Foundations of Machine Learning: Week 1: Total Probability and Bayes' Theorem

Professor Christopher Yau
christopher.yau@manchester.ac.uk

October 23, 2020

Foundations of Machine Learning:
Week 1: Total Probability and Bayes' Theorem

Professor Christopher Yau
christopher.yau@manchester.ac.uk
October 23, 2020

Total (Marginal) Probability

For two disjoint events E_1 and E_2 and event F the **law of total probability** says:

$$P(F) = P(F|E_1)P(E_1) + P(F|E_2)P(E_2)$$

Example: What is the total probability of winning a tennis match?

$$P(\text{Win}) = P(\text{Win}|\text{Outplay})P(\text{Outplay}) + P(\text{Win}|\text{Injury})P(\text{Injury})$$

Total (Marginal) Probability

For two disjoint events E_1 and E_2 and event F the **law of total probability** says:

$$P(F) = P(F|E_1)P(E_1) + P(F|E_2)P(E_2)$$

Example: What is the total probability of winning a tennis match?

$$P(\text{Win}) = P(\text{Win}|\text{Outplay})P(\text{Outplay}) + P(\text{Win}|\text{Injury})P(\text{Injury})$$

In this lecture, we will consider the concept of *total probability* otherwise known as marginal probability.

This applies when we are interested in the probability of an event F which can arise after other events have happened E_1 and E_2 but, we do not care specifically about E_1 or E_2 only whether F has occurred.

If the preceding events are disjoint, i.e. they do not overlap, the total probability of F is given by the sum of the probability of F due to E_1 and F due to E_2 weighted by the probabilities of E_1 and E_2.

For example, you may win a tennis match because you outplay your opponent or your opponent retires injured.

The total probability of a victory needs to weight and combine these two preceding possibilities.

Total (Marginal) Probability

In general, for n disjoint events $E_1, \ldots, E_n$, and event F:

$$P(F) = \sum_{i=1}^{n} P(F|E_i)P(E_i).$$

If event F occurs as a consequence of either E_1, $E_2, \ldots,$ or E_n but we are only interested exclusively in F then we can take a weighted average over the possible preceding events.

Example: What is the probability that a machine learning classifier makes an error?

$$\begin{aligned} P(\text{Error}) = & P(\text{Error}|\text{Class A})P(\text{Class A}) \\ & + P(\text{Error}|\text{Class B})P(\text{Class B}) \\ & + P(\text{Error}|\text{Class C})P(\text{Class C}) \end{aligned}$$

Total (Marginal) Probability

In general, for n disjoint events $E_1, \ldots, E_n$, and event F:

$$P(F) = \sum_{i=1}^{n} P(F|E_i)P(E_i).$$

If event F occurs as a consequence of either E_1, $E_2, \ldots,$ or E_n but we are only interested exclusively in F then we can take a weighted average over the possible preceding events.

Example: What is the probability that a machine learning classifier makes an error?

$$\begin{aligned} P(\text{Error}) = & P(\text{Error}|\text{Class A})P(\text{Class A}) \\ & + P(\text{Error}|\text{Class B})P(\text{Class B}) \\ & + P(\text{Error}|\text{Class C})P(\text{Class C}) \end{aligned}$$

In general, if an events depends on multiple disjoint preceding events, we need to sum and weight over all the preceding possibilities.
For example, suppose we are interested in characterising the total error probability of a three-class classification algorithm where the probability of an error varies by the classification.
In order to obtain the total probability, we need to take a weighted-average over the classification possibilities where the weights are the probability of each of the three classes.
So the total error probability is the probability that the classifier calls class A and then an error occurs as a result of that, or the classifiers calls B and then an error occurs because of this and so on for class C.
The preceding events are disjoint since the classifier can only call one of the three classes.

Numerical Example

Suppose machines X, Y and Z are used to manufacture a device.

1. Machine X produces 50% of the devices of which 3% are defective
2. Machine Y produces 30% of the devices of which 4% are defective
3. Machine Z produces 20% of the devices of which 5% are defective

What is the probability that any given device is defective?

Numerical Example

Suppose machines X, Y and Z are used to manufacture a device.
1. Machine X produces 50% of the devices of which 3% are defective
2. Machine Y produces 30% of the devices of which 4% are defective
3. Machine Z produces 20% of the devices of which 5% are defective

What is the probability that any given device is defective?

Lets consider a numerical example.
Suppose machines X, Y and Z are used to manufacture a device.

1. Machine X produces 50% of the devices of which 3% are defective
2. Machine Y produces 30% of the devices of which 4% are defective
3. Machine Z produces 20% of the devices of which 5% are defective

This is a total probability question because we are interested in the probability of a device being defective irrespective of which machine made it.

Solution

Let D denote the event that a device is defective.

We want $P(D)$ and we are given:

- The probabilities that a device is made by X, Y and Z:

  $P(X) = 0.5$, $P(Y) = 0.3$ and $P(Z) = 0.2$.

- The conditional probability that a device is defective given that it is made by a specific machine:

  $P(D|X) = 0.03$, $P(D|Y) = 0.04$ and $P(D|Z) = 0.05$.

From the Law of Total Probability:

$$\begin{aligned} P(D) &= P(D|X)P(X) + P(D|Y)P(Y) + P(D|Z)(Z), \\ &= 0.03 \times 0.5 + 0.04 \times 0.3 + 0.05 \times 0.2, \\ &= 0.037. \end{aligned}$$

Solution
Let D denote the event that a device is defective.
We want $P(D)$ and we are given:
- The probabilities that a device is made by X, Y and Z: $P(X) = 0.5$, $P(Y) = 0.3$ and $P(Z) = 0.2$.
- The conditional probability that a device is defective given that it is made by a specific machine: $P(D|X) = 0.03$, $P(D|Y) = 0.04$ and $P(D|Z) = 0.05$.

From the Law of Total Probability:
$P(D) = P(D|X)P(X) + P(D|Y)P(Y) + P(D|Z)(Z)$, $= 0.03 \times 0.5 + 0.04 \times 0.3 + 0.05 \times 0.2$, $= 0.037$.

Denote by D the event that a device is defective.
We want to find the total probability $P(D)$.
We start by identifying the probabilities that each machine made the device and call these events$X$, Y and Z.
$P(X) = 0.5$, $P(Y) = 0.3$ and $P(Z) = 0.2$.
We can extract what the conditional probabilities of a defective device are under each machine from the question details.
$P(D|X) = 0.03$, $P(D|Y) = 0.04$ and $P(D|Z) = 0.05$.
Then use the law of total probability to work out the required probability.

$$P(D) = P(D|X)P(X) + P(D|Y)P(Y) + P(D|Z)(Z),$$

using these numerical quantities.

Bayes Theorem

Bayes Theorem for two events A and B states that:

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

Derivation:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A),$$
$$\Rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Bayes Theorem

Bayes Theorem for two events A and B states that:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Derivation:

$$P(A \cap B) = P(A|B)P(B) = P(B|A)P(A),$$
$$\Rightarrow P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Now that we have discussed conditional probability and total probability, we now turn to a very important probability theorem known as Bayes' Theorem.
The purpose of Bayes' Theorem is to allow us to *reverse* conditional probabilities.
Suppose I am interested in the probability of an event A given B.
Bayes' Theorem says that I compute this if I know the probability of B given A and the marginal probabilities of A and B.
You may read about $P(A)$ and $P(B)$ described as the *prior* probabilities of the events A and B.
This definition applies when Bayes' Theorem is used for Bayesian statistical modelling purposes. We will not be covering this topic in this course.

Bayes Theorem

In more general terms, for events $A_1, \ldots, A_n$ and event B:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

where in the denominator we use the total probability law in place of $P(B)$.

Bayes Theorem allows us to “flip” probabilities by allowing to answer questions about A given B by examining what happens to B given A.

Bayes Theorem

In more general terms, for events $A_1, \ldots, A_n$ and event B:

$$P(A_k|B) = \frac{P(B|A_k)P(A_k)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

where in the denominator we use the total probability law in place of $P(B)$.

Bayes Theorem allows us to “flip” probabilities by allowing to answer questions about A given B by examining what happens to B given A.

If we have multiple events A_1 to A_n given B, the general form of Bayes Theorem extends to considering all the cases.
The numerator depends on the event A_k we are interested in.
While the denominator uses a general form of total probability to give the total probability of event B averaging over all the possible preceding events.
Bayes Theorem allows us to “flip” probabilities by allowing to answer questions about A given B by examining what happens to B given A.

Numerical Example

Suppose machines X, Y and Z are used to manufacture a device.

1. Machine X produces 50% of the devices of which 3% are defective
2. Machine Y produces 30% of the devices of which 4% are defective
3. Machine Z produces 20% of the devices of which 5% are defective

What is the probability that given a defective device, it is made by machine X, Y or Z?

Numerical Example

Suppose machines X, Y and Z are used to manufacture a device.

1. Machine X produces 50% of the devices of which 3% are defective
2. Machine Y produces 30% of the devices of which 4% are defective
3. Machine Z produces 20% of the devices of which 5% are defective

What is the probability that given a defective device, it is made by machine X, Y or Z?

Lets return to our previous example and now ask what is the probability that, when given a defective device, it is made by a particular machine?
Note we need the probability of the machine *given* that a device is defective. However, the information we are given only tells us the probability of a defective device given the machine that made it.
This indicates we will need Bayes Theorem to "flip" the probabilities.

Solution

From the Law of Total Probability:

$$\begin{aligned} P(D) &= P(D|X)P(X) + P(D|Y)P(Y) + P(D|Z)P(Z), \\ &= 0.03 \times 0.5 + 0.04 \times 0.3 + 0.05 \times 0.2, \\ &= 0.037 \end{aligned}$$

From Bayes Theorem:

$$P(X|D) = P(D|X)P(X)/P(D) = 0.03 \times 0.5/0.037 = 0.405$$

and similarly for Y and Z.

You should find that $P(X|D) + P(Y|D) + P(Z|D) = 1$, i.e. given a defective device one of the machines must have made it! This is a useful sanity check for your calculations.

Solution

From the Law of Total Probability:

$$P(D) = P(D|X)P(X) + P(D|Y)P(Y) + P(D|Z)P(Z), \\ = 0.03 \times 0.5 + 0.04 \times 0.3 + 0.05 \times 0.2, \\ = 0.037$$

From Bayes Theorem:

$$P(X|D) = P(D|X)P(X)/P(D) = 0.03 \times 0.5/0.037 = 0.405$$

and similarly for Y and Z.

You should find that $P(X|D) + P(Y|D) + P(Z|D) = 1$, i.e. given a defective device one of the machines must have made it! This is a useful sanity check for your calculations.

Previously, we computed the total probability of a defective device. We will need this result to apply Bayes Theorem.
If we denote by D the event of a defective device and X the event that machine X made the device.
Using Bayes Theorem, the probability of machine X given D is given by the probability that X made a defective device times the probability that X makes the device.
These probabilities are given to us as part of the question.
The denominator is the total probability $P(D)$ we just calculated.
If we repeat the calculation for Y and Z, we should see that they add up to one since the entire sample space has been enumerated and therefore the sum of all possibilities must add to one.

Recap: Probability

Total probability refers to the probability of an event when all its dependencies have been averaged out.

Bayes' Theorem allows us to "flip" probabilities to answer conditional probability questions.

Recap: Probability

Total probability refers to the probability of an event when all its dependencies have been averaged out.

Bayes' Theorem allows us to "flip" probabilities to answer conditional probability questions.

So to summarise:
Total or marginal probability refers to the probability of an event when all its dependencies have been averaged out.
Bayes' Theorem allows us to "flip" probabilities to answer conditional probability questions.

END LECTURE

END LECTURE