Probability for Machine Learning

Basic Class Notes

1 Why Do We Need Probability in Machine Learning?

Machine learning is often about making predictions when we do not know everything.

For example:

- Will a customer buy a product?
- Is an email spam?
- Is a picture a cat or a dog?
- Will a stock price go up or down?
- Is a credit card transaction fraudulent?

The important idea is that we usually do not know the answer with certainty.

Probability gives us a mathematical language for describing **uncertainty**.

Instead of saying:

> "This email is definitely spam."

a machine learning model might say:

$$P(\text{spam}) = 0.90$$

which means that the model estimates a 90% probability that the email is spam.

Probability is therefore useful for:

- representing uncertainty,
- making predictions,
- classifying observations,
- dealing with noisy data,
- combining evidence, and
- making decisions.

2 Basic Probability

A probability is a number between 0 and 1.

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

where A is an **event**.

We can also express probability as a percentage.

$$0.5 = 50\%$$

2.1 Example: Rolling a Die

Suppose we roll a fair six-sided die.

The possible results are:

$$\{1, 2, 3, 4, 5, 6\}$$

The probability of rolling a 3 is:

$$P(3) = \frac{1}{6}$$

The probability of rolling an even number is:

$$P(\text{even}) = P(2) + P(4) + P(6) = \frac{3}{6} = \frac{1}{2}$$

So:

$$P(\text{even}) = 0.5 = 50\%$$

3 Events

An **event** is something that we are interested in.

For example:

$$A = \text{rolling an even number}$$

and

$$B = \text{rolling a number greater than 3.}$$

There are two useful operations on events.

3.1 Intersection: AND

The intersection

$$A \cap B$$

means "A AND B."

For example:

$$A = \text{even}$$

$$B = \text{greater than 3}$$

Then:

$$A \cap B = \{4, 6\}$$

because 4 and 6 are both even and greater than 3.

3.2 Union: OR

The union

$$A \cup B$$

means "A OR B."

The word "OR" normally includes the possibility that both events happen.

4 Independent Events

Two events are **independent** if knowing that one happened does not change the probability of the other.

For independent events:

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

4.1 Example: Two Coin Flips

Suppose we flip a fair coin twice.

The probability of heads on the first flip is:

$$P(H_1) = \frac{1}{2}$$

The probability of heads on the second flip is also:

$$P(H_2) = \frac{1}{2}$$

The probability of getting heads twice is:

$$P(H_1 \cap H_2) = P(H_1)P(H_2) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$$

So there is a 25% chance of getting two heads.

5 Conditional Probability

One of the most important ideas in machine learning is **conditional probability**.

Conditional probability asks:

> What is the probability of A, given that we already know B?

We write this as:

$$P(A \mid B)$$

and read it as:

> "The probability of A given B."

The formula is:

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

provided that $P(B) > 0$.

5.1 Example: Students and Exams

Suppose there are 100 students.

- 60 students studied for an exam.
- 50 students passed.
- 40 students both studied and passed.

Let:

$$S = \text{student studied}$$

and

$$P = \text{student passed.}$$

We want to know:

$$P(P \mid S)$$

In words:

> What is the probability that a student passed, given that the student studied?

There are 60 students who studied.

Of those 60 students, 40 passed.

Therefore:

$$P(P \mid S) = \frac{40}{60} = \frac{2}{3} \approx 0.667$$

So approximately 67% of students who studied passed.

5.2 Why This Matters for Machine Learning

Machine learning frequently asks questions like:

$$P(\text{buy} \mid \text{visited website})$$

or:

$$P(\text{fraud} \mid \text{transaction information})$$

or:

$$P(\text{cat} \mid \text{image})$$

The information after the vertical bar is what we already know.

6 Bayes' Rule

Bayes' rule is one of the most important formulas in probability and machine learning.

It allows us to reverse a conditional probability.

The formula is:

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

The four parts have useful names.

Quantity	**Common name**
$P(A)$	Prior probability
$P(B \mid A)$	Likelihood
$P(B)$	Evidence
$P(A \mid B)$	Posterior probability

The basic idea is:

$$\boxed{\text{Posterior} \propto \text{Likelihood} \times \text{Prior}}$$

where "$\propto$" means "proportional to."

7 Bayes' Rule: Spam Example

Suppose we have 1,000 emails.

- 200 are spam.
- 800 are not spam.

Therefore:

$$P(\text{spam}) = \frac{200}{1000} = 0.20$$

and:

$$P(\text{not spam}) = 0.80.$$

Now suppose the word "FREE" appears in:

- 80% of spam emails.
- 10% of non-spam emails.

Therefore:

$$P(\text{FREE} \mid \text{spam}) = 0.80$$

and:

$$P(\text{FREE} \mid \text{not spam}) = 0.10.$$

We receive an email containing "FREE."

We want:

$$P(\text{spam} \mid \text{FREE})$$

Using Bayes' rule:

$$P(\text{spam} \mid \text{FREE}) = \frac{P(\text{FREE} \mid \text{spam})P(\text{spam})}{P(\text{FREE})}$$

We first calculate how many emails contain "FREE."

Among spam:

$$200(0.80) = 160$$

Among non-spam:

$$800(0.10) = 80$$

Therefore 240 emails contain "FREE."

Thus:

$$P(\text{spam} \mid \text{FREE}) = \frac{160}{240} = \frac{2}{3} \approx 0.667$$

So an email containing "FREE" has approximately a 67% probability of being spam under this simple model.

8 A Very Important Distinction

Do not confuse:

$$P(A \mid B)$$

with:

$$P(B \mid A).$$

They are generally different.

For example:

$$P(\text{spam} \mid \text{FREE})$$

asks:

> Given that an email contains "FREE," how likely is it to be spam?

But:

$$P(\text{FREE} \mid \text{spam})$$

asks:

> Given that an email is spam, how likely is it to contain "FREE"?

These are different questions.

Bayes' rule provides the mathematical connection between them.

9 Maximum Likelihood Estimation

Probability can also be used to estimate unknown parameters.

Suppose we have a coin, but we do not know whether it is fair.

Let:

$$p = P(\text{heads})$$

Then:

$$P(\text{tails}) = 1 - p.$$

Suppose we flip the coin 10 times and observe:

$$H, H, T, H, H, T, H, H, H, T$$

There are:

$$7 \text{ heads}$$

and:

$$3 \text{ tails}.$$

We want to estimate p.

9.1 Likelihood

If the probability of heads is p, the probability of observing 7 heads and 3 tails is proportional to:

$$L(p) = p^7(1 - p)^3.$$

This is called the **likelihood function**.

Maximum likelihood estimation chooses the value of p that makes the observed data most likely.

In this example, the maximum occurs at:

$$\boxed{p = 0.7}$$

So our maximum likelihood estimate is:

$$P(\text{heads}) \approx 0.7.$$

This makes intuitive sense: 7 out of 10 observed flips were heads.

10 Probability vs. Likelihood

The words "probability" and "likelihood" are related, but they are used differently.

Suppose:

$$P(\text{heads}) = p.$$

If p is known, probability tells us how likely an observation is.

For example:

$$P(7 \text{ heads in } 10 \text{ flips} \mid p = 0.5).$$

With likelihood, we observe the data and ask which value of p best explains it.

$$L(p \mid \text{observed data})$$

So:

Probability	**Likelihood**
Parameter is treated as known	Data is treated as known
Ask about possible data	Ask about possible parameter values
Used for prediction	Used for estimation

Maximum likelihood estimation is therefore a way of learning parameters from data.

11 Naive Bayes Classifiers

A **Naive Bayes classifier** is a simple machine learning algorithm based on Bayes' rule.

It can be used for classification.

For example:

- spam vs. not spam,
- positive vs. negative review,

- cat vs. dog,
- disease vs. no disease,
- document category classification.

Suppose we want to classify an email.
Let:

$$C = \text{class}$$

where the possible classes are:

$$C \in \{\text{spam}, \text{not spam}\}.$$

We observe some features:

$$X = (x_1, x_2, \ldots, x_n).$$

For example:

$$X = (\text{FREE}, \text{MONEY}, \text{WIN})$$

We want to calculate:

$$P(\text{spam} \mid X)$$

and:

$$P(\text{not spam} \mid X).$$

We choose the class with the larger probability.

12 The "Naive" Assumption

The difficult part is calculating:

$$P(x_1, x_2, x_3 \mid C).$$

Naive Bayes makes a simplifying assumption.
It assumes that the features are conditionally independent once we know the class.

Therefore:

$$P(x_1, x_2, x_3 \mid C) \approx P(x_1 \mid C)P(x_2 \mid C)P(x_3 \mid C).$$

For many features:

$$\boxed{P(X \mid C) \approx \prod_i P(x_i \mid C)}$$

This assumption is often not completely true.

That is why the algorithm is called "naive."

However, the simplification makes the calculation very easy.

13 Naive Bayes Example

Suppose we have two classes:

spam

and:

not spam.

Suppose:

$$P(\text{spam}) = 0.4$$

and:

$$P(\text{not spam}) = 0.6.$$

An email contains the words:

FREE

and:

MONEY.

Suppose our training data tells us:

$$P(\text{FREE} \mid \text{spam}) = 0.8$$

$$P(\text{MONEY} \mid \text{spam}) = 0.7$$

and:

$$P(\text{FREE} \mid \text{not spam}) = 0.1$$

$$P(\text{MONEY} \mid \text{not spam}) = 0.2.$$

Using the naive independence assumption:

$$P(X \mid \text{spam}) \approx P(\text{FREE} \mid \text{spam})P(\text{MONEY} \mid \text{spam})$$

Therefore:

$$P(X \mid \text{spam}) \approx 0.8(0.7) = 0.56.$$

Multiply by the prior:

$$P(X \mid \text{spam})P(\text{spam}) = 0.56(0.4) = 0.224.$$

For not spam:

$$P(X \mid \text{not spam}) \approx 0.1(0.2) = 0.02.$$

Therefore:

$$P(X \mid \text{not spam})P(\text{not spam}) = 0.02(0.6) = 0.012.$$

We compare:

$$0.224$$

with:

$$0.012.$$

Since the spam score is much larger, the Naive Bayes classifier classifies the email as:

$$\boxed{\text{spam}}$$

14 Why Can We Ignore the Denominator?

Bayes' rule says:

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

Suppose we are comparing two possible classes.

For spam:

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

For not spam:

$$P(\text{not spam} \mid X) = \frac{P(X \mid \text{not spam})P(\text{not spam})}{P(X)}.$$

Notice that $P(X)$ is the same in both cases.

Therefore, when we only want to know which class has the larger probability, we can compare:

$$P(X \mid C)P(C)$$

without calculating $P(X)$.

This gives the classification rule:

$$\boxed{\hat{C} = \arg\max_{C} P(X \mid C)P(C)}$$

where $\hat{C}$ means our predicted class.

15 Training a Naive Bayes Classifier

How do we obtain all these probabilities?

We use training data.

Suppose we have many labeled emails:

Email	Contains FREE	Class
1	Yes	Spam
2	No	Not spam
3	Yes	Spam
4	No	Not spam
5	Yes	Spam

We can count examples.

For example:

$$P(\text{spam}) = \frac{\text{number of spam emails}}{\text{number of emails}}$$

and:

$$P(\text{FREE} \mid \text{spam}) = \frac{\text{spam emails containing FREE}}{\text{spam emails}}.$$

The probabilities are therefore estimated from the training data.

This connects Naive Bayes to maximum likelihood estimation: we are estimating probability parameters from observed data.

16 Probability, Bayes, and Machine Learning

The ideas in this chapter fit together.

Probability

Probability describes uncertainty:

$$P(A).$$

Conditional Probability

Conditional probability lets us use information we already know:

$$P(A \mid B).$$

Bayes' Rule

Bayes' rule lets us reverse conditional probabilities:

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

Maximum Likelihood

Maximum likelihood estimates unknown parameters from data.

Naive Bayes

Naive Bayes uses Bayes' rule plus a simplifying independence assumption to classify data. The overall idea is:

$$\boxed{\text{Data} \longrightarrow \text{Probability estimates} \longrightarrow \text{Bayesian reasoning} \longrightarrow \text{Prediction}}$$

17 Why Is Naive Bayes Useful?

Naive Bayes is relatively simple, but it illustrates several important machine learning ideas.

1. **Learning from data**
   We estimate probabilities from examples.
2. **Using prior knowledge**
   The class prior $P(C)$ affects the prediction.
3. **Combining evidence**
   Multiple features can contribute to a prediction.
4. **Handling uncertainty**
   The result can be expressed as probabilities rather than absolute certainty.
5. **Classification**
   We can select the class with the highest posterior probability.

18 A Simple Mental Model

A useful way to think about Bayesian classification is:

> Start with what you believed before seeing the new evidence. Then look at the evidence and update your belief.

For example:

Prior belief

becomes:

Updated belief after seeing evidence.

Mathematically:

$$\boxed{\text{Posterior} \propto \text{Prior} \times \text{Evidence}}$$

This "update your belief using new evidence" idea appears throughout machine learning.

19 Key Ideas to Remember

1. A probability is a number between 0 and 1.
2. Conditional probability asks:
$$P(A \mid B)$$
"What is the probability of A, given B?"
3. Bayes' rule is:
$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)}.$$
4. Do not confuse:
$$P(A \mid B)$$
with:
$$P(B \mid A).$$
5. Maximum likelihood chooses the parameter value that makes the observed data most likely.
6. Naive Bayes uses:
$$P(C \mid X) \propto P(X \mid C)P(C).$$

7. Naive Bayes assumes features are conditionally independent:

$$P(X \mid C) \approx \prod_i P(x_i \mid C).$$

8. Probability provides a way for machine learning models to reason about uncertainty.

20 Quick Practice Questions

1. Basic Probability

A fair six-sided die is rolled.

What is the probability of rolling an odd number?

$$\text{Answer: } \frac{3}{6} = \frac{1}{2}.$$

2. Conditional Probability

In a class of 100 students:

- 40 are freshmen.
- 60 are sophomores.
- 30 freshmen passed an exam.

What is:

$$P(\text{passed} \mid \text{freshman})?$$

3. Bayes' Rule

Suppose:

$$P(A) = 0.2$$

$$P(B \mid A) = 0.8$$

and:

$$P(B) = 0.4.$$

Calculate:

$$P(A \mid B).$$

Using Bayes' rule:

$$P(A \mid B) = \frac{0.8(0.2)}{0.4} = 0.4.$$

4. Maximum Likelihood

A coin is flipped 20 times.

The result is:

$$12 \text{ heads}, \qquad 8 \text{ tails}.$$

What is the maximum likelihood estimate of the probability of heads?

$$\text{Answer: } \hat{p} = \frac{12}{20} = 0.6.$$

5. Naive Bayes

Suppose a classifier compares:

$$P(\text{cat} \mid X)$$

with:

$$P(\text{dog} \mid X).$$

If:

$$P(\text{cat} \mid X) = 0.75$$

and:

$$P(\text{dog} \mid X) = 0.25,$$

which class does the classifier choose?

Answer: cat.

21 Summary

Probability gives machine learning a way to represent uncertainty.

The most important progression to understand is:

$$\boxed{\begin{array}{c}\text{Probability}\\ \downarrow \\ \text{Conditional Probability}\\ \downarrow \\ \text{Bayes' Rule}\\ \downarrow \\ \text{Maximum Likelihood}\\ \downarrow \\ \text{Naive Bayes Classification}\end{array}}$$

These ideas are simple enough to understand with coins, dice, and email examples, but they form the foundation for much more advanced machine learning methods.