Fundamentals of Machine Learning

Chapter 6: Probability-based Learning

Sections 6.1, 6.2, 6.3

Big Idea

Two Aces, One Queen

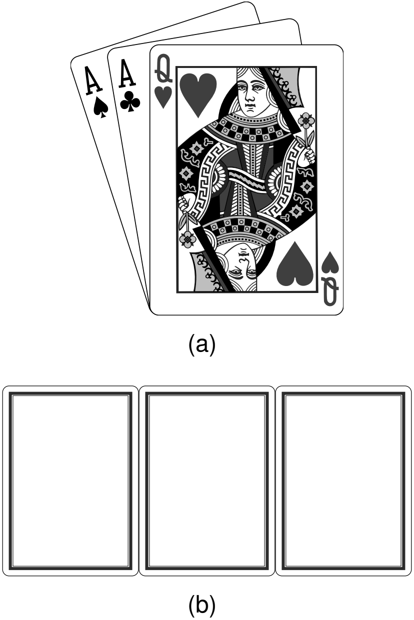
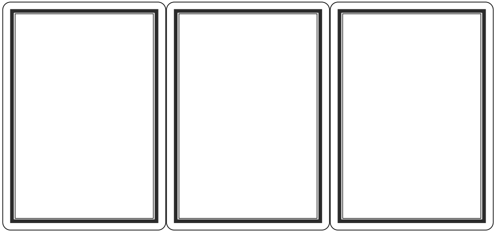
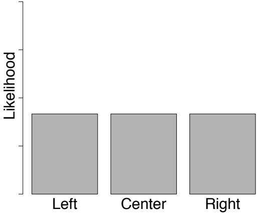


Figure: A game of *find the lady*

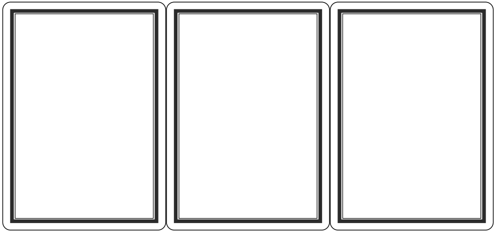


(a)



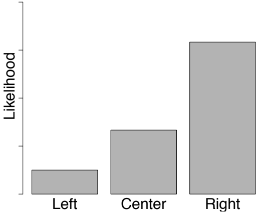
(b)

Figure: A game of *find the lady*: (a) the cards dealt face down on a table; and (b) the initial likelihoods of the queen ending up in each position.



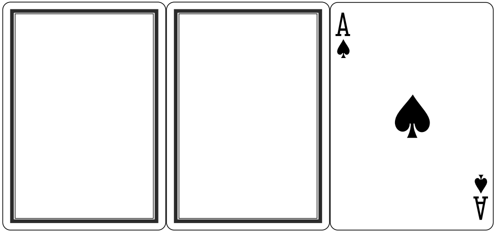
(a)

You watch the guy for 30 hands and notice that he has a tendency to place the Queen on the right:

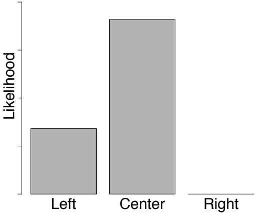


(b)

Figure: A game of *find the lady*: (a) the cards dealt face down on a table; and (b) a revised set of likelihoods for the position of the queen based on evidence collected.



(a)



(b)

Figure: A game of *find the lady*: (a) The set of cards after the wind blows over the one on the right; (b) the revised likelihoods for the position of the queen based on this new evidence.

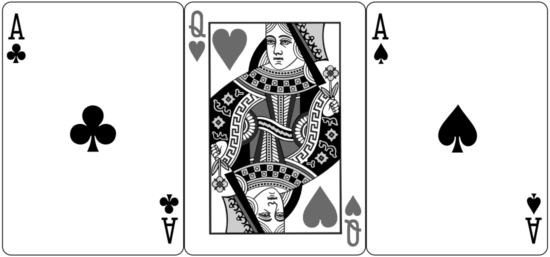


Figure: A game of *find the lady*: The final positions of the cards in the game.

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.

Fundamentals

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.

Joint probability distribution of events H, F, V, and M:

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

Examples of “Summing Out”:

P(not f) = sum of all cells where “not f” hold.

P(h and not f) = sum of all cells where “h and not f” hold.

- 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**.

Notation: P(X and Y) = P(XY)
Definition of conditional probability: P(X|Y) = P(XY)/P(Y)
Reorganising terms gives
P(XY) = P(X|Y) P(Y) = P(Y|X) P(X)
(The above implies that it doesn't matter how you name the events!)
Moving P(Y) to the RHS yields the Bayes' Theorem:

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?

Event "t": Test comes out positive

Theorem of Total Probability

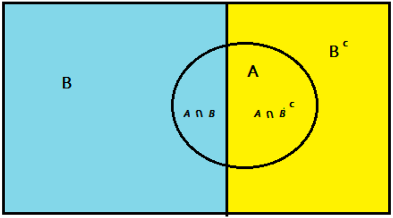
$$P(d|t) = \frac{P(t|d)P(d)}{P(t)}$$

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

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

Power of Bayes' Theorem: Even though the test is accurate 99% both ways, the probability that the patient has the disease conditioned upon the test coming out positive is just about 1% !!!
WHY: Because the disease is extremely rare to begin with.

Theorem of Total Probability



(Source: www.datasciencecentral.com)

Recall:
P(AB) = P(A|B) P(B)

"c" denotes complement, so B^c means "not B".
For event A, we have $A = (AB) \cup (AB^c)$ where U denotes the union operator.
Thus, $P(A) = P(AB) + P(AB^c)$

➔ $P(A) = P(A|B)\, P(B) + P(A|B^c)\, P(B^c)$

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 allows calculation of any joint distribution using only conditional probabilities:

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:

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

Simple Examples of Chain Rule:
Two Events: P(AB) = P(A) x P(B I A) = P(B) x P(A I B)
Three Events: P(ABC) = P(A) x P(B I A) x P(C I BA)

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:

$$P(h, \neg f, v \mid m) = \textcolor{red}{?}$$

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:

$$P(m|h, \neg f, v) = \frac{\begin{pmatrix} P(h|m) \times P(\neg f|h, m) \\ \times P(v|\neg f, h, m) \times P(m) \end{pmatrix}}{P(h, \neg f, v)}$$

$$= \frac{0.6666 \times 0.3}{0.6} = 0.3333$$

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

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

Apparently, for any events A and B, we have P(A | B) + P(not A | B) = 1:

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

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

- 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).

(MAP: Maximum A Posteriori)

Bayesian MAP Prediction Model

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

Bayesian MAP Prediction Model (without normalization)

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

Why we don't need normalisation when making a prediction:
For both "m" and "not m", we divide the probabilities by the same denominator, so ignoring the denominator will not change the outcome when we determine which one is bigger.
The denominator is needed only to make sure the probabilities add up to 1.

Let's do another prediction example using the Bayes' Theorem and the Chain Rule:

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

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

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

$$
\begin{aligned}
P(m \mid h, f, \neg v) &= \frac{\begin{pmatrix} P(h|m) \times P(f \mid h, m) \\ \quad \times P(\neg v \mid f, h, m) \times P(m) \end{pmatrix}}{P(h, f, \neg v)} \\
&= \frac{0.6666 \times 0 \times 0 \times 0.3}{0.1} = 0
\end{aligned}
$$

$$P(\neg m \mid h, f, \neg v) = \frac{\begin{pmatrix} P(h|\neg m) \times P(f \mid h, \neg m) \\ \times P(\neg v \mid f, h, \neg m) \times P(\neg m) \end{pmatrix}}{P(h, f, \neg v)}$$

$$= \frac{0.7143 \times 0.2 \times 1.0 \times 0.7}{0.1} = 1.0$$

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

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

- 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:

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

$$
\begin{aligned}
P(X|Y) &= \frac{P(X,Y)}{P(Y)} \\
P(X,Y) &= P(X|Y) \times P(Y) \\
&= P(Y|X) \times P(X)
\end{aligned}
$$

X and Y are dependent

$$
\begin{aligned}
P(X|Y) &= P(X) \\
P(X,Y) &= P(X) \times P(Y)
\end{aligned}
$$

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:

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

- 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:

$$
\begin{aligned}
Factor_1 &: < P(M) > \\
Factor_2 &: < P(h|m), P(h|\neg m) > \\
Factor_3 &: < P(f|m), P(f|\neg m) > \\
Factor_4 &: < P(v|m), P(v|\neg m) >
\end{aligned}
$$

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

WARNING:
It always holds that P(A I B) + P(not A I B) = 1.
However, P(A I B) + P(A I not B) doesn't have to add up to 1!
In fact, this sum can be bigger than 1.

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.

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

These four factors are ALL you need to make a prediction for ANY combination of descriptive feature values!

$$Factor_1 : < P(m) = 0.3 >$$
$$Factor_2 : < P(h|m) = 0.6666, P(h|\neg m) = 0.7143 >$$
$$Factor_3 : < P(f|m) = 0.3333, P(f|\neg m) = 0.4286 >$$
$$Factor_4 : < P(v|m) = 0.6666, P(v|\neg m) = 0.5714 >$$

- 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!

In this particular case, assuming conditional independence (given the patient does have meningitis) helps us with avoiding the overfitting problem that we faced when using the Chain Rule while using the Bayes' Theorem.

Standard Approach: The Naive Bayes' Classifier

Naive Bayes' Classifier

$$\mathbb{M}(\mathbf{q}) = \underset{l \in levels(t)}{\operatorname{argmax}} \left(\prod_{i=1}^{m} P(\mathbf{q}[i] \mid t = l) \right) \times P(t = l)$$

Naive Bayes' is simple to train!

1. calculate the priors for each of the target levels
2. calculate the conditional probabilities for each feature given each target level.

Example: Loan application fraud detection

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(\mathrm{CH} = \textit{'none'} \mid fr)$	$=$	0.1666	$P(\mathrm{CH} = \textit{'none'} \mid \neg fr)$	$=$	0
$P(\mathrm{CH} = \textit{'paid'} \mid fr)$	$=$	0.1666	$P(\mathrm{CH} = \textit{'paid'} \mid \neg fr)$	$=$	0.2857
$P(\mathrm{CH} = \textit{'current'} \mid fr)$	$=$	0.5	$P(\mathrm{CH} = \textit{'current'} \mid \neg fr)$	$=$	0.2857
$P(\mathrm{CH} = \textit{'arrears'} \mid fr)$	$=$	0.1666	$P(\mathrm{CH} = \textit{'arrears'} \mid \neg fr)$	$=$	0.4286
$P(\mathrm{GC} = \textit{'none'} \mid fr)$	$=$	0.8334	$P(\mathrm{GC} = \textit{'none'} \mid \neg fr)$	$=$	0.8571
$P(\mathrm{GC} = \textit{'guarantor'} \mid fr)$	$=$	0.1666	$P(\mathrm{GC} = \textit{'guarantor'} \mid \neg fr)$	$=$	0
$P(\mathrm{GC} = \textit{'coapplicant'} \mid fr)$	$=$	0	$P(\mathrm{GC} = \textit{'coapplicant'} \mid \neg fr)$	$=$	0.1429
$P(\mathrm{ACC} = \textit{'own'} \mid fr)$	$=$	0.6666	$P(\mathrm{ACC} = \textit{'own'} \mid \neg fr)$	$=$	0.7857
$P(\mathrm{ACC} = \textit{'rent'} \mid fr)$	$=$	0.3333	$P(\mathrm{ACC} = \textit{'rent'} \mid \neg fr)$	$=$	0.1429
$P(\mathrm{ACC} = \textit{'free'} \mid fr)$	$=$	0	$P(\mathrm{ACC} = \textit{'free'} \mid \neg fr)$	$=$	0.0714

Table: The probabilities needed by a Naive Bayes prediction model calculated from the dataset. Notation key: FR=FRAUDULENT, CH=CREDIT HISTORY, GC = GUARANTOR/COAPPLICANT, ACC = ACCOMODATION, T=*'true'*, F=*'false'*.

$$
\begin{array}{rclrcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(\text{CH} = \textit{'none'} \mid fr) & = & 0.1666 & P(\text{CH} = \textit{'none'} \mid \neg fr) & = & 0 \\
P(\text{CH} = \textit{'paid'} \mid fr) & = & 0.1666 & P(\text{CH} = \textit{'paid'} \mid \neg fr) & = & 0.2857 \\
P(\text{CH} = \textit{'current'} \mid fr) & = & 0.5 & P(\text{CH} = \textit{'current'} \mid \neg fr) & = & 0.2857 \\
P(\text{CH} = \textit{'arrears'} \mid fr) & = & 0.1666 & P(\text{CH} = \textit{'arrears'} \mid \neg fr) & = & 0.4286 \\
P(\text{GC} = \textit{'none'} \mid fr) & = & 0.8334 & P(\text{GC} = \textit{'none'} \mid \neg fr) & = & 0.8571 \\
P(\text{GC} = \textit{'guarantor'} \mid fr) & = & 0.1666 & P(\text{GC} = \textit{'guarantor'} \mid \neg fr) & = & 0 \\
P(\text{GC} = \textit{'coapplicant'} \mid fr) & = & 0 & P(\text{GC} = \textit{'coapplicant'} \mid \neg fr) & = & 0.1429 \\
P(\text{ACC} = \textit{'own'} \mid fr) & = & 0.6666 & P(\text{ACC} = \textit{'own'} \mid \neg fr) & = & 0.7857 \\
P(\text{ACC} = \textit{'rent'} \mid fr) & = & 0.3333 & P(\text{ACC} = \textit{'rent'} \mid \neg fr) & = & 0.1429 \\
P(\text{ACC} = \textit{'free'} \mid fr) & = & 0 & P(\text{ACC} = \textit{'free'} \mid \neg fr) & = & 0.0714
\end{array}
$$

Credit History	Guarantor/CoApplicant	Accomodation	Fraudulent
paid	none	rent	?

$$
\begin{aligned}
P(fr) &= 0.3 & P(\neg fr) &= 0.7 \\
P(\text{CH} = \textit{'paid'} \mid fr) &= 0.1666 & P(\text{CH} = \textit{'paid'} \mid \neg fr) &= 0.2857 \\
P(\text{GC} = \textit{'none'} \mid fr) &= 0.8334 & P(\text{GC} = \textit{'none'} \mid \neg fr) &= 0.8571 \\
P(\text{ACC} = \textit{'rent'} \mid fr) &= 0.3333 & P(\text{ACC} = \textit{'rent'} \mid \neg fr) &= 0.1429
\end{aligned}
$$

$$
\left(\prod_{k=1}^{m} P\left(\mathbf{q}\left[k\right] \mid fr\right) \right) \times P\left(fr\right) = 0.0139
$$

$$
\left(\prod_{k=1}^{m} P\left(\mathbf{q}\left[k\right] \mid \neg fr\right) \right) \times P(\neg fr) = 0.0245
$$

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMODATION	FRAUDULENT
paid	none	rent	?

$$P(fr) = 0.3 \qquad P(\neg fr) = 0.7$$

$$P(\text{CH} = \textit{'paid'} \mid fr) = 0.1666 \qquad P(\text{CH} = \textit{'paid'} \mid \neg fr) = 0.2857$$

$$P(\text{GC} = \textit{'none'} \mid fr) = 0.8334 \qquad P(\text{GC} = \textit{'none'} \mid \neg fr) = 0.8571$$

$$P(\text{ACC} = \textit{'rent'} \mid fr) = 0.3333 \qquad P(\text{ACC} = \textit{'rent'} \mid \neg fr) = 0.1429$$

$$\left(\prod_{k=1}^{m} P\left(\mathbf{q}\left[k\right] \mid fr\right)\right) \times P\left(fr\right) = 0.0139$$

$$\left(\prod_{k=1}^{m} P\left(\mathbf{q}\left[k\right] \mid \neg fr\right)\right) \times P(\neg fr) = 0.0245$$

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMODATION	FRAUDULENT
paid	none	rent	*'false'*

The model is generalizing beyond the dataset!

ID	Credit History	Guarantor/ CoApplicant	Accommodation	Fraud
1	current	none	own	true
2	paid	none	own	false
3	paid	none	own	false
4	paid	guarantor	rent	true
5	arrears	none	own	false
6	arrears	none	own	true
7	current	none	own	false
8	arrears	none	own	false
9	current	none	rent	false
10	none	none	own	true
11	current	coapplicant	own	false
12	current	none	own	true
13	current	none	rent	true
14	paid	none	own	false
15	arrears	none	own	false
16	current	none	own	false
17	arrears	coapplicant	rent	false
18	arrears	none	free	false
19	arrears	none	own	false
20	paid	none	own	false

Credit History	Guarantor/CoApplicant	Accommodation	Fraudulent
paid	none	rent	*'false'*

Summary

$$P(t|\mathbf{d}) = \frac{P(\mathbf{d}|t) \times P(t)}{P(\mathbf{d})} \tag{2}$$

- A Naive Bayes' classifier naively assumes that each of the descriptive features in a domain is conditionally independent of all of the other descriptive features, given the state of the target feature.
- This assumption, although often wrong, enables the Naive Bayes' model to maximally factorise the representation that it uses of the domain.
- Surprisingly, given the naivety and strength of the assumption it depends upon, a Naive Bayes' model often performs reasonably well.

A downside of Naive Bayes is that it ignores interactions between descriptive features (when conditioned upon a given target feature level).