Probability-based Learning

Chapter 6

Péter Molnár[1]

J. Mack Robinson College of Business
Georgia State University

MSA8150 – Spring 2016

[1] Original slides from John Kelleher, Brian Mac Namee, and Aoife D'Arcy

Big Idea

- We can use estimates of likelihoods to determine the most likely prediction that should be made.
- More importantly, we revise these predictions based on data we collect and whenever extra evidence becomes available.

Table: A simple dataset for MENINGITIS diagnosis with descriptive features that describe the presence or absence of three common symptoms of the disease: HEADACHE, FEVER, and VOMITING.

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- A **probability function**, $P()$, returns the probability of a feature taking a specific value.
- A **joint probability** refers to the probability of an assignment of specific values to multiple different features.
- A **conditional probability** refers to the probability of one feature taking a specific value given that we already know the value of a different feature
- A **probability distribution** is a data structure that describes the probability of each possible value a feature can take. The sum of a probability distribution must equal 1.0.
- A **joint probability distribution** is a probability distribution over more than one feature assignment and is written as a multi-dimensional matrix in which each cell lists the probability of a particular combination of feature values being assigned.
- The sum of all the cells in a joint probability distribution must be 1.0.

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

- Given a joint probability distribution, we can compute the probability of any event in the domain that it covers by summing over the cells in the distribution where that event is true.
- Calculating probabilities in this way is known as **summing out**.

Bayes' Theorem

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

Example

After a yearly checkup, a doctor informs their patient that he has both bad news and good news. The bad news is that the patient has tested positive for a serious disease and that the test that the doctor has used is 99% accurate (i.e., the probability of testing positive when a patient has the disease is 0.99, as is the probability of testing negative when a patient does not have the disease). The good news, however, is that the disease is extremely rare, striking only 1 in 10,000 people.

- What is the actual probability that the patient has the disease?
- Why is the rarity of the disease good news given that the patient has tested positive for it?

$$P(d|t) = \frac{P(t|d)P(d)}{P(t)}$$

$$\begin{aligned} P(t) &= P(t|d)P(d) + P(t|\neg d)P(\neg d) \\ &= (0.99 \times 0.0001) + (0.01 \times 0.9999) = 0.0101 \end{aligned}$$

$$\begin{aligned} P(d|t) &= \frac{0.99 \times 0.0001}{0.0101} \\ &= 0.0098 \end{aligned}$$

Deriving Bayes theorem

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

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

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

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

- The divisor is the prior probability of the evidence
- This division functions as a normalization constant.

$$
\begin{aligned}
0 \leq P(X|Y) &\leq 1 \\
\sum_i P(X_i|Y) &= 1.0
\end{aligned}
$$

- We can calculate this divisor directly from the dataset.

$$P(Y) = \frac{|\{\text{rows where Y is the case}\}|}{|\{\text{rows in the dataset}\}|}$$

- Or, we can use the **Theorem of Total Probability** to calculate this divisor.

$$P(Y) = \sum_i P(Y|X_i)P(X_i) \tag{1}$$

Generalized Bayes' Theorem

$$P(t = l|\mathbf{q}[1], \ldots, \mathbf{q}[m]) = \frac{P(\mathbf{q}[1], \ldots, \mathbf{q}[m]|t = l)P(t = l)}{P(\mathbf{q}[1], \ldots, \mathbf{q}[m])}$$

Chain Rule

$$
\begin{aligned}
P(\mathbf{q}[1], \ldots, \mathbf{q}[m]) = \\
& P(\mathbf{q}[1]) \times P(\mathbf{q}[2]|\mathbf{q}[1]) \times \\
& \cdots \times P(\mathbf{q}[m]|\mathbf{q}[m-1], \ldots, \mathbf{q}[2], \mathbf{q}[1])
\end{aligned}
$$

- To apply the chain rule to a conditional probability we just add the conditioning term to each term in the expression:

$$
\begin{aligned}
P(\mathbf{q}[1], \ldots, \mathbf{q}[m]|t = l) = \\
& P(\mathbf{q}[1]|t = l) \times P(\mathbf{q}[2]|\mathbf{q}[1], t = l) \times \ldots \\
& \cdots \times P(\mathbf{q}[m]|\mathbf{q}[m-1], \ldots, \mathbf{q}[3], \mathbf{q}[2], \mathbf{q}[1], t = l)
\end{aligned}
$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

HEADACHE	FEVER	VOMITING	MENINGITIS
true	false	true	?

$$P(M|h, \neg f, v) = ?$$

- In the terms of Bayes' Theorem this problem can be stated as:

$$P(M|h, \neg f, v) = \frac{P(h, \neg f, v|M) \times P(M)}{P(h, \neg f, v)}$$

- There are two values in the domain of the MENINGITIS feature, *'true'* and *'false'*, so we have to do this calculation twice.

- We will do the calculation for m first
- To carry out this calculation we need to know the following probabilities: $P(m)$, $P(h, \neg f, v)$ and $P(h, \neg f, v \mid m)$.

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- We can calculate the required probabilities directly from the data. For example, we can calculate $P(m)$ and $P(h, \neg f, v)$ as follows:

$$P(m) = \frac{|\{\mathbf{d}_5, \mathbf{d}_8, \mathbf{d}_{10}\}|}{|\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_6, \mathbf{d}_7, \mathbf{d}_8, \mathbf{d}_9, \mathbf{d}_{10}\}|} = \frac{3}{10} = 0.3$$

$$P(h, \neg f, v) = \frac{|\{\mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_6, \mathbf{d}_7, \mathbf{d}_8, \mathbf{d}_{10}\}|}{|\{\mathbf{d}_1, \mathbf{d}_2, \mathbf{d}_3, \mathbf{d}_4, \mathbf{d}_5, \mathbf{d}_6, \mathbf{d}_7, \mathbf{d}_8, \mathbf{d}_9, \mathbf{d}_{10}\}|} = \frac{6}{10} = 0.6$$

- However, as an exercise we will use the chain rule calculate:

$$P(h, \neg f, v \mid m) = ?$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- Using the chain rule calculate:

$$
\begin{aligned}
P(h, \neg f, v \mid m) &= P(h \mid m) \times P(\neg f \mid h, m) \times P(v \mid \neg f, h, m) \\
&= \frac{|\{\mathbf{d}_8, \mathbf{d}_{10}\}|}{|\{\mathbf{d}_5, \mathbf{d}_8, \mathbf{d}_{10}\}|} \times \frac{|\{\mathbf{d}_8, \mathbf{d}_{10}\}|}{|\{\mathbf{d}_8, \mathbf{d}_{10}\}|} \times \frac{|\{\mathbf{d}_8, \mathbf{d}_{10}\}|}{|\{\mathbf{d}_8, \mathbf{d}_{10}\}|} \\
&= \frac{2}{3} \times \frac{2}{2} \times \frac{2}{2} = 0.6666
\end{aligned}
$$

- So the calculation of $P(m|h, \neg f, v)$ is:

$$P(m|h, \neg f, v) = \frac{\left(\begin{array}{l} P(h|m) \times P(\neg f|h, m) \\ \quad \times P(v|\neg f, h, m) \times P(m) \end{array} \right)}{P(h, \neg f, v)}$$
$$= \frac{0.6666 \times 0.3}{0.6} = 0.3333$$

- The corresponding calculation for $P(\neg m|h, \neg f, v)$ is:

$$
\begin{aligned}
P(\neg m \mid h, \neg f, v) &= \frac{P(h, \neg f, v \mid \neg m) \times P(\neg m)}{P(h, \neg f, v)} \\
&= \frac{\left(\begin{array}{l} P(h|\neg m) \times P(\neg f \mid h, \neg m) \\ \quad \times P(v|\neg f, h, \neg m) \times P(\neg m) \end{array} \right)}{P(h, \neg f, v)} \\
&= \frac{0.7143 \times 0.8 \times 1.0 \times 0.7}{0.6} = 0.6667
\end{aligned}
$$

$$P(m|h, \neg f, v) = 0.3333$$

$$P(\neg m|h, \neg f, v) = 0.6667$$

- These calculations tell us that it is twice as probable that the patient does not have meningitis than it is that they do even though the patient is suffering from a headache and is vomiting!

The Paradox of the False Positive

- The mistake of forgetting to factor in the prior gives rise to the **paradox of the false positive** which states that in order to make predictions about a rare event the model has to be as accurate as the prior of the event is rare or there is a significant chance of **false positives** predictions (i.e., predicting the event when it is not the case).

Bayesian MAP Prediction Model

$$\begin{aligned}\mathbb{M}_{MAP}(\mathbf{q}) &= \underset{l \in levels(t)}{\text{argmax}}\, P(t = l \mid \mathbf{q}[1], \ldots, \mathbf{q}[m]) \\ &= \underset{l \in levels(t)}{\text{argmax}}\, \frac{P(\mathbf{q}[1], \ldots, \mathbf{q}[m] \mid t = l) \times P(t = l)}{P(\mathbf{q}[1], \ldots, \mathbf{q}[m])}\end{aligned}$$

Bayesian MAP Prediction Model (without normalization)

$$\mathbb{M}_{MAP}(\mathbf{q}) = \underset{l \in levels(t)}{\text{argmax}}\, P(\mathbf{q}[1], \ldots, \mathbf{q}[m] \mid t = l) \times P(t = l)$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

$$P(m \mid h, f, \neg v) = ?$$

$$P(\neg m \mid h, f, \neg v) = ?$$

$$P(m \mid h, f, \neg v) = \frac{\left(\begin{array}{l} P(h|m) \times P(f \mid h, m) \\ \quad \times P(\neg v \mid f, h, m) \times P(m) \end{array} \right)}{P(h, f, \neg v)}$$
$$= \frac{0.6666 \times 0 \times 0 \times 0.3}{0.1} = 0$$

$$P(\neg m \mid h, f, \neg v) = \frac{\left(\begin{array}{l} P(h|\neg m) \times P(f \mid h, \neg m) \\ \quad \times P(\neg v \mid f, h, \neg m) \times P(\neg m) \end{array} \right)}{P(h, f, \neg v)}$$
$$= \frac{0.7143 \times 0.2 \times 1.0 \times 0.7}{0.1} = 1.0$$

$$P(m \mid h, f, \neg v) = 0$$

$$P(\neg m \mid h, f, \neg v) = 1.0$$

- There is something odd about these results!

Curse of Dimensionality

As the number of descriptive features grows the number of potential conditioning events grows. Consequently, an exponential increase is required in the size of the dataset as each new descriptive feature is added to ensure that for any conditional probability there are enough instances in the training dataset matching the conditions so that the resulting probability is reasonable.

- The probability of a patient who has a headache and a fever having meningitis should be greater than zero!
- Our dataset is not large enough $\rightarrow$ our model is over-fitting to the training data.
- The concepts of conditional independence and factorization can help us overcome this flaw of our current approach.

- If knowledge of one event has no effect on the probability of another event, and *vice versa*, then the two events are **independent** of each other.
- If two events X and Y are independent then:

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

- Recall, that when two event are dependent these rules are:

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

- Full independence between events is quite rare.
- A more common phenomenon is that two, or more, events may be independent if we know that a third event has happened.
- This is known as conditional independence.

- For two events, X and Y, that are conditionally independent given knowledge of a third events, here Z, the definition of the probability of a joint event and conditional probability are:

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

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

X and Y are dependent

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

X and Y are independent

- If the event $t = l$ causes the events $\mathbf{q}[1], \ldots, \mathbf{q}[m]$ to happen then the events $\mathbf{q}[1], \ldots, \mathbf{q}[m]$ are conditionally independent of each other given knowledge of $t = l$ and the chain rule definition can be simplified as follows:

$$
\begin{aligned}
&P(\mathbf{q}[1], \ldots, \mathbf{q}[m] \mid t = l) \\
&\quad = P(\mathbf{q}[1] \mid t = l) \times P(\mathbf{q}[2] \mid t = l) \times \cdots \times P(\mathbf{q}[m] \mid t = l) \\
&\quad = \prod_{i=1}^{m} P(\mathbf{q}[i] \mid t = l)
\end{aligned}
$$

- Using this we can simplify the calculations in Bayes' Theorem, under the assumption of conditional independence between the descriptive features given the level l of the target feature:

$$P(t = l \mid \mathbf{q}[1], \ldots, \mathbf{q}[m]) = \frac{\left(\prod_{i=1}^{m} P(\mathbf{q}[i] \mid t = l) \right) \times P(t = l)}{P(\mathbf{q}[1], \ldots, \mathbf{q}[m])}$$

Withouth conditional independence

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

With conditional independence

$$P(X, Y, Z|W) = \underbrace{P(X|W)}_{Factor1} \times \underbrace{P(Y|W)}_{Factor2} \times \underbrace{P(Z|W)}_{Factor3} \times \underbrace{P(W)}_{Factor4}$$

- The joint probability distribution for the meningitis dataset.

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

- Assuming the descriptive features are conditionally independent of each other given MENINGITIS we only need to store four factors:

$$Factor_1 : < P(M) >$$
$$Factor_2 : < P(h|m), P(h|\neg m) >$$
$$Factor_3 : < P(f|m), P(f|\neg m) >$$
$$Factor_4 : < P(v|m), P(v|\neg m) >$$

$$P(H, F, V, M) = P(M) \times P(H|M) \times P(F|M) \times P(V|M)$$

ID	HEADACHE	FEVER	VOMITING	MENINGITIS
1	true	true	false	false
2	false	true	false	false
3	true	false	true	false
4	true	false	true	false
5	false	true	false	true
6	true	false	true	false
7	true	false	true	false
8	true	false	true	true
9	false	true	false	false
10	true	false	true	true

- Calculate the factors from the data.

$$Factor_1 : < P(M) >$$
$$Factor_2 : < P(h|m), P(h|\neg m) >$$
$$Factor_3 : < P(f|m), P(f|\neg m) >$$
$$Factor_4 : < P(v|m), P(v|\neg m) >$$

$$Factor_1 : < P(m) = 0.3 >$$
$$Factor_2 : < P(h|m) = 0.6666, P(h|\neg m) = 0.7413 >$$
$$Factor_3 : < P(f|m) = 0.3333, P(f|\neg m) = 0.4286 >$$
$$Factor_4 : < P(v|m) = 0.6666, P(v|\neg m) = 0.5714 >$$

$$Factor_1 : < P(m) = 0.3 >$$
$$Factor_2 : < P(h|m) = 0.6666, P(h|\neg m) = 0.7413 >$$
$$Factor_3 : < P(f|m) = 0.3333, P(f|\neg m) = 0.4286 >$$
$$Factor_4 : < P(v|m) = 0.6666, P(v|\neg m) = 0.5714 >$$

- Using the factors above calculate the probability of MENINGITIS=*'true'* for the following query.

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

$$P(m|h,f,\neg v) = \frac{P(h|m) \times P(f|m) \times P(\neg v|m) \times P(m)}{\sum_i P(h|M_i) \times P(f|M_i) \times P(\neg v|M_i) \times P(M_i)} =$$

$$\frac{0.6666 \times 0.3333 \times 0.3333 \times 0.3}{(0.6666 \times 0.3333 \times 0.3333 \times 0.3) + (0.7143 \times 0.4286 \times 0.4286 \times 0.7)} = 0.1948$$

$$Factor_1 : < P(m) = 0.3 >$$
$$Factor_2 : < P(h|m) = 0.6666, P(h|\neg m) = 0.7413 >$$
$$Factor_3 : < P(f|m) = 0.3333, P(f|\neg m) = 0.4286 >$$
$$Factor_4 : < P(v|m) = 0.6666, P(v|\neg m) = 0.5714 >$$

- Using the factors above calculate the probability of MENINGITIS=*'false'* for the same query.

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

$$P(\neg m|h,f,\neg v) = \frac{P(h|\neg m) \times P(f|\neg m) \times P(\neg v|\neg m) \times P(\neg m)}{\sum_i P(h|M_i) \times P(f|M_i) \times P(\neg v|M_i) \times P(M_i)} =$$

$$\frac{0.7143 \times 0.4286 \times 0.4286 \times 0.7}{(0.6666 \times 0.3333 \times 0.3333 \times 0.3) + (0.7143 \times 0.4286 \times 0.4286 \times 0.7)} = 0.8052$$

$$P(m|h,f,\neg v) = 0.1948$$

$$P(\neg m|h,f,\neg v) = 0.8052$$

- As before, the MAP prediction would be MENINGITIS = *'false'*
- The posterior probabilities are not as extreme!