Basic Probability and Its Applications in AI

Bui Van Tai

In this section, we'll examine several examples of probability applied to different types of data.

1 Classic Probability and Geometric Probability

1. Classic (Theoretical) Probability

Definition:

Classic probability (also known as theoretical probability) is the probability of an event occurring when all possible outcomes are equally likely.

Formula:

$$P(E) = \frac{\text{Number of favorable outcomes}}{\text{Total number of possible outcomes}}$$

Example:
If a fair 6-sided die is rolled, the probability of getting a 4 is:

$$P(4) = \frac{1}{6}$$

2. Geometric Probability

Definition:
Geometric probability is the probability that involves geometric measures such as length, area, or volume. It is used when outcomes are continuous and distributed over a region.

Formula:

$$P(E) = \frac{\text{Measure of favorable region}}{\text{Measure of total region}}$$

Example 1 (Length):
A point is randomly chosen on a line segment of length 10. What is the probability it falls on a sub-segment of length 4?

$$P = \frac{4}{10} = 0.4$$

Example 2 (Area):
A dart lands randomly on a circular target of radius 2. What is the probability it lands in a smaller concentric circle of radius 1?

$$P = \frac{\pi \cdot 1^2}{\pi \cdot 2^2} = \frac{1}{4}$$

3. Comparison Table

Feature	Classic Probability	Geometric Probability
Outcomes Type	Discrete, countable	Continuous, measurable
Based On	Counting outcomes	Measuring length/area/volume
Formula	$\frac{\text{favorable outcomes}}{\text{total outcomes}}$	$\frac{\text{favorable region}}{\text{total region}}$
Example	Rolling a die, flipping a coin	Random point on a line, dart on a target

2 Experimental Probability?

Experimental Probability is the probability based on the outcomes of an actual experiment or observation, rather than theoretical reasoning.

Definition:

> The experimental probability of an event is the ratio of the number of times the event occurs to the total number of trials.

$$P(E) = \frac{\text{Number of times event } E \text{ occurs}}{\text{Total number of trials}}$$

Example:
Suppose you flip a coin 100 times:

- Heads comes up 56 times,
- Tails comes up 44 times.

Then the experimental probability of getting heads is:

$$P(\text{heads}) = \frac{56}{100} = 0.56$$

Use Cases of Experimental Probability:

- When theoretical probability is too complex or unknown,
- In simulations, surveys, or real-world data collection,
- To test or compare with theoretical probability.

Comparison with Classic Probability:

Aspect	Classic Probability	Experimental Probability
Basis	Logical reasoning	Actual data or observation
Formula	$P(E) = \frac{\text{Favorable outcomes}}{\text{Total possible outcomes}}$	$P(E) = \frac{\text{Observed frequency}}{\text{Total trials}}$
Example	Coin flip: $P(\text{heads}) = \frac{1}{2}$	56 heads in 100 flips: $P(\text{heads}) = \frac{56}{100}$

3 Law of Total Probability

3.1 Definition

The Law of Total Probability allows us to calculate the probability of an event by dividing the sample space into non-overlapping partitions and computing the weighted sum of conditional probabilities.

If $\{B_1, B_2, \ldots, B_n\}$ is a partition of the sample space (meaning the B_i are mutually exclusive and their union is the entire sample space), then the probability of event A is:

$$P(A) = \sum_{i=1}^{n} P(A|B_i) \times P(B_i)$$

3.2 Significance

- Allows calculation of an event's probability through different "scenarios" (the partitions B_i)
- Each scenario contributes a portion to the total probability, proportional to the probability of that scenario
- Particularly useful when conditional probabilities $P(A|B_i)$ are easier to calculate than the direct probability $P(A)$

4 Rules of Probability

Probability theory is governed by a few fundamental rules that apply to all kinds of events and probability models.

1. Probability Values are Between 0 and 1

For any event A:

$$0 \leq P(A) \leq 1$$

- $P(A) = 0$: the event cannot occur.
- $P(A) = 1$: the event is certain to occur.

2. Probability of the Sample Space is 1

Let S be the sample space (the set of all possible outcomes):

$$P(S) = 1$$

3. Complement Rule

The complement of event A, denoted A^c, is the event that A does not occur:

$$P(A^c) = 1 - P(A)$$

4. Addition Rule

a) For Mutually Exclusive Events:
If A and B are mutually exclusive (they cannot both occur):

$$P(A \cup B) = P(A) + P(B)$$

b) General Addition Rule (Not Mutually Exclusive):

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

5. Multiplication Rule

a) For Independent Events:
If A and B are independent:

$$P(A \cap B) = P(A) \cdot P(B)$$

b) For Dependent Events:

$$P(A \cap B) = P(A) \cdot P(B \mid A)$$

Summary Table

Rule	Formula	Description
Boundaries	$0 \leq P(A) \leq 1$	Probabilities are between 0 and 1
Sample Space	$P(S) = 1$	The total probability is 1
Complement	$P(A^c) = 1 - P(A)$	The probability that event A does not occur
Addition (Disjoint)	$P(A \cup B) = P(A) + P(B)$	If A and B cannot both occur
Addition (General)	$P(A \cup B) = P(A) + P(B) - P(A \cap B)$	General case for any two events
Multiplication (Independent)	$P(A \cap B) = P(A) \cdot P(B)$	If events A and B are independent
Multiplication (Dependent)	$P(A \cap B) = P(A) \cdot P(B \mid A)$	If event B depends on A

5 Law of Total Probability

The **Law of Total Probability** allows us to compute the probability of an event by considering all the distinct and mutually exclusive ways the event can happen.

Formal Definition:
Let $B_1, B_2, \ldots, B_n$ be a *partition* of the sample space (i.e., the B_i's are mutually exclusive and exhaustive), and let A be any event. Then the probability of A is:

$$P(A) = \sum_{i=1}^{n} P(A \mid B_i) \cdot P(B_i)$$

Example:
Suppose a factory has 3 machines: M_1, M_2, M_3 producing 30%, 50%, and 20% of the items, respectively. Their defect rates are:

$$P(\text{defect} \mid M_1) = 0.01, \quad P(\text{defect} \mid M_2) = 0.02, \quad P(\text{defect} \mid M_3) = 0.05$$

Then the total probability of selecting a defective item is:

$$\begin{aligned} P(\text{defect}) &= P(\text{defect} \mid M_1) \cdot P(M_1) + P(\text{defect} \mid M_2) \cdot P(M_2) + P(\text{defect} \mid M_3) \cdot P(M_3) \\ &= 0.01 \cdot 0.3 + 0.02 \cdot 0.5 + 0.05 \cdot 0.2 \\ &= 0.003 + 0.01 + 0.01 \\ &= 0.023 \end{aligned}$$

So, the total probability of picking a defective item is 2.3%.

6 Bayes' Theorem

Bayes' Theorem is a fundamental rule in probability theory that allows us to update the probability of an event based on new evidence.

Definition

Let A and B be two events with $P(B) \neq 0$. Then:

$$P(A \mid B) = \frac{P(B \mid A) \cdot P(A)}{P(B)}$$

- $P(A \mid B)$: Posterior probability — probability of A given B
- $P(B \mid A)$: Likelihood — probability of B given A
- $P(A)$: Prior probability — initial belief about A
- $P(B)$: Evidence — total probability of observing B

3.1 Naive Bayes Classifier for Discrete Random Variables

Problem Example: Predicting Whether to Play Tennis

We are given a small dataset to predict whether a person will play tennis based on four **discrete features** (a "naive" hypothesis, yet effective): *Outlook*, *Temperature*, *Humidity* and *Wind*.

Dataset (5 Samples)

Outlook	Temp	Humidity	Wind	Play
Sunny	Hot	High	Weak	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	High	Strong	No
Sunny	Cool	High	Strong	???

Step 1: Bayes's Theorem Overview

We want to compute:

$$P(\text{Play=Yes} \mid X), \quad P(\text{Play=No} \mid X)$$

where

$$X = (\text{Outlook} = \text{Sunny}, \text{Temp} = \text{Cool}, \text{Humidity} = \text{High}, \text{Wind} = \text{Strong})$$

According to Bayes's Theorem:

$$P(\text{Class} \mid X) = \frac{P(X \mid \text{Class}) \cdot P(\text{Class})}{P(X)}$$

Since $P(X)$ is constant for both classes, we compare:

$$P(X \mid \text{Yes}) \cdot P(\text{Yes}) \quad \text{and} \quad P(X \mid \text{No}) \cdot P(\text{No})$$

Step 1: Compute Prior Probabilites

From the dataset:

$$P(\text{Yes}) = \frac{3}{5} = 0.6 \quad , \quad P(\text{No}) = \frac{2}{5} = 0.4$$

Step 2: Likelihoods (with Laplace Smoothing)

Apply Laplace smoothing (add-one) to avoid zero probabilities

$$P(x_i \mid \text{Class}) = \frac{\text{count}(x_i, \text{Class}) + 1}{\text{count}(\text{Class}) + N_i}$$

Where N_i is the number of possible values for feature i.
- Outlook: 3 values (Sunny, Overcast, Rain)
- Temperature: 3 values (Hot, Mild, Cool)
- Humidity: 2 values (High, Normal)
- Wind: 2 values (Weak, Strong)

Step 3: Compute Likelihoods

For Class = Yes (3 samples)

$$P(X \mid \text{Yes}) = P(\text{Sunny} \mid \text{Yes}) \cdot P(\text{Cool} \mid \text{Yes}) \cdot P(\text{High} \mid \text{Yes}) \cdot P(\text{Strong} \mid \text{Yes})$$

$$= \frac{1+1}{3+3} \cdot \frac{1+1}{3+3} \cdot \frac{2+1}{3+2} \cdot \frac{0+1}{3+2} = \frac{2}{6} \cdot \frac{2}{6} \cdot \frac{3}{5} \cdot \frac{1}{5} = \frac{12}{900}$$

$$P(X \mid \text{Yes}) \cdot P(\text{Yes}) = \frac{12}{900} \cdot 0.6 = \frac{7.2}{900} \approx 0.008$$

For Class = No (2 samples)

$$P(X \mid \text{No}) = P(\text{Sunny} \mid \text{No}) \cdot P(\text{Cool} \mid \text{No}) \cdot P(\text{High} \mid \text{No}) \cdot P(\text{Strong} \mid \text{No})$$

$$= \frac{1+1}{2+3} \cdot \frac{0+1}{2+3} \cdot \frac{2+1}{2+2} \cdot \frac{1+1}{2+2} = \frac{2}{5} \cdot \frac{1}{5} \cdot \frac{3}{4} \cdot \frac{2}{4} = \frac{12}{400}$$

$$P(X \mid \text{No}) \cdot P(\text{No}) = \frac{12}{400} \cdot 0.4 = \frac{4.8}{400} = 0.012$$

Conclusion

Since:

$$P(\text{No} \mid X) > P(\text{Yes} \mid X)$$

We conclude that the person will **not play tennis**.

Python Code

```python
# Training data (5 samples)
data = [
    ['Sunny', 'Hot', 'High', 'Weak', 'No'],
    ['Overcast', 'Hot', 'High', 'Weak', 'Yes'],
    ['Rain', 'Mild', 'High', 'Weak', 'Yes'],
    ['Sunny', 'Cool', 'Normal', 'Weak', 'Yes'],
    ['Rain', 'Mild', 'High', 'Strong', 'No']
]

# Input to predict
X = ['Sunny', 'Cool', 'High', 'Strong']

# Extract labels and features
labels = [row[-1] for row in data]
features = [row[:-1] for row in data]

# Calculate prior probability P(Class)
def prior_prob(class_label):
    return sum(1 for label in labels if label == class_label) / len(labels)

# Calculate conditional probability P(x_i | Class) with Laplace smoothing
def cond_prob(feature_idx, feature_val, class_label):
    count = 0
    total = 0
    unique_vals = set(row[feature_idx] for row in features)

    for i, row in enumerate(features):
        if labels[i] == class_label:
            total += 1
            if row[feature_idx] == feature_val:
                count += 1

    # Apply Laplace smoothing
    return (count + 1) / (total + len(unique_vals))

# Compute P(X | Yes) * P(Yes)
yes_prob = prior_prob('Yes')
for i in range(len(X)):
    yes_prob *= cond_prob(i, X[i], 'Yes')

# Compute P(X | No) * P(No)
no_prob = prior_prob('No')
for i in range(len(X)):
    no_prob *= cond_prob(i, X[i], 'No')

# Final prediction
print(f"P(X | Yes) * P(Yes) = {yes_prob}")
print(f"P(X | No) * P(No) = {no_prob}")
if yes_prob > no_prob:
    print("=> Prediction: Play")
```

```
else:
    print("=> Prediction: Do not play")
```

Listing 1: Naive Bayes Classifier

Output

```
P(X | Yes) * P(Yes) = 0.008
P(X | No) * P(No) = 0.012
=> Prediction: Do not play
```

3.2 Naive Bayes for Continuous Random Variable

Unlike discrete variables, which can take only a finite or countable set of distinct values, *continuous variables* can take infinite number of possible values (height, weight, temperature,...).

This example shows how to apply Navie Bayes Classifier to continuous data using Gaussian distribution.

We are given 5 BMI values and their corresponding class labels:

BMI (x)	Class (y)
18	0 (Healthy)
20	0
22	0
26	1 (Sick)
28	1
24	???

We want to classify a new sample with BMI=24.

Step 1: Compute mean and variance for each class

For class C_0 (Healthy):

General formulas:

$$\mu_0 = \frac{1}{n_0}\sum_{i=1}^{n_0} x_i \quad and \quad \sigma_0^2 = \frac{1}{n_0}\sum_{i=1}^{n_0}(x_i - \mu_0)^2$$

Apply to data: [18, 20, 22]

$$\mu_0 = \frac{18+20+22}{3} = 20$$

$$\sigma_0^2 = \frac{(18-20)^2 + (20-20)^2 + (22-20)^2}{3} = \frac{8}{3} \approx 2.67$$

$$\sigma_0 = \sqrt{2.67} \approx 1.63$$

For Class C_1 (Sick):

Apply to data: [26, 28]

$$\mu_1 = \frac{26+28}{2} = 27 \quad , \quad \sigma_1^2 = \frac{(26-27)^2 + (28-27)^2}{2} = 1 \quad , \quad \sigma_1 = \sqrt{1} = 1$$

Step 2: Compute Pior Probabilities

$$P(C_0) = \frac{n_0}{n} = \frac{3}{5} = 0.6$$

$$P(C_1) = \frac{n_1}{n} = \frac{2}{5} = 0.4$$

Step 3: Use Gaussian Probability Density Function

General formula:

$$P(x|C_k) = \frac{1}{\sqrt{2\pi\sigma_k^2}} \cdot e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

For class 0:

$$P(x = 24|C_0) = \frac{1}{\sqrt{2\pi \cdot 2.67}} \cdot e^{-\frac{(24-20)^2}{2 \cdot 2.67}} \approx 0.0122$$

$$P(C_0|x = 24) \propto 0.0122 \cdot 0.6 = 0.0073$$

For class 1:

$$P(x = 24|C_1) = \frac{1}{\sqrt{2\pi \cdot 1}} \cdot e^{-\frac{(24-27)^2}{2 \cdot 1}} \approx 0.0044$$

$$P(C_1|x = 24) \propto 0.0044 \cdot 0.4 = 0.0018$$

Step 4: Final Prediction

$$P(C_0|x = 24) \propto 0.0073 \quad P(C_1|x = 24) \propto 0.0018 \Rightarrow \text{Predict Class 0 (Healthy)}$$

Python Code

We use scikit-learn library, which provides efficient and easy-to-use tools for implementing Naive Bayes Classifier models, making the process simpler.

```
1
2  from sklearn.naive_bayes import GaussianNB
3  import numpy as np
4
5  # Input data
6  X = np.array([[18], [20], [22], [26], [28]])
7  y = np.array([0, 0, 0, 1, 1])
8
9  # Train model
10 model = GaussianNB()
11 model.fit(X, y)
12
13 # Predict for BMI = 24
14 x_test = np.array([[24]])
15 predicted_class = model.predict(x_test)
16 print("Predicted class:", predicted_class[0])
```

Listing 2: NBC For Continuous Random Variable

```
1 Predicted class: 0
```

Bayes'theorem demonstrates its wide-ranging and adaptable applicability in various problem domains and data types.