, ..., X_n|Y) = \prod_{i=1}^{n} P(X_i|Y)$$

This assumption forms the basis of the **Naive Bayes classifier**, reducing the number of required estimates **from** $2(2^n - 1) + 2$ **to** $2n + 2$.

2.1 Conditional Independence Breakdown

$$P(X_1|X_2, Y) = P(X_1|Y) \tag{1}$$

This implies that X_1 and X_2 are conditionally independent given Y, which can be represented as:

$$X_1 \perp X_2 \mid Y$$

We factorize the probabilities as below:

$$P(X_1, X_2|Y) = P(X_1|Y)P(X_2|Y) \tag{2}$$

Using the definition of conditional probability,

$$P(X_1, X_2, Y) = P(X_1|Y)P(X_2|Y)P(Y) \tag{3}$$

Dividing both sides by $P(Y)$:

$$P(X_1, X_2) = \frac{P(X_1|Y)P(X_2|Y)P(Y)}{P(Y)} \tag{4}$$

Since $P(Y)$ cancels out in numerator and denominator:

$$P(X_1, X_2) = P(X_1|Y)P(X_2|Y) \tag{5}$$

3 Naive Bayes Classification - Solving Data Sparsity

Let us better understand this concept with an example.

Consider a classification problem where we want to predict whether a student will pass or fail based on the following 3 attributes - **Attendance (Poor(P)/Average(A)/High(H)), Reads (Y/N), and Assignment Solving (Low(L)/Medium(M)/High(H))**.

Using Naive Bayes, we compute:

$$P(+|A, N, M) = \frac{P(+) \times P(A|+) \times P(N|+) \times P(M|+)}{P(M|+)} \tag{6}$$

Expanding the denominator:

$$P(+)\left(P(A|+)P(N|+)P(M|+)\right) + P(-)P(A|-)P(N|-)P(M|-) \tag{7}$$

3.1 Training Data Table

Attendance	Reads	Assignment Solving	Grade
H	Y	M	+
A	N	L	-
A	Y	H	+
P	Y	L	-
P	N	H	-
H	N	M	+

3.2 Test Case

We need to predict the grade for the following test case:

$$(A, N, M) \text{ (Attendance = A, Reading = N, Assignment Solving = M)}$$

We use:

$$P(+) = \frac{1}{2}, \quad P(-) = \frac{1}{2} \tag{8}$$

$$P(A|+) = \frac{P(A,+)}{P(+)} = \frac{1/6}{1/2} = \frac{1}{3} \tag{9}$$

$$P(N|+) = \frac{1}{2}, \quad P(M|+) = \frac{2}{3} \tag{10}$$

Similarly,

$$P(A|-) = \frac{1}{3}, \quad P(N|-) = \frac{2}{3}, \quad P(M|-) = \frac{0}{3} \tag{11}$$

Thus, using these values, we can compute the following:

$$P(+|A, N, M) = \frac{P(+)P(A|+)P(N|+)P(M|+)}{P(A, N, M)} \tag{12}$$

$$P(+|A, N, M) = \frac{P(+)P(A|+)P(N|+)P(M|+)}{P(+)P(A|+)P(N|+)P(M|+) + P(-)P(A|-)P(N|-)P(M|-)} \tag{13}$$

$$P(+|A, N, M) = \frac{(1/2).(1/3).(1/2).(2/3)}{(1/2).(1/3).(1/2).(2/3) + (1/2).(1/3).(2/3).(0)} \tag{14}$$

$$P(+|A, N, M) = 1 \tag{15}$$

4 Decision Boundaries and Log-Linear Models

4.1 Decision Boundary in Naive Bayes

To classify an observation, we compare the posterior probabilities.

$$\frac{P(Y|X_1, X_2, ..., X_n)}{P(\overline{Y}|X_1, ..., X_n)} \geq 1 \tag{16}$$

Taking the natural logarithm:

$$\ln\left(\frac{P(Y)\prod_{i=1}^{n} P(X_i|Y)}{P(\overline{Y})\prod_{i=1}^{n} P(X_i|\overline{Y})}\right) \geq 0 \tag{17}$$

Rewriting:

$$\ln P(Y) - \ln P(\overline{Y}) + \sum_{i=1}^{n} \ln\left(\frac{P(X_i|Y)}{P(X_i|\overline{Y})}\right) \geq 0 \tag{18}$$

Since $\ln P(Y) - \ln P(\overline{Y})$ is a constant, we define:

$$C = \ln P(Y) - \ln P(\overline{Y}) \tag{19}$$

Thus, the decision boundary is given by:

$$C + \sum_{i=1}^{n} \ln\left(\frac{P(X_i|Y)}{P(X_i|\overline{Y})}\right) \geq 0 \tag{20}$$

4.2 Log-Linear Model Interpretation

This shows that the Naive Bayes classifier produces a log-linear model, meaning:

- Probabilities are transformed into log-space to form a linear decision boundary.
- When mapped back to the original probability space, the boundary can become non-linear.

4.3 Real-World Applications

Naive Bayes has been extensively researched in applications such as:

- Spam Filtering: The most common application of Bayes' classifier is in spam filtering and text classification. All our emails are all built on top of a Bayes' classifier model
- Medical Image Diagnostics Cancer Research: Another application is to classify cancer status of a person based on their scans like MRI.

5 Gaussian Naive Bayes for Continuous Features

- Up to this point, we have primarily dealt with **discrete-valued features**. However, many real-world applications involve features that vary continuously over a range.
- To extend **Naive Bayes classification** to such cases, we assume that each feature follows a **specific probability distribution**.
- The most common type of distribution is Gaussian (Normal) Distribution:

$$P(X_i|Y=k) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(X-\mu)^2}{2\sigma^2}} \tag{21}$$

  where:

  - μ is the mean of the feature for class k
  - σ^2 is the variance of the feature for class k

- For predicting for a new data point, we just substitute the values of X for each of the features and find the class which has highest probability

5.1 Classification Problem: Basketball Player Classification

- To demonstrate **Gaussian Naive Bayes classification**, consider a dataset where students are described using two **continuous** features:
  1. **Height (H)**
  2. **Marks (M)**
- We want to classify whether a student is a basketball player (BB) or not

  Using Bayes' Theorem:

$$P(BB|H,M) = \frac{P(H|BB)P(M|BB)P(BB)}{P(H,M)} \tag{22}$$

  Expanding the denominator:

$$P(H,M) = P(H,M|BB)P(BB) + P(H,M|\overline{BB})P(\overline{BB}) \tag{23}$$

$$P(BB|H,M) = \frac{P(H|BB)P(M|BB)P(BB)}{P(H,M|BB)P(BB) + P(H,M|\overline{BB})P(\overline{BB})} \tag{24}$$

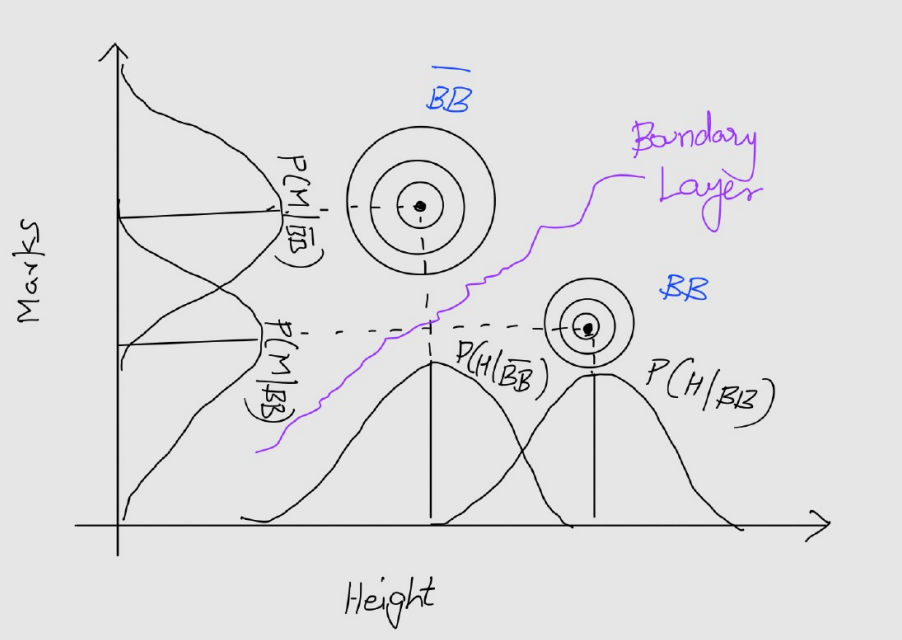

Figure 1: Gaussian Distributions and Decision Boundary for Basketball Player Classification

Alternatively, we can use the ratio:

$$\frac{P(BB)P(H|BB)P(M|BB)}{P(\overline{BB})P(H|\overline{BB})P(M|\overline{BB})} \tag{25}$$

- If the ratio is **greater than 1**, the student is classified as a **basketball player (BB)**.
- Otherwise, the student is classified as **not a basketball player $\overline{\mathbf{BB}}$**.

5.2 Observations from the figure

The **decision boundary and probability distributions** for this classification problem can be visualized as follows:

1. **Gaussian Distributions for Marks and Height**:
   - Two **Gaussian distributions** exist for **Marks**, representing probabilities conditioned on whether a student is a basketball player or not.
   - Similarly, two **Gaussian distributions** exist for **Height**, representing conditional probabilities given the class.
2. **Formation of Concentric Regions**:
   - The combination of **Height and Marks** distributions results in a **region of concentric circles** in the graph.
   - These circles represent contours of equal probability for classifying a student as **BB or $\overline{\mathbf{BB}}$**.
3. **Decision Boundary**:
   - The classification **boundary** is depicted in **violet**.
   - This boundary separates **basketball players** from **non-basketball players** based on **height and marks**.
4. **Elliptical Decision Regions**:
   - Depending on the **mean and variance** of Height and Marks distributions, the concentric circles can **deform into ellipses**.
   - This is because Gaussian distributions in **two dimensions** often lead to **elliptical decision boundaries** rather than perfect circles.

6 Key Takeaways

- **Computational Efficiency**: Naive Bayes significantly reduces computational complexity from exponential to linear scale.
- **Log-Linear Decision Boundaries**: The classifier maps feature distributions into log-space, forming a linear decision boundary in this transformed space.
- **Gaussian Naive Bayes for Continuous Data**: By assuming Gaussian distributions for continuous features, the model generalizes well for real-world applications.
- **Practical Applications**: Naive Bayes is widely used in spam filtering, medical diagnostics, and text classification due to its ability to handle **uncertainty and sparse data effectively**.