Week 4: Naive Bayes Classification

Ekarat Rattagan

September 12, 2025

Overview

Review: Given a data set,

$X \in \mathbb{R}^{N \times d}$ that has N rows and d dimensions. $y_i \in \{C_1, C_2, \ldots, C_k\}$

where C_k is a class k.

- Churn prediction
- Credit score prediction
- Email classification (link)

Bayes' Rule

$$\underbrace{P(Y \mid X)}_{\text{Posterior}} = \frac{\underbrace{P(X \mid Y)}_{\text{Likelihood}} \cdot \underbrace{P(Y)}_{\text{Prior}}}{\underbrace{P(X)}_{\text{Marginal}}}$$

There are four parts:

- **Posterior probability** (the conditional probability of Y given X)
- **Prior probability** (the initial belief regarding the truth of a statement)
- **Likelihood** (the chance of observing a particular sample X when the parameter is equal to Y)
- **Marginal probability** (the unconditional probability (overall populations))

Bayes' Rule (Conditional Probabilities)

$$P(Y|X) = \frac{P(Y \cap X)}{P(X)} \tag{1}$$

This says that conditional probability is the probability that both X and Y occur divided by the unconditional probability that X occurs.

$$P(X|Y) = \frac{P(X \cap Y)}{P(Y)} \tag{2}$$

$$P(X|Y)P(Y) = P(X \cap Y) \tag{3}$$

$$P(Y \cap X) = P(X \cap Y) \tag{4}$$

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{5}$$

Bayes Classifier:

$$P(Y|X) = \frac{P(X|Y)P(Y)}{P(X)} \tag{6}$$

Bayes' Rule Classification Example

X_1 (Name)	Y (Sex)
Drew	M
Claudia	F
Drew	F
Drew	F
Alberto	M
Karin	F
Nina	F
Sergio	M
Drew	?

Goal: Predict gender of the last "Drew" using Bayes' Rule.

$$P(Y_i = M \mid X_i = \text{Drew}) = \frac{P(\text{Drew} \mid M) \cdot P(M)}{P(\text{Drew})} = \frac{\frac{1}{3} \cdot \frac{3}{8}}{\frac{3}{8}} = \frac{1}{3}$$

$$P(Y_i = F \mid X_i = \text{Drew}) = \frac{P(\text{Drew} \mid F) \cdot P(F)}{P(\text{Drew})} = \frac{\frac{2}{5} \cdot \frac{5}{8}}{\frac{3}{8}} = \frac{2}{3}$$

Prediction: Female (F) with probability $\frac{2}{3}$

Naive Bayes with Multiple Features

X_1 (Name)	X_2 (Over 170cm)	X_3 (Eye Color)	X_4 (Hair Length)	Y (Sex)
Drew	No	Blue	Short	M
Claudia	Yes	Brown	Long	F
Drew	No	Blue	Long	F
Drew	No	Blue	Long	F
Alberto	Yes	Brown	Short	M
Karin	No	Blue	Long	F
Nina	Yes	Brown	Short	F
Sergio	Yes	Blue	Long	M

Example: Predict Y for X_1=Drew, X_2=No, X_3=Brown, X_4=Short

$$P(Y \mid x_1, x_2, x_3, x_4) = \frac{P(x_1, x_2, x_3, x_4 \mid Y) \cdot P(Y)}{P(x_1, x_2, x_3, x_4)}$$

Concept of Bayes' Rule

Features (x_i) are dependent (Assumption)

For a joint distribution of d variables,
all $2^d - 1$ combinations must be known.

General form of likelihood (Chain rule of probability):

$$P(x_1, x_2, \ldots, x_d \mid Y) = P(x_1 \mid Y)P(x_2 \mid x_1, Y)P(x_3 \mid x_2, x_1, Y)\ldots$$

$$\ldots P(x_d \mid x_{d-1}, x_{d-2}, \ldots, x_2, x_1, Y)$$

Naive Bayes Assumption

Assumption: Attributes x_i are conditionally independent given class y.

Likelihood:

$$P(\mathbf{X} \mid y_i) = P(x_1 \mid y_i) \cdot P(x_2 \mid y_i) \cdot \ldots \cdot P(x_d \mid y_i)$$

Example:

$$P(x_1, x_2, x_3, x_4 \mid Y) = P(\text{name} = \text{Drew} \mid Y = M) \cdot P(\text{over 170} = \text{No} \mid Y = M$$

$$\cdot P(\text{eye color} = \text{brown} \mid Y = M) \cdot P(\text{hair} = \text{short} \mid Y = M)$$

Evidence:

$$\text{evidence} = \sum_{i=1}^{K} P(\mathbf{X} \mid y_i) \cdot P(y_i) \quad \text{where } K \text{ is the number of classes}$$

Example computation

$$P(x_1 = \text{Drew},\ x_2 = \text{No},\ x_3 = \text{brown},\ x_4 = \text{short} \mid \text{Male})$$

$$P(x_1 = \text{Drew} \mid \text{Male})$$

$$P(x_2 = \text{No} \mid \text{Male})$$

$$P(x_3 = \text{brown} \mid \text{Male})$$

$$P(x_4 = \text{short} \mid \text{Male})$$

Advantages and Disadvantages

Advantages:

- Fast to train/classify
- Handles streaming data (e.g., Email spam detection)

Disadvantage:

- Assumes independence of features (X)
- Link: Does assumption work or not

Handling Continuous Values

Previous Example: $x_i \in \{\text{Categories}\}$ — Compute probability by counting.

If $x_i \in \{\text{Continuous-values}\}$, we will assume the values follow a Gaussian distribution.

Example: $x_i = \text{Height (cm)}$

$$\text{Values:} \quad 100, \; 101, \; 99, \; 120, \; \ldots, \; 160$$

Gaussian Probability Density Function:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

where $e = 2.7183$

Laplacian Correction (Estimator)

Dealing with zero probability values:
Example: Training set contains 1000 samples

- Income = low: 0 samples, medium: 990 samples, high: 10 samples

Without Laplace:

$$P(\text{low}) = 0$$
$$P(\text{medium}) = \frac{990}{1000}$$
$$P(\text{high}) = \frac{10}{1000}$$

With Laplace Correction:

$$P(\text{low}) = \frac{0+1}{1000+3}$$
$$P(\text{medium}) = \frac{990+1}{1000+3}$$
$$P(\text{high}) = \frac{10+1}{1000+3}$$

Reference

- Olabenjo, Babatunde. "Applying naive bayes classification to google play apps categorization." arXiv preprint arXiv:1608.08574 (2016).