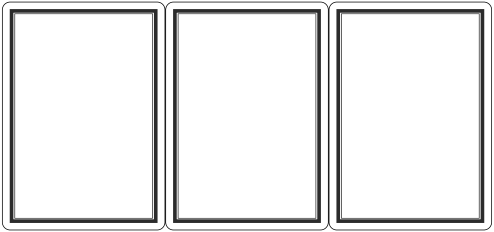
Probability-based Learning
Sections 6.1, 6.2, 6.3

John D. Kelleher and Brian Mac Namee and Aoife D'Arcy

Big Idea

(a)



(b)

Figure: A game of *find the lady*

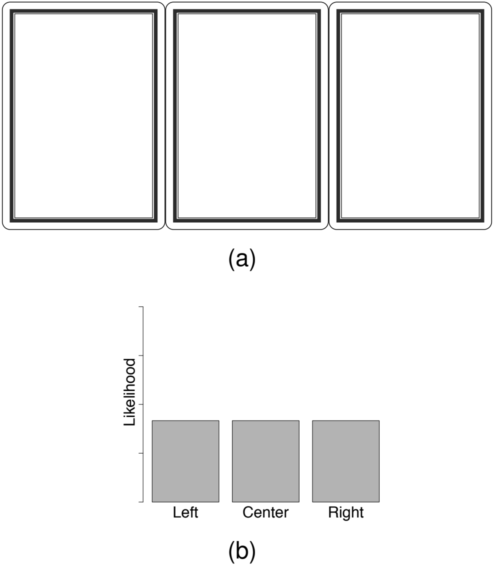
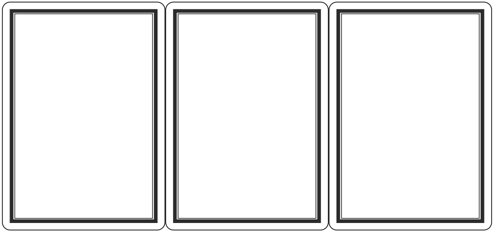
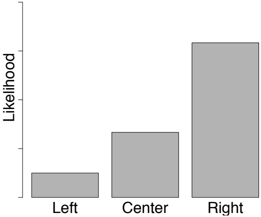


Figure: A game of *find the lady*: (a) the cards dealt face down on a table; and (b) the initial likelihoods of the queen ending up in each position.

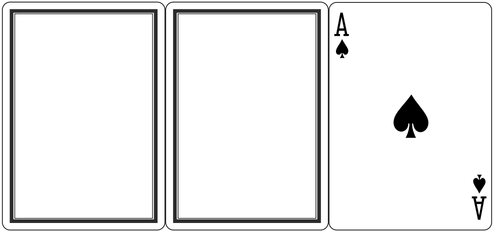


(a)

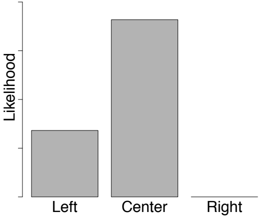


(b)

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.

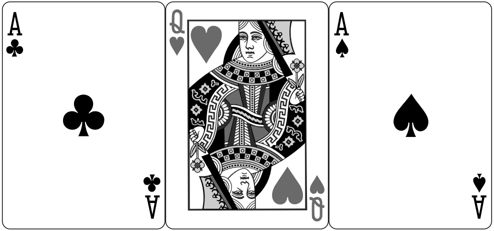


Figure: A game of *find the lady*: The final positions of the cards in the game.

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.

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

- 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?

Bayes' Theorem

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

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

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

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

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

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

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

Bayesian Prediction

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:

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

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

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

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

$$P(m|h,\neg f,v) = 0.3333$$
$$P(\neg m|h,\neg f,v) = 0.6667$$

- 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

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

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

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

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

Bayesian Prediction

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

$$
\begin{aligned}
P(\neg m \mid h, f, \neg v) &= \frac{\begin{pmatrix} P(h|\neg m) \times P(f \mid h, \neg m) \\ \times P(\neg v \mid f, h, \neg m) \times P(\neg m) \end{pmatrix}}{P(h, f, \neg v)} \\
&= \frac{0.7143 \times 0.2 \times 1.0 \times 0.7}{0.1} = 1.0
\end{aligned}
$$

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

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

- 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:

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

- 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:

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

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

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!

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.

Table: A dataset from a loan application fraud detection domain.

ID	Credit History	Guarantor/ CoApplicant	Accomodation	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
$P(\text{CH} = \textit{'none'} \mid fr)$	$=$	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

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'*.

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

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	?

A Worked Example

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

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

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

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.

Probability-based Learning
Sections 6.4, 6.5

John D. Kelleher and Brian Mac Namee and Aoife D'Arcy

Smoothing

$P(fr)$	$=$	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

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMMODATION	FRAUDULENT
paid	guarantor	free	?

$$
\begin{aligned}
P(fr) &= 0.3 & P(\neg fr) &= 0.7 \\
P(CH = paid \mid fr) &= 0.1666 & P(CH = paid \mid \neg fr) &= 0.2857 \\
P(GC = guarantor \mid fr) &= 0.1666 & P(GC = guarantor \mid \neg fr) &= 0 \\
P(ACC = free \mid fr) &= 0 & P(ACC = free \mid \neg fr) &= 0.0714
\end{aligned}
$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.0$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.0$$

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMMODATION	FRAUDULENT
paid	guarantor	free	?

- The standard way to avoid this issue is to use **smoothing**.
- Smoothing takes some of the probability from the events with lots of the probability share and gives it to the other probabilities in the set.

- There are several different ways to smooth probabilities, we will use **Laplacian smoothing**.

Laplacian Smoothing (conditional probabilities)

$$P(f = v|t) = \frac{count(f = v|t) + k}{count(f|t) + (k \times |Domain(f)|)}$$

Raw	$P(GC = none \mid \neg fr)$	$=$	0.8571
Probabilities	$P(GC = guarantor \mid \neg fr)$	$=$	0
	$P(GC = coapplicant \mid \neg fr)$	$=$	0.1429
Smoothing	k	$=$	3
Parameters	$count(GC \mid \neg fr)$	$=$	14
	$count(GC = none \mid \neg fr)$	$=$	12
	$count(GC = guarantor \mid \neg fr)$	$=$	0
	$count(GC = coapplicant \mid \neg fr)$	$=$	2
	$\lvert Domain(GC) \rvert$	$=$	3
Smoothed	$P(GC = none \mid \neg fr) = \frac{12+3}{14+(3\times3)}$	$=$	0.6522
Probabilities	$P(GC = guarantor \mid \neg fr) = \frac{0+3}{14+(3\times3)}$	$=$	0.1304
	$P(GC = coapplicant \mid \neg fr) = \frac{2+3}{14+(3\times3)}$	$=$	0.2174

Table: Smoothing the posterior probabilities for the GUARANTOR/COAPPLICANT feature conditioned on FRAUDULENT being False.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = Free \mid fr)$	$=$	0.2	$P(ACC = Free \mid \neg fr)$	$=$	0.1739

Table: The Laplacian smoothed, with $k = 3$, probabilities needed by a Naive Bayes prediction model calculated from the fraud detection dataset. Notation key: FR=FRAUDULENT, CH=CREDIT HISTORY, GC = GUARANTOR/COAPPLICANT, ACC = ACCOMODATION, T=*'True'*, F=*'False'*.

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMMODATION	FRAUDULENT
paid	guarantor	free	?

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(ACC = Free \mid fr)$	$=$	0.2	$P(ACC = Free \mid \neg fr)$	$=$	0.1739

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[m] \mid fr)\right) \times P(fr) = 0.0036$$

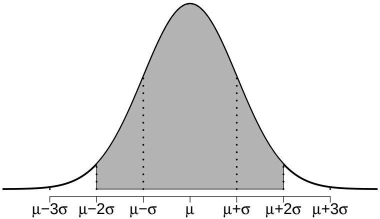
$$\left(\prod_{k=1}^{m} P(\mathbf{q}[m] \mid \neg fr)\right) \times P(\neg fr) = 0.0043$$

Table: The relevant smoothed probabilities, from Table 2 [9], needed by the Naive Bayes prediction model in order to classify the query from the previous slide and the calculation of the scores for each candidate classification.

Continuous Features: Probability Density Functions

- A **probability density function** (PDF) represents the probability distribution of a continuous feature using a mathematical function, such as the normal distribution.

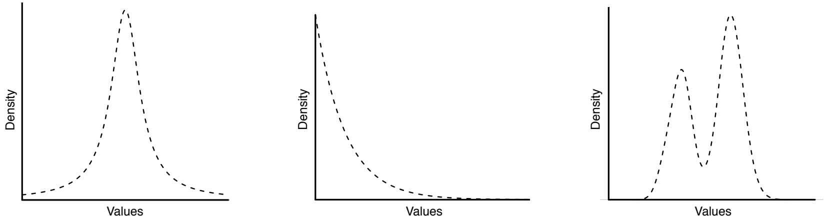
$$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



- A PDF defines a density curve and the shape of the of the curve is determined by:
  - the statistical distribution that is used to define the PDF
  - the values of the statistical distribution parameters

Table: Definitions of some standard probability distributions.

Normal $x \in \mathbb{R}$ $\mu \in \mathbb{R}$ $\sigma \in \mathbb{R}_{>0}$	$N(x, \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$
Student-t $x \in \mathbb{R}$ $\phi \in \mathbb{R}$ $\rho \in \mathbb{R}_{>0}$ $\kappa \in \mathbb{R}_{>0}$ $z = \frac{x-\phi}{\rho}$	$\tau(x, \phi, \rho, \kappa) = \frac{\Gamma(\frac{\kappa+1}{2})}{\Gamma(\frac{\kappa}{2}) \times \sqrt{\pi\kappa} \times \rho} \times \left(1 + \left(\frac{1}{\kappa} \times z^2\right)\right)^{-\frac{\kappa+1}{2}}$
Exponential $x \in \mathbb{R}$ $\lambda \in \mathbb{R}_{>0}$	$E(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$
Mixture of n Gaussians $x \in \mathbb{R}$ $\{\mu_1, \ldots, \mu_n \| \mu_i \in \mathbb{R}\}$ $\{\sigma_1, \ldots, \sigma_n \| \sigma_i \in \mathbb{R}_{>0}\}$ $\{\omega_1, \ldots, \omega_n \| \omega_i \in \mathbb{R}_{>0}\}$ $\sum_{i=1}^{n} \omega_i = 0$	$N(x, \mu_1, \sigma_1, \omega_1, \ldots, \mu_n, \sigma_n, \omega_n) = \sum_{i=1}^{n} \frac{\omega_i}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$



(a) Normal/Student-t (b) Exponential (c) Mixture of Gaussians

Figure: Plots of some well known probability distributions.

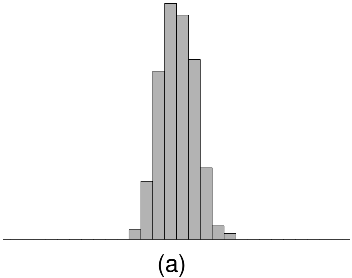


Figure: Histograms of two unimodal datasets: (a) the distribution has light tails; (b) the distribution has fat tails.

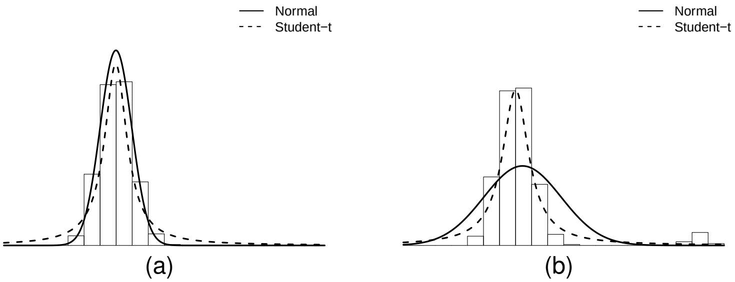


Figure: Illustration of the robustness of the student-*t* distribution to outliers: (a) a density histogram of a unimodal dataset overlaid with the density curves of a normal and a student-*t* distribution that have been fitted to the data; (b) a density histogram of the same dataset with outliers added, overlaid with the density curves of a normal and a student-*t* distribution that have been fitted to the data. The student-*t* distribution is less affected by the introduction of outliers. (This figure is inspired by Figure 2.16 in (Bishop, 2006).)

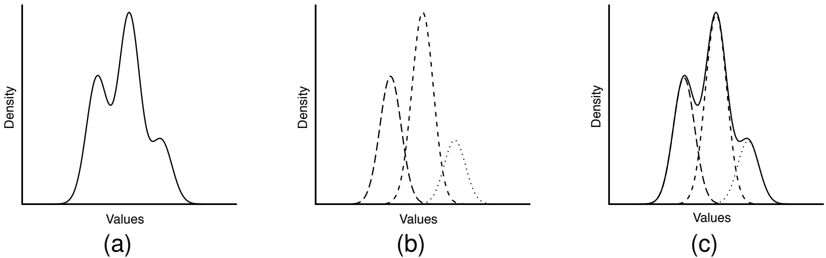


Figure: Illustration of how a mixture of Gaussians model is composed of a number of normal distributions. The curve plotted using a solid line is the mixture of Gaussians density curve, created using an appropriately weighted summation of the three normal curves, plotted using dashed and dotted lines.

- A PDF is an abstraction over a density histogram and consequently PDF represents probabilities in terms of area under the curve.
- To use a PDF to calculate a probability we need to think in terms of the area under an interval of the PDF curve.
- We can calculate the area under a PDF by looking this up in a probability table or to use integration to calculate the area under the curve within the bounds of the interval.

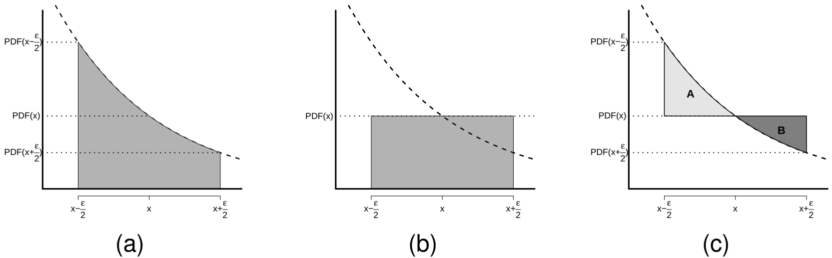


Figure: (a) The area under a density curve between the limits $x - \frac{\epsilon}{2}$ and $x + \frac{\epsilon}{2}$; (b) the approximation of this area computed by $PDF(x) \times \epsilon$; and (c) the error in the approximation is equal to the difference between area A, the area under the curve omitted from the approximation, and area B, the area above the curve erroneously included in the approximation. Both of these areas will get smaller as the width of the interval gets smaller, resulting in a smaller error in the approximation.

- There is no hard and fast rule for deciding on **interval size** - instead, this decision is done on a case by case basis and is dependent on the precision required in answering a question.
- To illustrate how PDFs can be used in Naive Bayes models we will extend our loan application fraud detection query to have an ACCOUNT BALANCE feature

Table: The dataset from the loan application fraud detection domain with a new continuous descriptive features added: ACCOUNT BALANCE

ID	CREDIT HISTORY	GUARANTOR/ COAPPLICANT	ACCOMMODATION	ACCOUNT BALANCE	FRAUD
1	current	none	own	56.75	true
2	current	none	own	1,800.11	false
3	current	none	own	1,341.03	false
4	paid	guarantor	rent	749.50	true
5	arrears	none	own	1,150.00	false
6	arrears	none	own	928.30	true
7	current	none	own	250.90	false
8	arrears	none	own	806.15	false
9	current	none	rent	1,209.02	false
10	none	none	own	405.72	true
11	current	coapplicant	own	550.00	false
12	current	none	free	223.89	true
13	current	none	rent	103.23	true
14	paid	none	own	758.22	false
15	arrears	none	own	430.79	false
16	current	none	own	675.11	false
17	arrears	coapplicant	rent	1,657.20	false
18	arrears	none	free	1,405.18	false
19	arrears	none	own	760.51	false
20	current	none	own	985.41	false

- We need to define two PDFs for the new ACCOUNT BALANCE (AB) feature with each PDF conditioned on a different value in the domain or the target:
  - $P(AB = X|fr) = PDF_1(AB = X|fr)$
  - $P(AB = X|\neg fr) = PDF_2(AB = X|\neg fr)$
- Note that these two PDFs do not have to be defined using the same statistical distribution.

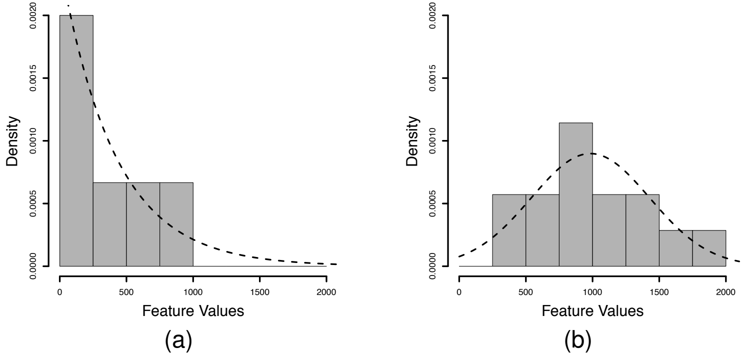


Figure: Histograms, using a bin size of 250 units, and density curves for the ACCOUNT BALANCE feature: (a) the fraudulent instances overlaid with a fitted exponential distribution; (b) the non-fraudulent instances overlaid with a fitted normal distribution.

- From the shape of these histograms it appears that
  - the distribution of values taken by the ACCOUNT BALANCE feature in the set of instances where the target feature FRAUDULENT=*'True'* follows an exponential distribution
  - the distributions of values taken by the ACCOUNT BALANCE feature in the set of instances where the target feature FRAUDULENT=*'False'* is similar to a normal distribution.
- Once we have selected the distributions the next step is to fit the distributions to the data.

- To fit the exponential distribution we simply compute the sample mean, $\bar{x}$, of the ACCOUNT BALANCE feature in the set of instances where FRAUDULENT=*'True'* and set the λ parameter equal to one divided by $\bar{x}$.
- To fit the normal distribution to the set of instances where FRAUDULENT=*'False'* we simply compute the sample mean and sample standard deviation, s, for the ACCOUNT BALANCE feature for this set of instances and set the parameters of the normal distribution to these values.

Table: Partitioning the dataset based on the value of the target feature and fitting the parameters of a statistical distribution to model the ACCOUNT BALANCE feature in each partition.

ID	...	ACCOUNT BALANCE	FRAUD
1		56.75	true
4		749.50	true
6		928.30	true
10	...	405.72	true
12		223.89	true
13		103.23	true
$\overline{\text{AB}}$		411.22	
$\lambda = {}^{1}/_{\overline{\text{AB}}}$		0.0024	

ID	...	ACCOUNT BALANCE	FRAUD
2		1 800.11	false
3		1 341.03	false
5		1 150.00	false
7		250.90	false
8		806.15	false
9		1 209.02	false
11		550.00	false
14		758.22	false
15		430.79	false
16		675.11	false
17		1 657.20	false
18		1 405.18	false
19		760.51	false
20		985.41	false
$\overline{\text{AB}}$		984.26	
$sd(\text{AB})$		460.94	

Table: The Laplace smoothed (with $k = 3$) probabilities needed by a naive Bayes prediction model calculated from the dataset in Table 5 [23], extended to include the conditional probabilities for the new ACCOUNT BALANCE feature, which are defined in terms of PDFs.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = x \mid fr)$			$P(AB = x \mid \neg fr)$		
	$\approx$	$E\begin{pmatrix} x, \\ \lambda = 0.0024 \end{pmatrix}$		$\approx$	$N\begin{pmatrix} x, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix}$

Table: A query loan application from the fraud detection domain.

Credit History	Guarantor/ CoApplicant	Accomodation	Account Balance	Fraudulent
paid	guarantor	free	759.07	?

Table: The probabilities, from Table 7 [29], needed by the naive Bayes prediction model to make a prediction for the query $\langle \text{CH} = \textit{'paid'}, \text{GC} = \textit{'guarantor'}, \text{ACC} = \textit{'free'}, \text{AB} = 759.07 \rangle$ and the calculation of the scores for each candidate prediction.

$$
\begin{array}{rclrcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = paid | fr) & = & 0.2222 & P(CH = paid | \neg fr) & = & 0.2692 \\
P(GC = guarantor | fr) & = & 0.2667 & P(GC = guarantor | \neg fr) & = & 0.1304 \\
P(ACC = free | fr) & = & 0.2 & P(ACC = free | \neg fr) & = & 0.1739 \\
P(AB = 759.07 | fr) & & & P(AB = 759.07 | \neg fr) & & \\
\approx E\left(\begin{array}{c} 759.07, \\ \lambda = 0.0024 \end{array}\right) & = & 0.00039 & \approx N\left(\begin{array}{c} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{array}\right) & = & 0.00077
\end{array}
$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] | fr)\right) \times P(fr) = 0.0000014$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] | \neg fr)\right) \times P(\neg fr) = 0.0000033$$

Continuous Features: Binning

- In Section 3.6.2 we explained two of the best known binning techniques **equal-width** and **equal-frequency**.
- We can use these techniques to *bin* continuous features into categorical features
- In general we recommend **equal-frequency binning**.

Table: The dataset from a loan application fraud detection domain with a second continuous descriptive feature added: LOAN AMOUNT

ID	CREDIT HISTORY	GUARANTOR/ COAPPLICANT	ACCOMMODATION	ACCOUNT BALANCE	LOAN AMOUNT	FRAUD
1	current	none	own	56.75	900	true
2	current	none	own	1 800.11	150 000	false
3	current	none	own	1 341.03	48 000	false
4	paid	guarantor	rent	749.50	10 000	true
5	arrears	none	own	1 150.00	32 000	false
6	arrears	none	own	928.30	250 000	true
7	current	none	own	250.90	25 000	false
8	arrears	none	own	806.15	18 500	false
9	current	none	rent	1 209.02	20 000	false
10	none	none	own	405.72	9 500	true
11	current	coapplicant	own	550.00	16 750	false
12	current	none	free	223.89	9 850	true
13	current	none	rent	103.23	95 500	true
14	paid	none	own	758.22	65 000	false
15	arrears	none	own	430.79	500	false
16	current	none	own	675.11	16 000	false
17	arrears	coapplicant	rent	1 657.20	15 450	false
18	arrears	none	free	1 405.18	50 000	false
19	arrears	none	own	760.51	500	false
20	current	none	own	985.41	35 000	false

Table: The LOAN AMOUNT continuous feature discretized into 4 equal-frequency bins.

ID	LOAN AMOUNT	BINNED LOAN AMOUNT	FRAUD	ID	LOAN AMOUNT	BINNED LOAN AMOUNT	FRAUD
15	500	bin1	false	9	20,000	bin3	false
19	500	bin1	false	7	25,000	bin3	false
1	900	bin1	true	5	32,000	bin3	false
10	9,500	bin1	true	20	35,000	bin3	false
12	9,850	bin1	true	3	48,000	bin3	false
4	10,000	bin2	true	18	50,000	bin4	false
17	15,450	bin2	false	14	65,000	bin4	false
16	16,000	bin2	false	13	95,500	bin4	true
11	16,750	bin2	false	2	150,000	bin4	false
8	18,500	bin2	false	6	250,000	bin4	true

- Once we have discretized the data we need to record the raw continuous feature threshold between the bins so that we can use these for query feature values.

Table: The thresholds used to discretize the LOAN AMOUNT feature in queries.

Bin Thresholds		
	Bin1	$\leq 9,925$
$9,925 <$	Bin2	$\leq 19,250$
$19,225 <$	Bin3	$\leq 49,000$
$49,000 <$	Bin4	

Table: The Laplace smoothed (with $k = 3$) probabilities needed by a naive Bayes prediction model calculated from the fraud detection dataset. Notation key: FR = FRAUD, CH = CREDIT HISTORY, AB = ACCOUNT BALANCE, GC = GUARANTOR/COAPPLICANT, ACC = ACCOMMODATION, BLA = BINNED LOAN AMOUNT.

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(CH = none \mid fr)$	$=$	0.2222	$P(CH = none \mid \neg fr)$	$=$	0.1154
$P(CH = paid \mid fr)$	$=$	0.2222	$P(CH = paid \mid \neg fr)$	$=$	0.2692
$P(CH = current \mid fr)$	$=$	0.3333	$P(CH = current \mid \neg fr)$	$=$	0.2692
$P(CH = arrears \mid fr)$	$=$	0.2222	$P(CH = arrears \mid \neg fr)$	$=$	0.3462
$P(GC = none \mid fr)$	$=$	0.5333	$P(GC = none \mid \neg fr)$	$=$	0.6522
$P(GC = guarantor \mid fr)$	$=$	0.2667	$P(GC = guarantor \mid \neg fr)$	$=$	0.1304
$P(GC = coapplicant \mid fr)$	$=$	0.2	$P(GC = coapplicant \mid \neg fr)$	$=$	0.2174
$P(ACC = own \mid fr)$	$=$	0.4667	$P(ACC = own \mid \neg fr)$	$=$	0.6087
$P(ACC = rent \mid fr)$	$=$	0.3333	$P(ACC = rent \mid \neg fr)$	$=$	0.2174
$P(ACC = free \mid fr)$	$=$	0.2	$P(ACC = free \mid \neg fr)$	$=$	0.1739
$P(AB = x \mid fr)$ $\approx E\left(x, \lambda = 0.0024\right)$			$P(AB = x \mid \neg fr)$ $\approx N\left(x, \mu = 984.26, \sigma = 460.94\right)$		
$P(BLA = bin1 \mid fr)$	$=$	0.3333	$P(BLA = bin1 \mid \neg fr)$	$=$	0.1923
$P(BLA = bin2 \mid fr)$	$=$	0.2222	$P(BLA = bin2 \mid \neg fr)$	$=$	0.2692
$P(BLA = bin3 \mid fr)$	$=$	0.1667	$P(BLA = bin3 \mid \neg fr)$	$=$	0.3077
$P(BLA = bin4 \mid fr)$	$=$	0.2778	$P(BLA = bin4 \mid \neg fr)$	$=$	0.2308

Table: A query loan application from the fraud detection domain.

Credit History	Guarantor/ CoApplicant	Accomodation	Account Balance	Loan Amount	Fraudulent
paid	guarantor	free	759.07	8,000	?

Table: The relevant smoothed probabilities, from Table 13 [37], needed by the naive Bayes model to make a prediction for the query $\langle \text{CH} = \textit{'paid'}, \text{GC} = \textit{'guarantor'}, \text{ACC} = \textit{'free'}, \text{AB} = 759.07, \text{LA} = 8\,000 \rangle$ and the calculation of the scores for each candidate prediction.

$$
\begin{array}{rclrcl}
P(fr) & = & 0.3 & P(\neg fr) & = & 0.7 \\
P(CH = paid | fr) & = & 0.2222 & P(CH = paid | \neg fr) & = & 0.2692 \\
P(GC = guarantor | fr) & = & 0.2667 & P(GC = guarantor | \neg fr) & = & 0.1304 \\
P(ACC = free | fr) & = & 0.2 & P(ACC = free | \neg fr) & = & 0.1739 \\
P(AB = 759.07 | fr) & & & P(AB = 759.07 | \neg fr) & & \\
\approx E\left(\begin{array}{c} 759.07, \\ \lambda = 0.0024 \end{array} \right) & = & 0.00039 & \approx N\left(\begin{array}{c} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{array} \right) & = & 0.00077 \\
P(BLA = bin1 | fr) & = & 0.3333 & P(BLA = bin1 | \neg fr) & = & 0.1923
\end{array}
$$

$$\left(\prod_{k=1}^{m} P(\mathbf{q}[k] \mid fr)\right) \times P(fr) = 0.000000462$$

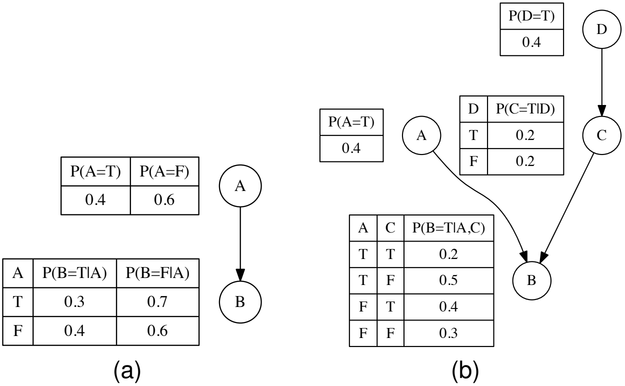
$$\left(\prod_{k=1}^{n} P(\mathbf{q}[k] \mid \neg fr)\right) \times P(\neg fr) = 0.000000633$$

Bayesian Networks

- **Bayesian networks** use a graph-based representation to encode the structural relationships—such as direct influence and conditional independence—between subsets of features in a domain.
- Consequently, a Bayesian network representation is generally more compact than a full joint distribution, yet is not forced to assert global conditional independence between all descriptive features.

A Bayesian Network is a directed acyclical graph that is composed of thee basic elements:

- nodes
- edges
- conditional probability tables (CPT)



(a)

P(A=T)	P(A=F)
0.4	0.6

A	P(B=T|A)	P(B=F|A)
T	0.3	0.7
F	0.4	0.6

(b)

P(D=T)
0.4

P(A=T)
0.4

D	P(C=T|D)
T	0.2
F	0.2

A	C	P(B=T|A,C)
T	T	0.2
T	F	0.5
F	T	0.4
F	F	0.3

Figure: (a) A Bayesian network for a domain consisting of two binary features. The structure of the network states that the value of feature A directly influences the value of feature B. (b) A Bayesian network consisting of 4 binary features with a path containing 3 generations of nodes: D, C, and B.

- In probability terms the directed edge from A to B in Figure (a) on the previous slide states that:

$$P(A, B) = P(B|A) \times P(A) \tag{1}$$

- For example, the probability of the event a and $\neg b$ is

$$P(a, \neg b) = P(\neg b|a) \times P(a) = 0.7 \times 0.4 = 0.28$$

- Equation (1)[44] can be generalized to the statement that for any network with N nodes, the probability of an event $x_1, \ldots, x_n$, can be computed using the following formula:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(x_i)) \tag{2}$$

- For example, using the more complex Bayesian network in figure (b) above, we can calculate the probability of the joint event $P(a, \neg b, \neg c, d)$ as follows:

$$
\begin{aligned}
P(a, \neg b, \neg c, d) &= P(\neg b | a, \neg c) \times P(\neg c | d) \times P(a) \times P(d) \\
&= 0.5 \times 0.8 \times 0.4 \times 0.4 = 0.064
\end{aligned}
$$

- We can uses Bayes' Theorem to invert the dependencies between nodes in a network.
- Returning to the simpler network in figure (a) above we can calculate $P(a|\neg b)$ as follows:

$$\begin{aligned}
P(a|\neg b) &= \frac{P(\neg b|a) \times P(a)}{P(\neg b)} = \frac{P(\neg b|a) \times P(a)}{\sum_i P(\neg b|A_i)} \\
&= \frac{P(\neg b|a) \times P(a)}{(P(\neg b|a) \times P(a)) + (P(\neg b|\neg a) \times P(\neg a))} \\
&= \frac{0.7 \times 0.4}{(0.7 \times 0.4) + (0.6 \times 0.6)} = 0.4375
\end{aligned}$$

- For conditional independence we need to take into account not only the parents of a node by also the state of its children and their parents.
- The set of nodes in a graph that make a node independent of the rest of the graph are known as the **Markov blanket** of a node.

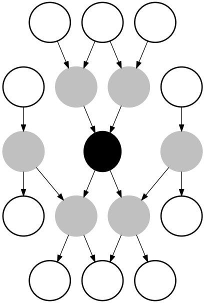


Figure: A depiction of the Markov blanket of a node. The gray nodes define the Markov blanket of the black node. The black node is conditionally independent of the white nodes given the state of the gray nodes.

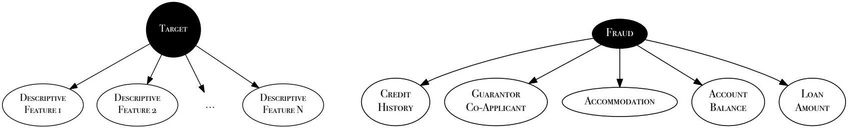
- The conditional independence of a node x_i in a graph with n nodes is defines as:

$$P(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) = P(x_i|Parents(x_i)) \prod_{j \in Children(x_i)} P(x_j|Parents(x_j)) \qquad (3)$$

- Applying the equation of the preceding slide to the network in figure (b) above we can calculate the probability of $P(c|\neg a, b, d)$ as

$$\begin{aligned} P(c|\neg a, b, d) &= P(c|d) \times P(b|c, \neg a) \\ &= 0.2 \times 0.4 = 0.08 \end{aligned}$$

- A naive Bayes classifier is a Bayesian network with a specific topological structure.



(a) (b)

Figure: (a) A Bayesian network representation of the conditional independence asserted by a naive Bayes model between the descriptive features given knowledge of the target feature; (b) a Bayesian network representation of the conditional independence assumption for the naive Bayes model in the fraud example.

- When we computed a conditional probability for a target feature using a naive Bayes model, we used the following calculation

$$P(t|\mathbf{d}[1], \ldots, \mathbf{d}[n]) = P(t) \prod_{j \in Children(t)} P(\mathbf{d}[j]|t)$$

- This equation is equivalent to Equation (3)[50] from earlier.

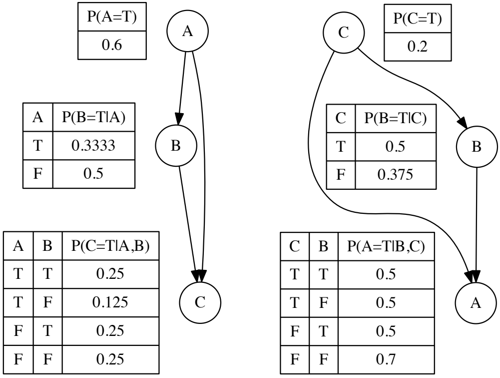
- Computing a conditional probability for a node becomes more complex if the value of one or more of the parent nodes is unknown.

- For example, in the context of the network in figure (b) above, to compute $P(b|a, d)$ where the status of node C in unknown we would do the following calculations:

1. Compute the distribution for C given D: $P(c \mid d) = 0.2$, $P(\neg c \mid d) = 0.8$
2. Compute $P(b \mid a, C)$ by summing out C: $P(b \mid a, C) = \sum_i P(b \mid a, C_i)$

$$
\begin{aligned}
P(b \mid a, C) &= \sum_i P(b \mid a, C_i) = \sum_i \frac{P(b, a, C_i)}{P(a, C_i)} \\
&= \frac{(P(b \mid a, c) \times P(a) \times P(c)) + (P(b \mid a, \neg c) \times P(a) \times P(\neg c))}{(P(a) \times P(c)) + (P(a) \times P(\neg c))} \\
&= \frac{(0.2 \times 0.4 \times 0.2) + (0.5 \times 0.4 \times 0.8)}{(0.4 \times 0.2) + (0.4 \times 0.8)} = 0.44
\end{aligned}
$$

- This example illustrates the power of Bayesian networks.
    - When complete knowledge of the state of all the nodes in the network is not available, we clamp the values of nodes that we do have knowledge of and sum out the unknown nodes.



P(A=T)
0.6

A	P(B=T|A)
T	0.3333
F	0.5

A	B	P(C=T|A,B)
T	T	0.25
T	F	0.125
F	T	0.25
F	F	0.25

(a)

P(C=T)
0.2

C	P(B=T|C)
T	0.5
F	0.375

C	B	P(A=T|B,C)
T	T	0.5
T	F	0.5
F	T	0.5
F	F	0.7

(b)

Figure: Two different Bayesian networks, each defining the same full joint probability distribution.

- We can illustrate that these two networks encode the same joint probability distribution by using each network to compute $P(\neg a, b, c)$
- Using network (a) we get:

$$P(\neg a, b, c) = P(c|\neg a, b) \times P(b|\neg a) \times P(\neg a)$$
$$= 0.25 \times 0.5 \times 0.4 = 0.05$$

- Using network (b) we get:

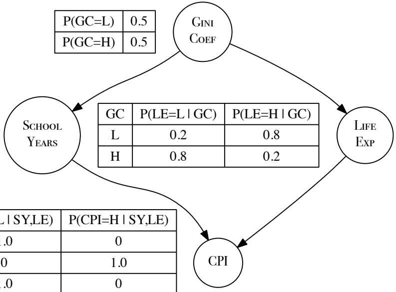
$$P(\neg a, b, c) = P(\neg a|c, b) \times P(b|c) \times P(c)$$
$$= 0.5 \times 0.5 \times 0.2 = 0.05$$

- The simplest was to construct a Bayesian network is to use a hybrid approach where:
    1. the topology of the network is given to the learning algorithm,
    2. and the learning task involves inducing the CPT from the data.

Table: (a) Some socio-economic data for a set of countries; (b) a binned version of the data listed in (a).

Country ID	Gini Coef	School Years	Life Exp	CPI	Gini Coef	School Years	Life Exp	CPI
Afghanistan	27.82	0.40	59.61	1.52	low	low	low	low
Argentina	44.49	10.10	75.77	3.00	high	low	low	low
Australia	35.19	11.50	82.09	8.84	low	high	high	high
Brazil	54.69	7.20	73.12	3.77	high	low	low	low
Canada	32.56	14.20	80.99	8.67	low	high	high	high
China	42.06	6.40	74.87	3.64	high	low	low	low
Egypt	30.77	5.30	70.48	2.86	low	low	low	low
Germany	28.31	12.00	80.24	8.05	low	high	high	high
Haiti	59.21	3.40	45.00	1.80	high	low	low	low
Ireland	34.28	11.50	80.15	7.54	low	high	high	high
Israel	39.2	12.50	81.30	5.81	low	high	high	high
New Zealand	36.17	12.30	80.67	9.46	low	high	high	high
Nigeria	48.83	4.10	51.30	2.45	high	low	low	low
Russia	40.11	12.90	67.62	2.45	high	high	low	low
Singapore	42.48	6.10	81.788	9.17	high	low	high	high
South Africa	63.14	8.50	54.547	4.08	high	low	low	low
Sweden	25.00	12.80	81.43	9.30	low	high	high	high
U.K.	35.97	13.00	80.09	7.78	low	high	high	high
U.S.A	40.81	13.70	78.51	7.14	high	high	high	high
Zimbabwe	50.10	6.7	53.684	2.23	high	low	low	low

(a) (b)



P(GC=L)	0.5
P(GC=H)	0.5

GC	P(SY=L I GC)	P(SY=H I GC)
L	0.2	0.8
H	0.8	0.2

GC	P(LE=L I GC)	P(LE=H I GC)
L	0.2	0.8
H	0.8	0.2

SY	LE	P(CPI=L I SY,LE)	P(CPI=H I SY,LE)
L	L	1.0	0
L	H	0	1.0
H	L	1.0	0
H	H	0	1.0

Figure: A Bayesian network that encodes the causal relationships between the features in the corruption domain. The CPT entries have been calculated using the data from Table 16 [61](b).

$$\mathbb{M}(\mathbf{q}) = \operatorname*{argmax}_{l \in levels(t)} BayesianNetwork(t = l, \mathbf{q}) \tag{4}$$

Example

- We wish to predict the CPI for a country with the follow profile:

  GINI COEF = *'high'*, SCHOOL YEARS = *'high'*

$$P(CPI = H | SY = H, GC = H) = \frac{P(CPI = H, SY = H, GC = H)}{P(SY = H, GC = H)}$$

$$= \frac{\sum_{i \in H,L} P(CPI = H, SY = H, GC = H, LE = i)}{P(SY = H, GC = H)}$$

$$
\begin{aligned}
&\sum_{i\in\{H,L\}} P(CPI = H, SY = H, GC = H, LE = i) \\
&\quad = \sum_{i\in\{H,L\}} P(CPI = H|SY = H, LE = i) \times P(SY = H|GC = H) \\
&\qquad\qquad \times P(LE = i|GC = H) \times P(GC = H) \\
&\quad = (P(CPI = H|SY = H, LE = H) \times P(SY = H|GC = H) \\
&\qquad \times P(LE = H|GC = H) \times P(GC = H)) \\
&\qquad + (P(CPI = H|SY = H, LE = L) \times P(SY = H|GC = H) \\
&\qquad \times P(LE = L|GC = H) \times P(GC = H)) \\
&\quad = (1.0 \times 0.2 \times 0.2 \times 0.5) + (0 \times 0.2 \times 0.8 \times 0.5) = 0.02
\end{aligned}
$$

$$
\begin{aligned}
P(SY = H, GC = H) &= P(SY = H | GC = H) \times P(GC = H) \\
&= 0.2 \times 0.5 = 0.1
\end{aligned}
$$

$$P(CPI = H | SY = H, GC = H) = \frac{0.02}{0.1} = 0.2$$

- Because of the calculation complexity that can arise when using Bayesian networks to do exact inference a popular approach is to approximate the required probability distribution using **Markov Chain Monte Carlo** algorithms.
- **Gibbs sampling** is one of the best known MCMC algorithms.
  1. Clamp the values of the evidence variables and randomly assign the values of the non-evidence variables.
  2. Generate samples by changing the value of one of the non-evidence variables using the distribution for the node conditioned on the state of the rest of the network.

Table: Examples of the samples generated using Gibbs sampling.

Sample Number	Gibbs Iteration	Feature Updated	GINI COEF	SCHOOL YEARS	LIFE EXP	CPI
1	37	CPI	high	high	high	low
2	44	LIFE EXP	high	high	high	low
3	51	CPI	high	high	high	low
4	58	LIFE EXP	high	high	low	high
5	65	CPI	high	high	high	low
6	72	LIFE EXP	high	high	high	low
7	79	CPI	high	high	low	high
8	86	LIFE EXP	high	high	low	low
9	93	CPI	high	high	high	low
10	100	LIFE EXP	high	high	high	low
11	107	CPI	high	high	low	high
12	114	LIFE EXP	high	high	high	low
13	121	CPI	high	high	high	low
14	128	LIFE EXP	high	high	high	low
15	135	CPI	high	high	high	low
16	142	LIFE EXP	high	high	low	low

...

$$\mathbb{M}(\mathbf{q}) = \operatorname*{argmax}_{l \in \mathit{levels}(t)} \mathit{Gibbs}\,(t = l, \mathbf{q}) \tag{5}$$

Summary

- Naive Bayes models can suffer from zero probabilities of relatively rare events. **Smoothing** is an easy way to combat this.
- Two ways to handle continuous features in probability-based models are: **Probability density functions** and **Binning**
- Using probability density functions requires that we match the observed data to an existing distribution.
- Although binning results in information loss it is a simple and effective way to handle continuous features in probability-based models.
- Bayesian network representation is generally more compact than a full joint distribution, yet is not forced to assert global conditional independence between all descriptive features.