Probability Basics

Martin Emms

September 15, 2022

Probability Background

Probability Background

- you have an variable/feature/attribute of a system and it takes on values in some specific set. The classic example is dice throwing, with the feature being the uppermost face of the dice, taking values in $\{1, 2, 3, 4, 5, 6\}$

- you have an variable/feature/attribute of a system and it takes on values in some specific set. The classic example is dice throwing, with the feature being the uppermost face of the dice, taking values in $\{1, 2, 3, 4, 5, 6\}$
- you talk of the probability of a particular feature value: $P(X = a)$

- you have an variable/feature/attribute of a system and it takes on values in some specific set. The classic example is dice throwing, with the feature being the uppermost face of the dice, taking values in $\{1, 2, 3, 4, 5, 6\}$
- you talk of the probability of a particular feature value: $P(X = a)$
- standard frequentist interpretation is that the systems can be observed over and over again, and that the relative frequency of $X = a$ in all the observations tends to a stable fixed value as the number of observations tends to infinity. $P(X = a)$ is this limit

$$P(X = a) = \lim_{N \to \infty} freq(X = a)/N$$

- on this frequentist interpretation you would definitely expect the sum over different outcomes to be 1, so where A is set of possible values for feature X, it is always assumed that

$$\sum_{a \in A} P(X = a) = 1$$

- on this frequentist interpretation you would definitely expect the sum over different outcomes to be 1, so where A is set of possible values for feature X, it is always assumed that

$$\sum_{a \in A} P(X = a) = 1$$

- typically also interested in types or kinds of outcome: not the probability of any particular value $X = a$. Jargon for this is event
- for example, the 'event' of dice throw being even can be described as $(X = 2 \vee X = 4 \vee X = 6)$

- on this frequentist interpretation you would definitely expect the sum over different outcomes to be 1, so where A is set of possible values for feature X, it is always assumed that

$$\sum_{a \in A} P(X = a) = 1$$

- typically also interested in types or kinds of outcome: not the probability of any particular value $X = a$. Jargon for this is event
- for example, the 'event' of dice throw being even can be described as $(X = 2 \vee X = 4 \vee X = 6)$
- the relative freq. of $(2 \; or \; 4 \; or \; 6)$ is by definition the same as the $(rel.freq \; 2) + (rel.freq. \; 4) + (rel.freq. \; 6)$. So its not surprising that by definition the probability of an 'event' is the sum of the mutually exclusive atomic possibilities that are contained within it (ie. ways for it to happen) so

$$P(X = 2 \vee X = 4 \vee X = 6) = P(X = 2) + P(X = 4) + P(X = 6)$$

Independence of two events

- suppose two 'events' A and B . If the probability of $A \wedge B$ occuring is just the probability A occuring times the probability of B occuring, you say the events A and B are independent

$$Independence : P(A \wedge B) = P(A) \times P(B)$$

Independence of two events

- suppose two 'events' A and B . If the probability of $A \wedge B$ occuring is just the probability A occuring times the probability of B occuring, you say the events A and B are independent

$$\textit{Independence} : P(A \wedge B) = P(A) \times P(B)$$

- Related idea is conditional probability, the probability of A given B: instead of considering how often A occurs, you just consider how often A occurs in situation which are already B situations.

Independence of two events

- suppose two 'events' A and B . If the probability of $A \wedge B$ occuring is just the probability A occuring times the probability of B occuring, you say the events A and B are independent

$$\mathit{Independence} : P(A \wedge B) = P(A) \times P(B)$$

- Related idea is conditional probability, the probability of A given B: instead of considering how often A occurs, you just consider how often A occurs in situation which are already B situations.
- This is defined to be

Conditional Prob

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

- there's a common-sense 'explanation' for the definition

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

- there's a common-sense 'explanation' for the definition

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

- you want to take the limit as N tends to infinity of

$$\lim_{N \to \infty} \left(\frac{count(A \wedge B)\ in\ N}{count(B)\ in\ N}\right)$$

- there's a common-sense 'explanation' for the definition

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

- you want to take the limit as N tends to infinity of

$$\lim_{N \to \infty} \left(\frac{count(A \wedge B)\ in\ N}{count(B)\ in\ N}\right)$$

you get the same thing if you divide top and bottom by N, so

$$\lim_{N \to \infty} \left(\frac{count(A \wedge B)\ in\ N}{count(B)\ in\ N}\right) \quad = \quad \lim_{N \to \infty} \frac{(count(A \wedge B)\ in\ N)/N}{(count(B)\ in\ N)/N}$$

- there's a common-sense 'explanation' for the definition

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

- you want to take the limit as N tends to infinity of

$$\lim_{N \to \infty} (\frac{count(A \wedge B)\ in\ N}{count(B)\ in\ N})$$

you get the same thing if you divide top and bottom by N, so

$$\begin{aligned}
\lim_{N \to \infty} (\frac{count(A \wedge B)\ in\ N}{count(B)\ in\ N}) &= \lim_{N \to \infty} \frac{(count(A \wedge B)\ in\ N)/N}{(count(B)\ in\ N)/N} \\
&= \frac{\lim_{N \to \infty}(count(A \wedge B)\ in\ N)/N}{\lim_{N \to \infty}(count(B)\ in\ N)/N}
\end{aligned}$$

- there's a common-sense 'explanation' for the definition

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

- you want to take the limit as N tends to infinity of

$$\lim_{N \to \infty} \left(\frac{count(A \wedge B)\ in\ N}{count(B)\ in\ N} \right)$$

you get the same thing if you divide top and bottom by N, so

$$\begin{aligned}
\lim_{N \to \infty} \left(\frac{count(A \wedge B)\ in\ N}{count(B)\ in\ N} \right) &= \lim_{N \to \infty} \frac{(count(A \wedge B)\ in\ N)/N}{(count(B)\ in\ N)/N} \\
&= \frac{\lim_{N \to \infty}(count(A \wedge B)\ in\ N)/N}{\lim_{N \to \infty}(count(B)\ in\ N)/N} \\
&= \frac{P(A \wedge B)}{P(B)}
\end{aligned}$$

- obviously given the definition of $P(A|B)$, you have the obvious but as it turns out very useful

Product Rule

$$P(A \wedge B) = P(A|B)P(B)$$

- obviously given the definition of $P(A|B)$, you have the obvious but as it turns out very useful

Product Rule

$$P(A \wedge B) = P(A|B)P(B)$$

- since $P(A|B)P(B) = P(B|A)P(A)$, you also get the famous

Bayesian Inversion

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

Alternative expressions of independence

- recall independence was defined to be $P(A \wedge B) = P(A) \times P(B)$. Given the definition of conditional probability there are equivalent formulations of independence in terms of conditional probability:

Alternative expressions of independence

- recall independence was defined to be $P(A \wedge B) = P(A) \times P(B)$. Given the definition of conditional probability there are equivalent formulations of independence in terms of conditional probability:

$$\textit{Independence} : P(A|B) = P(A)$$

$$\textit{Independence} : P(B|A) = P(B)$$

Alternative expressions of independence

- recall independence was defined to be $P(A \wedge B) = P(A) \times P(B)$. Given the definition of conditional probability there are equivalent formulations of independence in terms of conditional probability:

$$\mathit{Independence} : P(A|B) = P(A)$$

$$\mathit{Independence} : P(B|A) = P(B)$$

NOTE: each of these *on its own* is equivalent to $P(A \wedge B) = P(A) \times P(B)$

- Suppose > 1 feature/attribute of your system/situation eg. rolling a red & a green dice. Using X for red & Y for green can specify events with their values and their probs with expressions such as:[1]

$$P(X = 1, Y = 2)$$

[1] note comma often used instead of $\wedge$

- Suppose > 1 feature/attribute of your system/situation eg. rolling a red & a green dice. Using X for red & Y for green can specify events with their values and their probs with expressions such as:[1]

$$P(X = 1, Y = 2)$$

and the probability of such an event is called a joint probability

[1] note comma often used instead of $\wedge$

- Suppose > 1 feature/attribute of your system/situation eg. rolling a red & a green dice. Using X for red & Y for green can specify events with their values and their probs with expressions such as:[1]

$$P(X = 1, Y = 2)$$

  and the probability of such an event is called a joint probability
- if A is range of values for X & B is range for Y, the must have

$$\sum_{a \in A, b \in B} P(X = a, Y = b) = 1$$

[1] note comma often used instead of $\wedge$

- Suppose > 1 feature/attribute of your system/situation eg. rolling a red & a green dice. Using X for red & Y for green can specify events with their values and their probs with expressions such as:[1]

$$P(X = 1, Y = 2)$$

and the probability of such an event is called a joint probability

- if A is range of values for X & B is range for Y, the must have

$$\sum_{a \in A, b \in B} P(X = a, Y = b) = 1$$

- can wish to consider the probs of events specified by the value on just one feature (eg. those where X=1) and the probs. of these are called marginal probabilities and are obtained by summing the joints with all possible values of the other feature

$$P(X = 1) = \sum_{b \in B} P(X = 1, Y = b)$$

[1] note comma often used instead of $\wedge$

- the conditional probability function for two features X and Y is

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

so for any pair of values a for X and b for Y, the value of this function is $P(X = a, Y = b)/P(Y = b)$

- the conditional probability function for two features X and Y is

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

  so for any pair of values a for X and b for Y, the value of this function is $P(X = a, Y = b)/P(Y = b)$
- you say $P(X|Y) = P(X)$ and the features X and Y are independent in case for every value a for X and b for Y you have

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

Chain Rule

- generalising to more variables, you can derive the indispensable

chain rule

$$P(X, Y, Z) = P(Z|(X, Y)) \times P(X, Y) = P(Z|(X, Y)) \times P(Y|X) \times P(X)$$

$$P(X_1 \dots X_n) = P(X_n|(X_1 \dots X_{n-1})) \times \dots \times P(X_2|X_1) \times P(X_1)$$

Chain Rule

- generalising to more variables, you can derive the indispensable

chain rule

$$P(X, Y, Z) = P(Z|(X, Y)) \times P(X, Y) = P(Z|(X, Y)) \times P(Y|X) \times P(X)$$

$$P(X_1 \dots X_n) = P(X_n|(X_1 \dots X_{n-1})) \times \dots \times P(X_2|X_1) \times P(X_1)$$

important to note that this chain-rule re-expression of a joint probability as a product does not make any independence assumptions

Chain Rule

- generalising to more variables, you can derive the indispensable

chain rule

$$P(X,Y,Z) = P(Z|(X,Y)) \times P(X,Y) = P(Z|(X,Y)) \times P(Y|X) \times P(X)$$

$$P(X_1 \ldots X_n) = P(X_n|(X_1 \ldots X_{n-1})) \times \ldots \times P(X_2|X_1) \times P(X_1)$$

important to note that this chain-rule re-expression of a joint probability as a product does not make any independence assumptions

Notation: typically $P(Z|(X,Y))$ is written as $P(Z|X,Y)$

Conditional Independence

- there is a notion of conditional independence. It may be that two variables X and Y are not in general independent, but given a value for a third variable Z, X and Y become independent.

Conditional Indpt

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

Conditional Independence

- there is a notion of conditional independence. It may be that two variables X and Y are not in general independent, but given a value for a third variable Z, X and Y become independent.

Conditional Indpt

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

- as with straightforward independence there is an alternative expression for this, stating how a conditioning factor can be dropped

Conditional Indpt altern. def

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

Conditional Independence

- there is a notion of conditional independence. It may be that two variables X and Y are not in general independent, but given a value for a third variable Z, X and Y become independent.

Conditional Indpt

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

- as with straightforward independence there is an alternative expression for this, stating how a conditioning factor can be dropped

Conditional Indpt altern. def

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

- Real-life cases of this arise where Z describes a *cause*, which manifests itself into two *effects* X and Y, which though very dependent on Z, do not directly influence each other
- The theories behind Speech Recognition and Machine Translation typically make a lot of *conditional independence* assumptions