Fundamentals of Machine Learning

Appendix B - Introduction to Probability for Machine Learning

- Probability is the branch of mathematics that deals with measuring the likelihood (or uncertainty) around events.
- There are two ways of computing the likelihood of a future event:
  1. use the **relative frequency** of the event in the past,
  2. use a **subjective** estimate (ideally from an expert!)
- We will focus on using relative frequency.

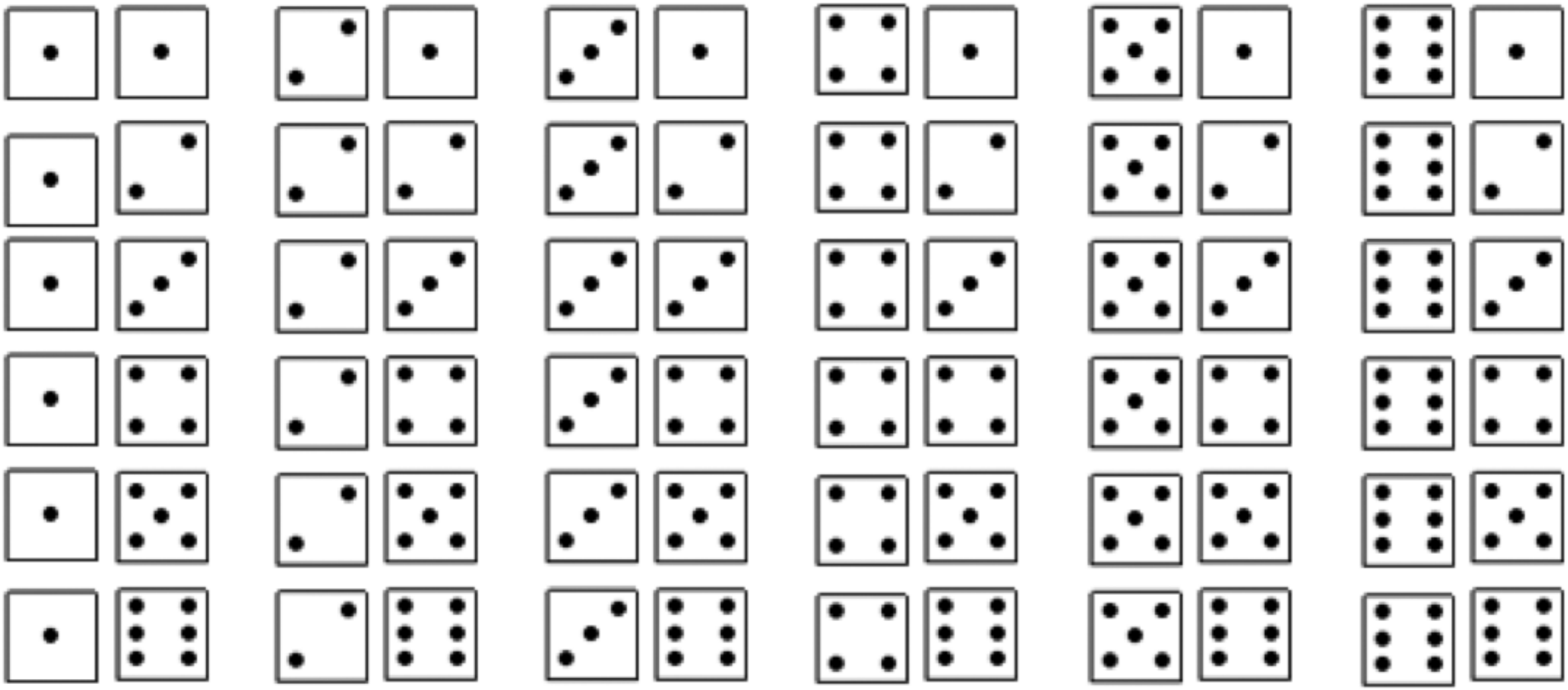

Figure: The **sample space** for the domain of two dice.

ID	Die1	Die2
1	3	4
2	1	5
3	6	5
4	3	3
5	1	1

Table: A dataset of instances from the sample space in Figure 1 [4].

- Throughout our discussions about probability we will be talking about **events** so it is very important that you understand this simple concept.

Events

- an **event** defines an assignment of values to the features in the domain; these assignments may define values for all the features in the domain (e.g. a full row in the dataset) or just to one or more features in the domain,

Example

- DIE1=*'3'*,
- DIE1=*'1'*, DIE2=*'5'*.

Probability Functions: $P()$

- A feature can take one or more values from a domain and we can find out the likelihood of a feature taking any particular value using a **probability function** $P()$.
- A probability function is a function that takes an event (an assignment of values to features) as a parameter and returns the likelihood of that event.

Example

- $P(\text{DIE1} = \textit{'3'})$ will return the likelihood of the event $\text{DIE1} = \textit{'3'}$
- $P(\text{DIE1} = \textit{'3'}, \text{DIE2} = \textit{'4'})$ will return the likelihood of the event where DIE1=*'3'* and DIE2=*'4'*.

Properties of Probability Functions

$$
\begin{aligned}
0 \leq P(f = level) &\leq 1 \\
\sum_i P(f = level_i) &= 1.0
\end{aligned}
$$

- Probability functions are very easy to create when you have a dataset.
- The value returned by a probability function for an event is simply the **relative frequency** of that event in the dataset – in other words, how often the event happened divided by how often could it have happened.

Example

- The relative frequency of the event $\text{DIE1} = \textit{'3'}$ is simply the count of all the rows in the dataset where the feature is assigned the relevant value divided by the number of rows in the dataset

Prior Probability (aka. Unconditional Probabilities)

- The probability of an event without any contextual information.
- The count of all the rows in the dataset where the feature(s) is assigned the relevant value(s) divided by the number of rows in the dataset.

Example

$$P(\text{DIE1} = \text{'3'}) = \frac{|\{\mathbf{d}_1, \mathbf{d}_4\}|}{|\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5\}|} = \frac{2}{5} = 0.4$$

Joint Probability

- The probability of two or more events happening together.
- The number of rows in the dataset where the set of assignments listed in the joint event holds divided by the total number of rows in the dataset.

Example

$$P(\text{DIE1} = \text{'6'}, \text{DIE2} = \text{'5'}) = \frac{|\{\mathbf{d}_3\}|}{|\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5\}|} = \frac{1}{5} = 0.2$$

Posterior Probabilities (aka. Conditional Probabilities)

- The probability of an event in a context where one or more events are known to have happened.
- The vertical bar symbol | can be read as *given that*.
- The number of rows in the dataset where both events are true divided by the number of rows in the dataset where just the given event is true.

Example

$$P(\text{DIE1} = \text{'6'} \mid \text{DIE2} = \text{'5'}) = \frac{|\{\mathbf{d}_3\}|}{|\{\mathbf{d}_2, \mathbf{d}_3\}|} = \frac{1}{2} = 0.5$$

Table: A simple dataset for MENINGITIS with three common symptoms of the disease listed as descriptive features: HEADACHE, FEVER and VOMITING.

ID	Headache	Fever	Vomiting	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

Your Turn!

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

- $P(h) = ?$
- $P(m|h) = ?$
- $P(m, h) = ?$

Your Turn!

$$P(h) = \frac{|\{\mathbf{d}_{11}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{37}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{54}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{89}, \mathbf{d}_{92}\}|} = \frac{7}{10} = 0.7$$

$$P(m|h) = \frac{|\{\mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{92}\}|} = \frac{2}{7} = 0.2857$$

$$P(m, h) = \frac{|\{\mathbf{d}_{75}, \mathbf{d}_{92}\}|}{|\{\mathbf{d}_{11}, \mathbf{d}_{37}, \mathbf{d}_{42}, \mathbf{d}_{49}, \mathbf{d}_{54}, \mathbf{d}_{57}, \mathbf{d}_{73}, \mathbf{d}_{75}, \mathbf{d}_{89}, \mathbf{d}_{92}\}|} = \frac{2}{10} = 0.2$$

Probability Distributions

- A probability distribution is a data structure that describes for all the values in the domain of a feature the probability of the feature taking that value.
- A probability distribution of a categorical feature is a vector that lists the probabilities associated with the values in the domain of the feature.
- We use bold notation $\mathbf{P}()$ to distinguish when we are talking about a probability distribution from a probability mass function $P()$.

Example

- Based on the following dataset the probability distribution for the binary feature, MENINGITIS, using the convention of the first element in the vector being the probability for *'True'*, would be:

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

$$\mathbf{P}(M) =< 0.3, 0.7 >$$

Joint Probability Distributions

- is a multi-dimensional matrix where each cell in the matrix list the probability for one of the events in the sample space defined by the combination of feature values.

Example

- The joint probability distribution for the four binary features HEADING, FEVER, VOMITING, MENINGITIS in the Meningitis domain would be:

$$\mathbf{P}(H, F, V, M) = \begin{bmatrix} P(h, f, v, m), & P(\neg h, f, v, m) \\ P(h, f, v, \neg m), & P(\neg h, f, v, \neg m) \\ P(h, f, \neg v, m), & P(\neg h, f, \neg v, m) \\ P(h, f, \neg v, \neg m), & P(\neg h, f, \neg v, \neg m) \\ P(h, \neg f, v, m), & P(\neg h, \neg f, v, m) \\ P(h, \neg f, v, \neg m), & P(\neg h, \neg f, v, \neg m) \\ P(h, \neg f, \neg v, m), & P(\neg h, \neg f, \neg v, m) \\ P(h, \neg f, \neg v, \neg m), & P(\neg h, \neg f, \neg v, \neg m) \end{bmatrix}$$

- A **full joint probability distribution** is simply a joint probability distribution over all the features in a domain.

Summing out (aka Marginalisation)

- Given a full joint probability we can compute the probability of any event in the domain by summing over the cells in the joint probability where that event is true.

Example

- Imagine we want to compute the probability of $P(h)$ in the domain specified by the joint probability distribution $\mathbf{P}(H, F, V, M)$.
- Simply sum the values in the cells containing h, in other words the cells in the first column in the full joint probability.

$$\mathbf{P}(H, F, V, M) = \begin{bmatrix} P(h, f, v, m), & P(\neg h, f, v, m) \\ P(h, f, v, \neg m), & P(\neg h, f, v, \neg m) \\ P(h, f, \neg v, m), & P(\neg h, f, \neg v, m) \\ P(h, f, \neg v, \neg m), & P(\neg h, f, \neg v, \neg m) \\ P(h, \neg f, v, m), & P(\neg h, \neg f, v, m) \\ P(h, \neg f, v, \neg m), & P(\neg h, \neg f, v, \neg m) \\ P(h, \neg f, \neg v, m), & P(\neg h, \neg f, \neg v, m) \\ P(h, \neg f, \neg v, \neg m), & P(\neg h, \neg f, \neg v, \neg m) \end{bmatrix}$$

- We can also use summing out to compute joint probabilities from a joint probability distribution.

Example

- Imagine we wish to calculate the probability of h and f when we don't care what value V and M take (here, V and M are examples of a hidden feature; a feature whose value is not specified as part of the evidence and which is not a target feature).

Example

- To calculate $P(h, V = ?, M = ?, f)$ from $\mathbf{P}(H, V, F, M)$ by summing the values in all the cells where h and f are the case (in other words summing the top four cells in column one).

$$\mathbf{P}(H, F, V, M) = \begin{bmatrix} P(h, f, v, m), & P(\neg h, f, v, m) \\ P(h, f, v, \neg m), & P(\neg h, f, v, \neg m) \\ P(h, f, \neg v, m), & P(\neg h, f, \neg v, m) \\ P(h, f, \neg v, \neg m), & P(\neg h, f, \neg v, \neg m) \\ P(h, \neg f, v, m), & P(\neg h, \neg f, v, m) \\ P(h, \neg f, v, \neg m), & P(\neg h, \neg f, v, \neg m) \\ P(h, \neg f, \neg v, m), & P(\neg h, \neg f, \neg v, m) \\ P(h, \neg f, \neg v, \neg m), & P(\neg h, \neg f, \neg v, \neg m) \end{bmatrix}$$

Conditional Probability

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{1}$$

- Use this rule to recalculate the probability of $P(m|h)$ (recall that $P(h) = 0.7$ and $P(m,h) = 0.2$

Conditional Probability

$$P(X|Y) = \frac{P(X,Y)}{P(Y)} \tag{1}$$

- Use this rule to recalculate the probability of $P(m|h)$ (recall that $P(h) = 0.7$ and $P(m,h) = 0.2$

Example

$$P(m|h) = \frac{P(m,h)}{P(h)} = \frac{0.2}{0.7} = 0.2857$$

Product Rule

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

- Note: $P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$

Example

- Use the product rule to recalculate $P(m, h)$.

Product Rule

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

- Note: $P(X, Y) = P(X|Y)P(Y) = P(Y|X)P(X)$

Example

- Use the product rule to recalculate $P(m, h)$.

$$P(m, h) = P(m|h) \times P(h) = 0.2857 \times 0.7 = 0.2$$

Chain Rule

- The Product Rule:

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

- generalizes to the Chain Rule:

$$\begin{aligned} P(A, B, C, \ldots, Z) =& P(Z) \times P(Y|Z) \times P(X|Y, Z) \times \ldots \\ & \times P(A|B, \ldots, X, Y, Z) \end{aligned}$$

Theorem of Total Probability

$$P(X) = \sum_{i=1}^{k} P(X|Y_i)P(Y_i)$$

Example

- Use the Theorem of Total Probability to recalculate $P(h)$ by summing out M.

$$P(h) = (P(h|m) \times P(m)) + (P(h|\neg m) \times P(\neg m))$$

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

- $P(h|m) =?$
- $P(m) =?$
- $P(h|\neg m) =?$
- $P(\neg m) =?$

ID	Headach	Fever	Vomit	Meningitis
11	True	True	False	False
37	False	True	False	False
42	True	False	True	False
49	True	False	True	False
54	False	True	False	True
57	True	False	True	False
73	True	False	True	False
75	True	False	True	True
89	False	True	False	False
92	True	False	True	True

- $P(h|m) = 0.6666$
- $P(m) = 0.3$
- $P(h|\neg m) = 0.7143$
- $P(\neg m) = 0.7$

$$
\begin{aligned}
P(h) &= (P(h|m) \times P(m)) + (P(h|\neg m) \times P(\neg m)) \\
&= ?
\end{aligned}
$$

$$
\begin{aligned}
P(h) &= (P(h|m) \times P(m)) + (P(h|\neg m) \times P(\neg m)) \\
&= (0.6666 \times 0.3) + (0.7143 \times 0.7) = 0.7
\end{aligned}
$$

- We can if we wish sum out more than one feature.

Example

- For example, we could compute $P(h)$ by summing out all the other features in the dataset:

$$P(h) = \sum_{i \in level(M)} \sum_{j \in level(Fev)} \sum_{k \in level(V)} P(h|M_i, Fev_j, V_k) \times P(M_i, Fev_j, V_k)$$

Summary