Lecture 13: BAYES' RULE [*]

10-606 MATHEMATICAL FOUNDATIONS FOR MACHINE LEARNING

1 Factoring with conditionals

Earlier we saw that we can write a conditional distribution as the ratio of a joint and a marginal:

$$P(X \mid Y) = \frac{P(X, Y)}{P(Y)}$$

A useful rearrangement of the above is

$$P(X, Y) = P(X \mid Y)P(Y)$$

That is, we can write a joint distribution as a product of a marginal and a conditional. By symmetry, it works the other way too: $P(X, Y) = P(Y \mid X)P(X)$.

Since a conditional distribution itself is just like any other distribution, the same identity works inside a conditional:

$$P(X, Y \mid Z) = P(X \mid Y, Z)P(Y \mid Z)$$

So, we can use this identity repeatedly to break up a large joint distribution into a product of factors, e.g.,

$$P(X, Y, Z) = P(X, Y \mid Z)P(Z) = P(X \mid Y, Z)P(Y \mid Z)P(Z)$$

2 Bayes' rule

From the identity above, we know

$$P(X \mid Y)P(Y) = P(X, Y) = P(Y \mid X)P(X)$$

Rearranging, we have

$$P(Y \mid X) = \frac{P(X \mid Y)P(Y)}{P(X)}$$

This equation is known as *Bayes' rule*.

Bayes' rule is really useful since it tells us how to incorporate new evidence or information about an uncertain variable.

[*] Compiled on Wednesday 19th July, 2023 at 14:26

For example, suppose there are two coins in a bag: one fair one, and one that's biased so that the probability of heads is 0.7. We draw one coin at random from the bag and start flipping it.

Initially we are equally likely to have drawn the fair coin or the biased one: $P(\text{fair}) = 0.5$. Now suppose we flip our coin and see heads. Intuitively, this outcome is more likely if we're holding the biased coin, so $P(\text{fair})$ should decrease.

Bayes' rule tells us that

$$P(\text{fair} \mid \text{flip} = H) = P(\text{flip} = H \mid \text{fair}) P(\text{fair}) / P(\text{flip} = H)$$

In words, we start from our *prior* probability table $P(\text{fair})$:

fair	0.5
¬fair	0.5

We multiply it by the *evidence* or likelihood of seeing a heads from our flip: $P(\text{flip} = H \mid \text{fair})$

$H \mid \text{fair}$	0.5
$H \mid \neg\text{fair}$	0.7

This gives us

$H \wedge \text{fair}$	$0.5 \cdot 0.5 = 0.25$
$H \wedge \neg\text{fair}$	$0.5 \cdot 0.7 = 0.35$

We then divide by the marginal probability of our observation $P(H)$. We can either calculate $P(H)$ directly, or use a shortcut: $P(H) = P(H \wedge \text{fair}) + P(H \wedge \neg\text{fair})$, so $P(H)$ is the sum of the entries in the table above. Either way we get $P(H) = 0.6$, so our final answer becomes:

$\text{fair} \mid H$	$0.25/0.6 = 5/12$
$\neg\text{fair} \mid H$	$0.35/0.6 = 7/12$

This is called the *posterior* probability of fair given the evidence of seeing H.

If we flip the coin again, we can repeat the exercise: our posterior after the first flip becomes our prior before the second flip, and we use Bayes' rule again to get our posterior after the second flip. This process, repeatedly updating a distribution over some variable using new evidence, is often called *tracking* or *filtering*.

Knowledge check

1. **Numerical answer**: Suppose we flip the coin again and see heads again. What is our new posterior $P(\text{fair} \mid H, H)$?
    - **Hint**: Start from the posterior distribution derived above, $P(\text{fair} \mid H)$.
    - **Answer**: Starting from the previous posterior distribution as our new prior, we can multiply by the evidence of seeing another heads, which gives:

$H \wedge$ (fair $\mid H$)	$5/12 \cdot 0.5 = 5/24$
$H \wedge$ ($\neg$fair $\mid H$)	$7/12 \cdot 0.7 = 49/120$

Normalizing once again gives us the (somewhat gnarly) new posterior distribution:

fair $\mid H, H$	$25/74$
$\neg$fair $\mid H, H$	$49/74$

3 Factoring a probability distribution

When we use probability spaces in practice, one of the most important tasks is to describe complicated relationships among many different possible events and random variables. A good way to organize these relationships is to *factor* our probability distribution i.e.,

$$P(X) = F_1(X)F_2(X)\dots F_n(X)$$

where X stands for a list of all the random variables or events that we might want to reason about, and each factor $F_k(X)$ encodes some understandable part of our overall model. This factorization means that the probability of X taking a given value x is

$$P(X = x) = F_1(x)F_2(x)\dots F_n(x)$$

There are lots of possible kinds of factors we might want to include. We can't cover them all, but the rest of this set of notes will look at a few useful kinds, and how to work with them.

3.1 Independence

The simplest kind of relationship is none at all: suppose we can split our random variables into two or more subsets that don't influence one another. Then these subsets are called *independent* of one another.

We can represent independence by writing our overall probability as a product of two or more factors, where the factors have *disjoint* sets of arguments, e.g.,

$$P(X_1, X_2) = P(X_1)P(X_2)$$

For example, if we flip a coin twice, it makes sense to assume that the two flips are independent: getting heads on one flip doesn't influence our chance of getting heads on the other. The above formula could represent our joint distribution, if we take X_1 to represent the first flip and X_2 to represent the second.

If we write x_1 for a value that X_1 might take, and x_2 for a value that X_2 might take, the above factorization stands for a table in which

$$P(X_1 = x_1, X_2 = x_2) = P(X_1 = x_1)P(X_2 = x_2)$$

For example, if our first coin has probability 0.7 of showing heads, while our second coin has probability 0.6, we get the following table:

X_1	X_2	p
H	H	$0.7 \cdot 0.6 = 0.42$
H	T	$0.7 \cdot 0.4 = 0.28$
T	H	$0.3 \cdot 0.6 = 0.18$
T	T	$0.3 \cdot 0.4 = 0.12$

We can see that this table encodes independence by calculating conditional distributions. Suppose we observe that $X_1 = T$. Then we can use the rule for conditional probabilities to find $P(X_2 \mid X_1 = T)$. To do so, we cross out the rows in the table that are inconsistent with our observation; in this case, we cross out the first and second rows, and keep the third and fourth. Then we renormalize the remaining rows: the two remaining entries are $0.18, 0.12$ and their sum is 0.3, so the result is

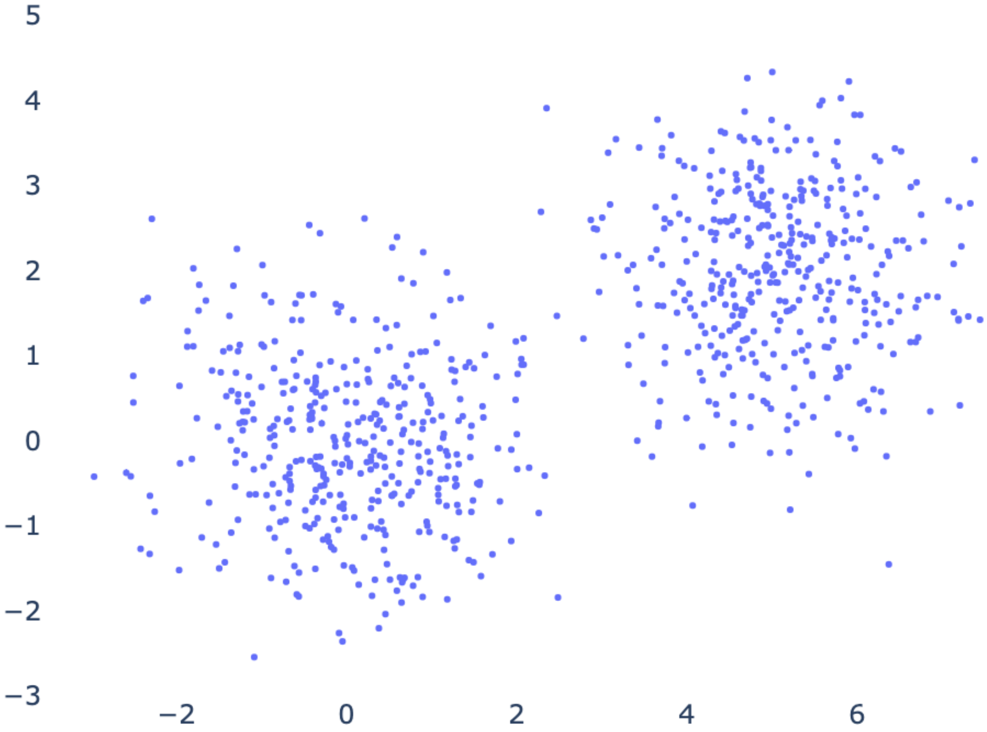
X_2	p
H	$0.18/0.3 = 0.6$
T	$0.12/0.3 = 0.4$

The distribution of the second coin remains the same: there is still a 60% probability of heads, unchanged from before we made our observation of the first coin. This property, the distribution of one random variable is unchanged when we make observations about the other, is what defines independence.

3.2 Conditional independence

Sometimes our random variables aren't independent to start out, but they **become** independent after we observe something. This is called *conditional independence*. Under conditional independence, our distribution isn't a product of independent factors to start, but it becomes a product of independent factors after we make an observation.

A good example is a clustered distribution:



In this distribution, the X and Y coordinates are not independent: if X is big, then we're more likely to be in the second cluster, meaning that Y is more likely to be big as well. But once we know which cluster we're in, the lower left one or the upper right one, X and Y are independent.

If the conditional distribution is independent, that means that it factors into the product of the probability of X and the probability of Y:

$$P(X, Y \mid Z) = P(X \mid Z)P(Y \mid Z)$$

where X, Y are the coordinates of a sample and Z is the indicator of which cluster the sample is in. Thus, we can express the full joint distribution as

$$P(X, Y, Z) = P(Z)P(X, Y \mid Z) = P(Z)P(X \mid Z)P(Y \mid Z)$$

We can test that this formula gives us conditional independence by conditioning on an observation, say $Z = 1$, i.e., the point is in the first cluster. The rule for conditional probabilities gives us

$$\begin{aligned} P(X, Y \mid Z = 1) &= P(X, Y, Z = 1)/P(Z = 1) \\ &= P(Z = 1)P(X \mid Z = 1)P(Y \mid Z = 1)/P(Z = 1) \\ &= P(X \mid Z = 1)P(Y \mid Z = 1) \end{aligned}$$

Since we've fixed a value for Z, $P(X \mid Z = 1)$ depends only on X, and $P(Y \mid Z = 1)$ depends only on Y. So, our conditional distribution is a product of two independent factors, as claimed.

3.3 Samples

One of the most common uses of independence or conditional independence is when we repeat an experiment many times to collect a sample. In this situation it makes sense to assume that each run of the experiment is independent from all of the other runs. If our sample is $X_1, X_2, \ldots, X_T$, that means our overall distribution factors as

$$P(X_1, X_2, \ldots, X_T) = P(X_1)P(X_2)\ldots P(X_T)$$

In a sample like this, we might have an unknown parameter vector $\theta \in \mathbb{R}^d$ that influences the distribution of our samples, e.g., θ might contain the mean and variance of a sample X_t. If we think of θ as fixed but unknown, we could emphasize the dependence on θ by writing

$$P_\theta(X_1, X_2, \ldots, X_T) = P_\theta(X_1)P_\theta(X_2)\ldots P_\theta(X_T)$$

On the other hand, we might want to think of θ itself as a random variable. In this case we would say that the samples X_t are conditionally independent given θ:

$$P(X_1, X_2, \ldots, X_T, \theta) = P(\theta)P(X_1 \mid \theta)P(X_2 \mid \theta)\ldots P(X_T \mid \theta)$$

These are two alternate views of the world: either the parameters take some fixed but unknown value, or the parameters are themselves random. Both views are reasonable; they often lead to similar conclusions about θ, but they can also be subtly different.

Knowledge check

1. **Math**: Using the probability identities we've defined in this reading, show that if two random variables are independent, then $P(X_1 \mid X_2) = P(X_1)$ i.e., the result we empirically demonstrated in subsection 3.1.
   - **Hint**: First use the definition of a conditional probability to express the quantity $P(X_1 \mid X_2)$ in terms of $P(X_1, X_2)$. Then apply the property of independent random variables that $P(X_1, X_2) = P(X_1)P(X_2)$.
   - **Answer**:

$$\begin{aligned} P(X_1 \mid X_2) &= P(X_1, X_2)/P(X_2) \\ &= P(X_1)P(X_2)/P(X_2) \\ &= P(X_1) \end{aligned}$$