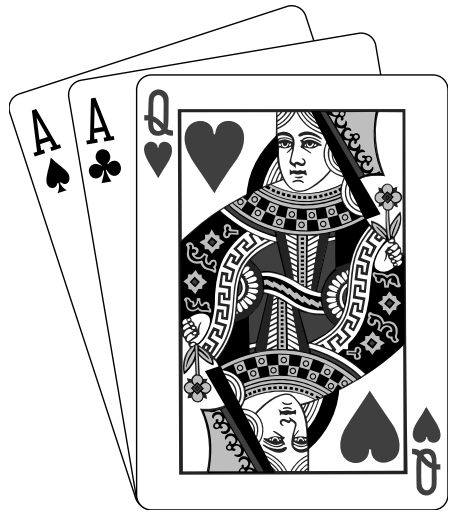
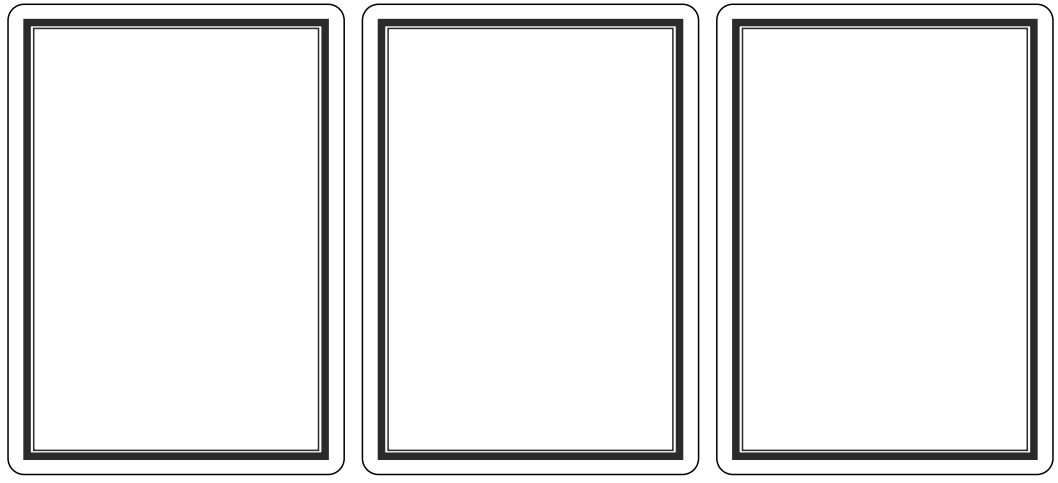
Probability-based Learning

Outline

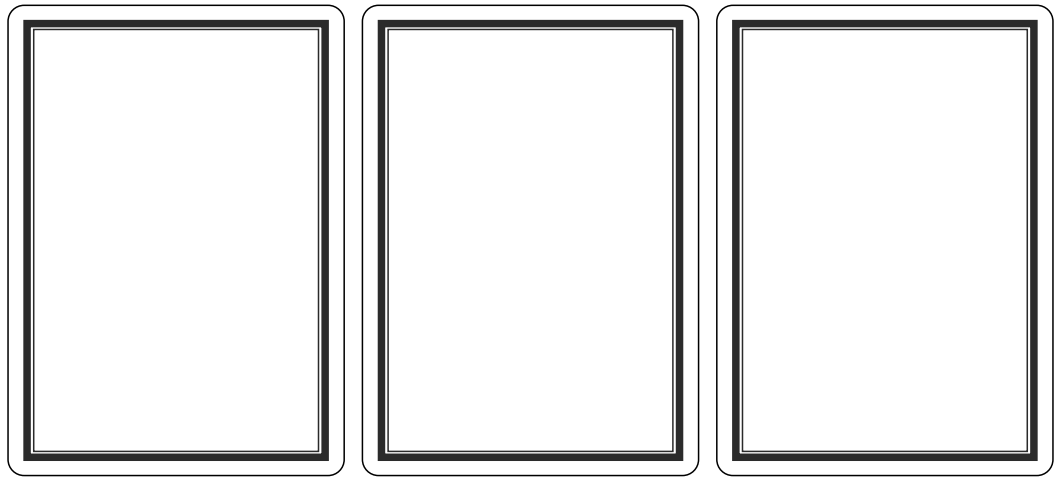


(a)

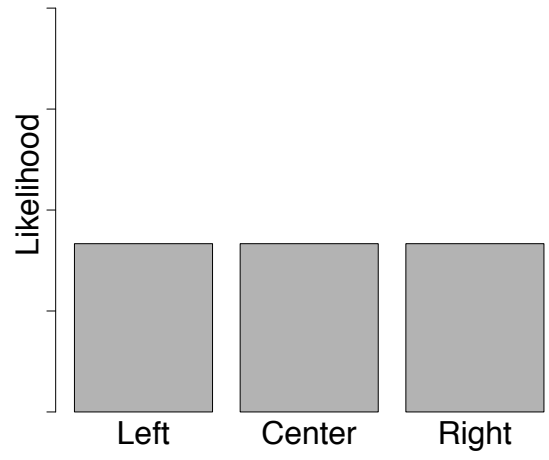


(b)

A game of *find the lady*

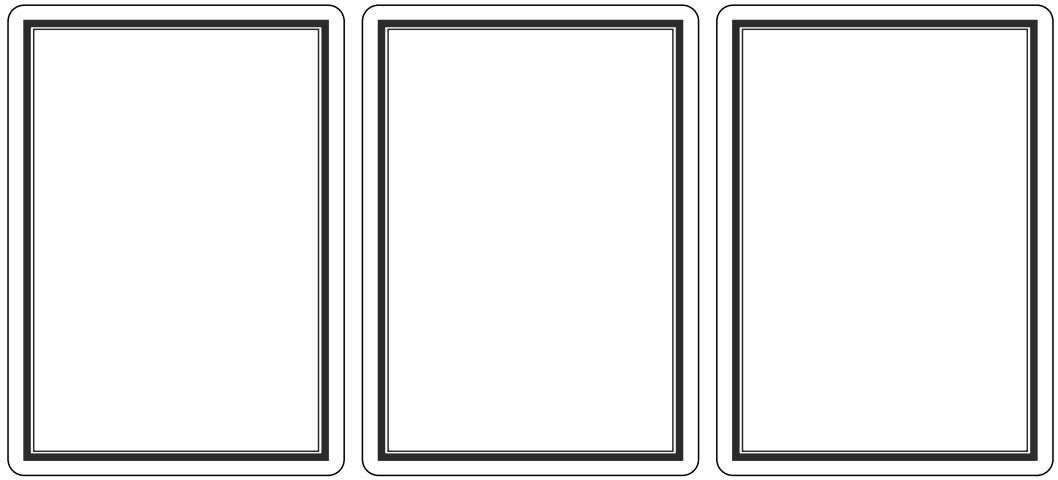


(c)

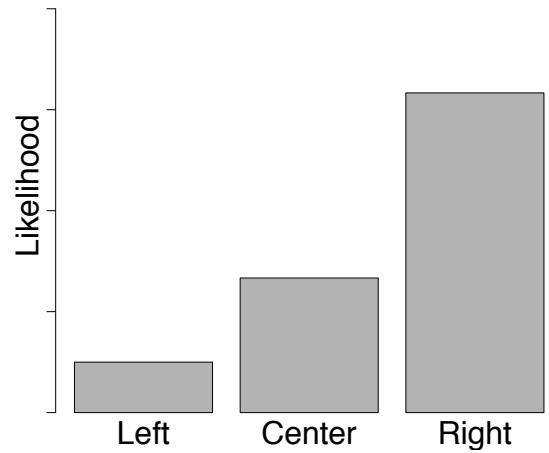


(d)

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

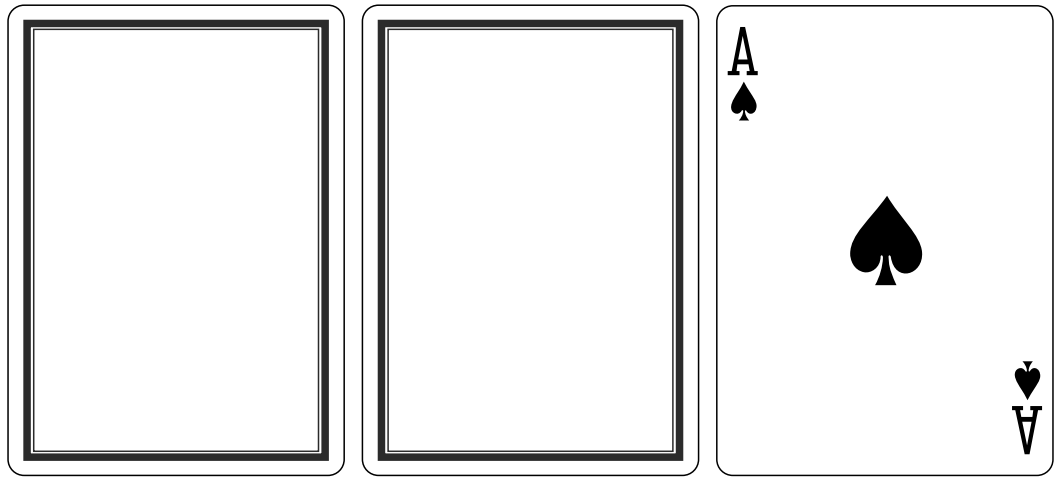


(e)

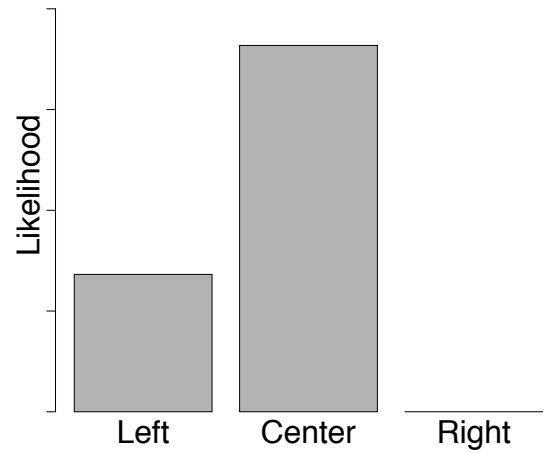


(f)

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

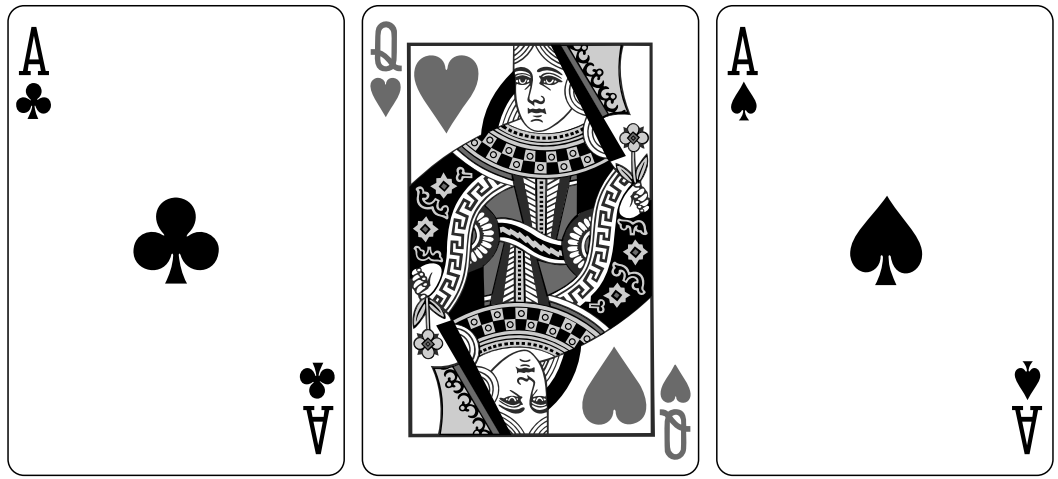


(g)



(h)

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



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

Big Idea

- Use likelihood estimates to determine the most likely prediction.
- Revise predictions based on new data and additional evidence.

A 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.
- **Joint probability** is the probability of specific values assigned to multiple features.
- **Conditional probability** is the probability of a feature's value given the value of another feature.
- A **probability distribution** describes the probability of each possible value a feature can take, summing to 1.0.
- A **joint probability distribution** is a multi-dimensional matrix showing probabilities of combinations of feature values, summing to 1.0.

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

- From a joint probability distribution, compute the probability of any event by summing the relevant cells.
- This calculation method is called **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 despite a positive test?

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

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

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

Deriving Bayes theorem

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

- 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 use the **Theorem of Total Probability**

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

Bayesian Prediction

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, add the conditioning term to each term in the expression.

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

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	?

- With Bayes' Theorem:

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

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

- We will calculate m first, and need to know $P(m)$, $P(h, \neg f, v)$ 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 $P(m)$ and $P(h, \neg f, v)$ directly from the data:

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

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

- However, as an exercise we use 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

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

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

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

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

- These calculations show it is twice as likely the patient does not have meningitis, despite having a headache and vomiting.

The Paradox of the False Positive

- Forgetting to factor in the prior leads to the **paradox of the false positive**, which states that predicting a rare event requires a model as accurate as the event is rare, or there is a significant chance of **false positives**.

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)

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

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

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

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

- There is something odd about these results!

Curse of Dimensionality

As the number of descriptive features grows, the number of potential conditioning events increases exponentially. Thus, the dataset size must also grow exponentially to ensure sufficient instances for each conditional probability.

- The probability of a patient with a headache and fever having meningitis should be greater than zero.
- Our dataset is too small, leading to over-fitting.
- Conditional independence and factorization can help address this flaw.

Conditional Independence and Factorization

- 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:

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

- Full independence between events is quite rare.
- More commonly, two or more events are independent if a third event has occurred, which is known as conditional independence.

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

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

- If $t = l$ causes $\mathbf{q}[1], \ldots, \mathbf{q}[m]$ to happen, then the events $\mathbf{q}[1], \ldots, \mathbf{q}[m]$ are conditionally independent of each other given $t = l$. The chain rule is:

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

- 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:

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

Without 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

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

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

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

$$\begin{aligned} \text{Factor}_1 &: P(M) \\ \text{Factor}_2 &: P(h|m),\ P(h|\neg m) \\ \text{Factor}_3 &: P(f|m),\ P(f|\neg m) \\ \text{Factor}_4 &: P(v|m),\ P(v|\neg m) \end{aligned}$$

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

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.

$Factor_1 : < P(m) = 0.3 >$

$Factor_2 : < P(h|m) = 0.6666, P(h|\neg m) = 0.7413 >$

$Factor_3 : < P(f|m) = 0.3333, P(f|\neg m) = 0.4286 >$

$Factor_4 : < P(v|m) = 0.6666, P(v|\neg m) = 0.5714 >$

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

- Calculate the probability of MENINGITIS=*'true'*.

$$P(m|h,f,\neg v) = \frac{P(h|m) \times P(f|m) \times P(\neg v|m) \times P(m)}{\sum_i P(h|M_i) \times P(f|M_i) \times P(\neg v|M_i) \times P(M_i)} =$$
$$\frac{0.6666 \times 0.3333 \times 0.3333 \times 0.3}{(0.6666 \times 0.3333 \times 0.3333 \times 0.3) + (0.7143 \times 0.4286 \times 0.4286 \times 0.7)} = 0.1948$$

- Calculate the probability of MENINGITIS=*'false'*.

$$P(\neg m|h,f,\neg v) = \frac{P(h|\neg m) \times P(f|\neg m) \times P(\neg v|\neg m) \times P(\neg m)}{\sum_i P(h|M_i) \times P(f|M_i) \times P(\neg v|M_i) \times P(M_i)} =$$
$$\frac{0.7143 \times 0.4286 \times 0.4286 \times 0.7}{(0.6666 \times 0.3333 \times 0.3333 \times 0.3) + (0.7143 \times 0.4286 \times 0.4286 \times 0.7)} = 0.8052$$

- As before, the MAP prediction MENINGITIS $=$ *'false'*, but the posterior probabilities are not as extreme!

Standard Approach: The Naive Bayes' Classifier

Naive Bayes' Classifier

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

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.

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

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMODATION	FRAUDULENT
paid	none	rent	?

$P(fr)$	$=$	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

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

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

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

The model is generalizing beyond the dataset!

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	?

$P(fr)$	$=$	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

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

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

- 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.

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

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

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

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	?

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

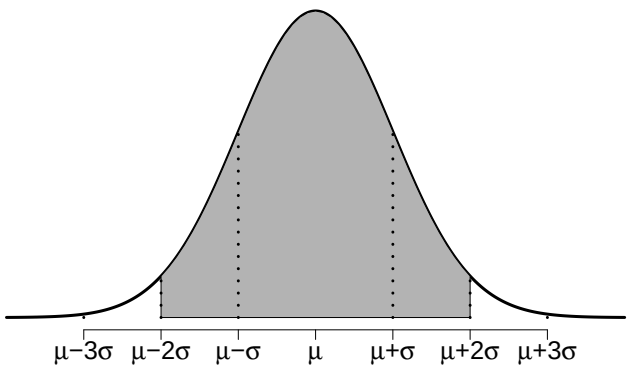
$$\left(\prod_{k=1}^{m} P(\mathbf{q}[m]|fr)\right) \times P(fr) = 0.0036$$
$$\left(\prod_{k=1}^{m} P(\mathbf{q}[m]|\neg fr)\right) \times P(\neg fr) = 0.0043$$

The relevant smoothed probabilities needed by the Naive Bayes prediction model to classify the query 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 is determined by:
  - the statistical distribution defining the PDF
  - the values of the distribution parameters

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
$x \in \mathbb{R}$
$\phi \in \mathbb{R}$
$\rho \in \mathbb{R}_{>0}$
$\kappa \in \mathbb{R}_{>0}$
$z = \dfrac{x - \phi}{\rho}$

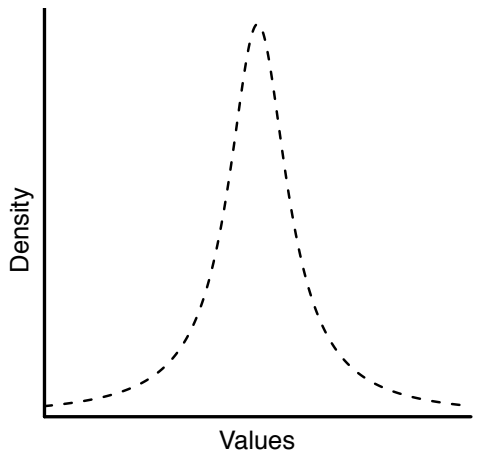
$$\tau(x, \phi, \rho, \kappa) = \frac{\Gamma(\frac{\kappa+1}{2})}{\Gamma(\frac{\kappa}{2}) \times \sqrt{\pi\kappa} \times \rho} \times \left(1 + \left(\frac{1}{\kappa} \times z^2\right)\right)^{-\frac{\kappa+1}{2}}$$

Exponential
$x \in \mathbb{R}$
$\lambda \in \mathbb{R}_{>0}$

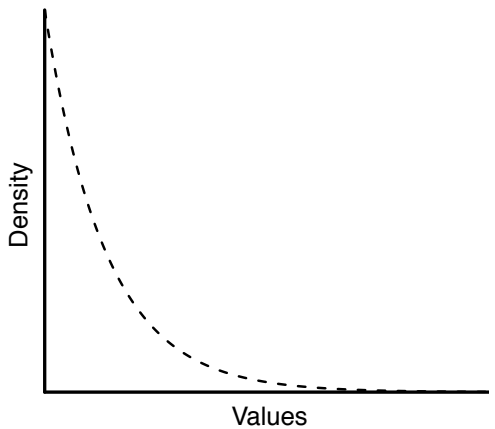
$$E(x, \lambda) = \begin{cases} \lambda e^{-\lambda x} & \text{for } x > 0 \\ 0 & \text{otherwise} \end{cases}$$

Mixture of n Gaussians
$x \in \mathbb{R}$
$\{\mu_1, \ldots, \mu_n | \mu_i \in \mathbb{R}\}$
$\{\sigma_1, \ldots, \sigma_n | \sigma_i \in \mathbb{R}_{>0}\}$
$\{\omega_1, \ldots, \omega_n | \omega_i \in \mathbb{R}_{>0}\}$
$\sum_{i=1}^{n} \omega_i = 0$

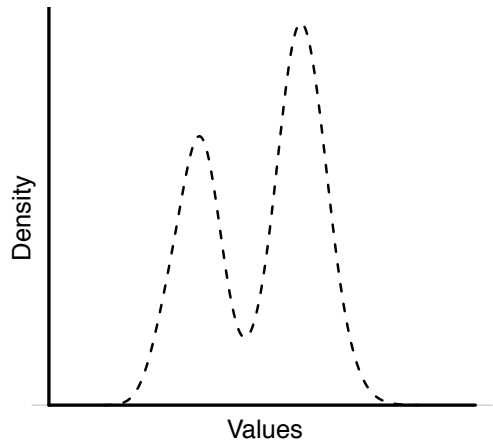
$$N(x, \mu_1, \sigma_1, \omega_1, \ldots, \mu_n, \sigma_n, \omega_n) = \sum_{i=1}^{n} \frac{\omega_i}{\sigma_i\sqrt{2\pi}} e^{-\frac{(x-\mu_i)^2}{2\sigma_i^2}}$$



(i) Normal/Student-t



(j) Exponential



(k) Mixture of Gaussians

Figure: Plots of some well known probability distributions.

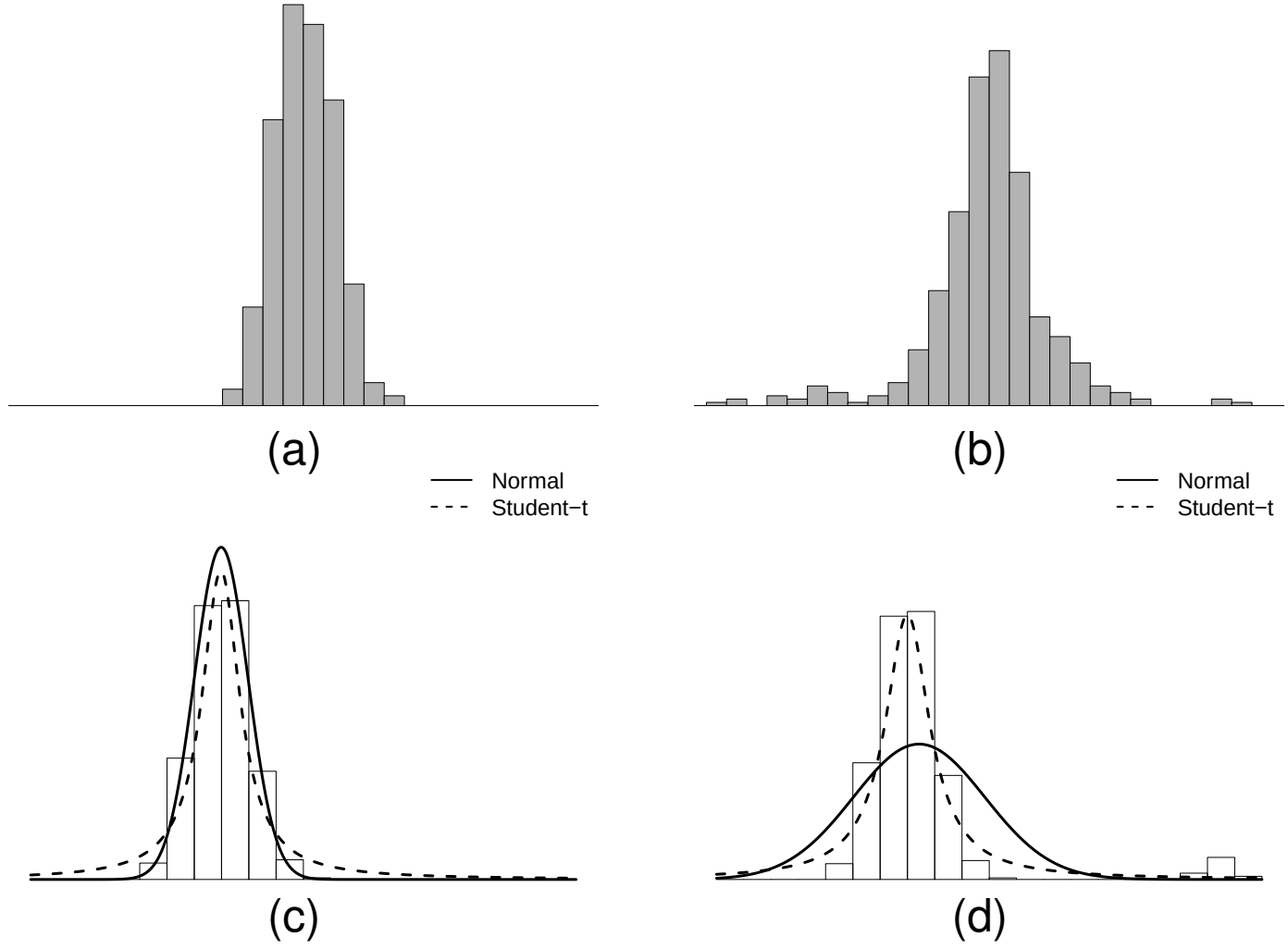


Illustration of the robustness of the student-*t* distribution to outliers: (c) a density histogram of a unimodal dataset overlaid with the density curves of a normal and a student-*t* distribution that have been fitted to the data; (d) a density histogram of the same dataset with outliers added, overlaid 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.

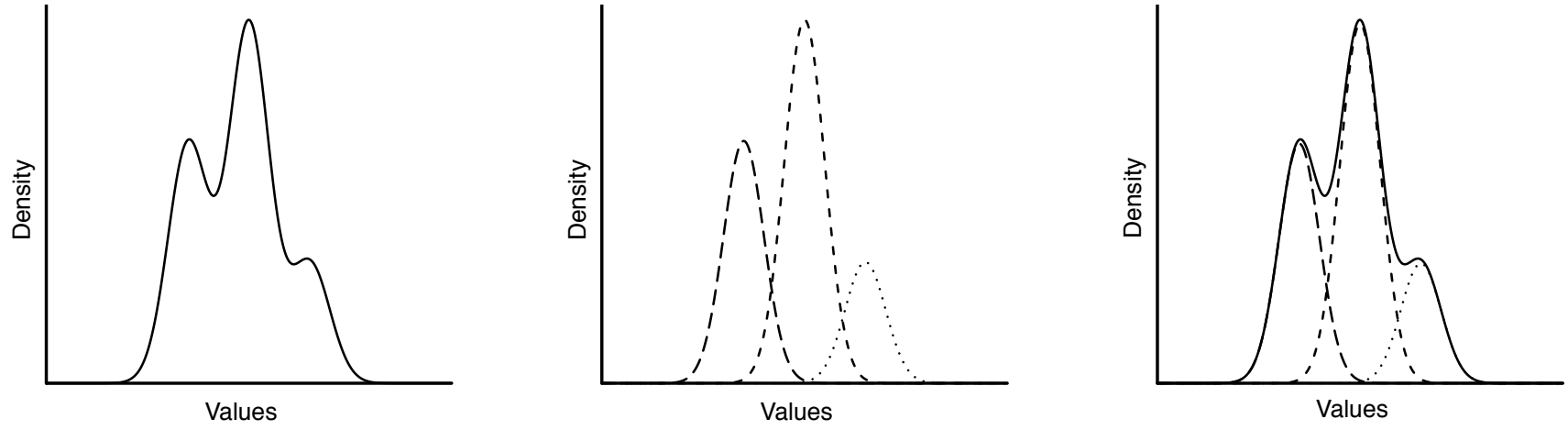
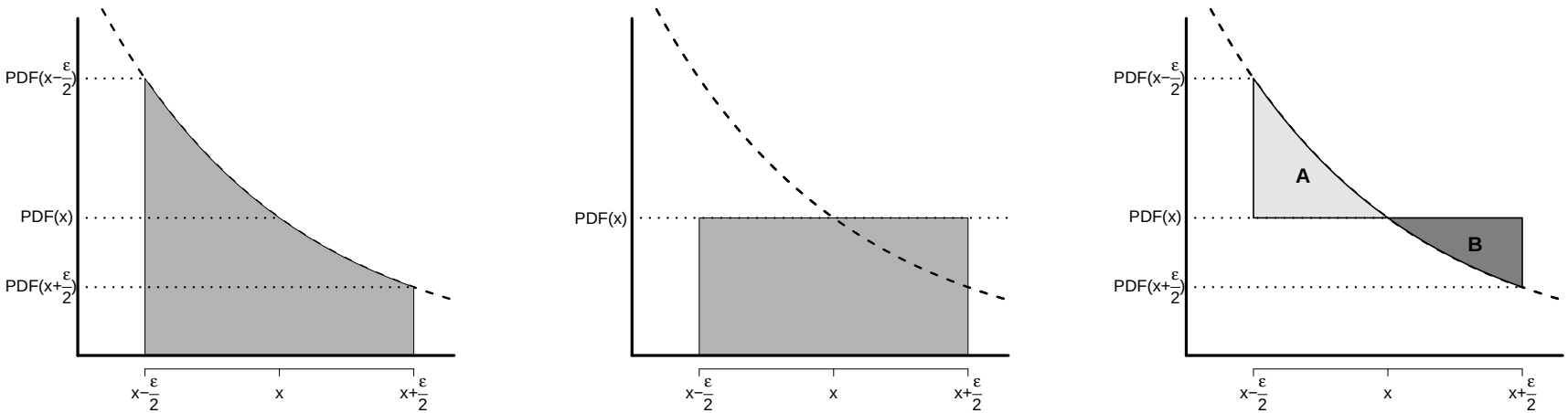


Illustration of how a mixture of Gaussians model is composed of a number of normal distributions. The curve plotted using a solid line is the mixture of Gaussians density curve, created using an appropriately weighted summation of the three normal curves, plotted using dashed and dotted lines.

- A PDF abstracts a density histogram, representing probabilities as areas under the curve.
- To calculate a probability, consider the area under the PDF curve over an interval.
- Find this area using probability tables or integration within the interval bounds.



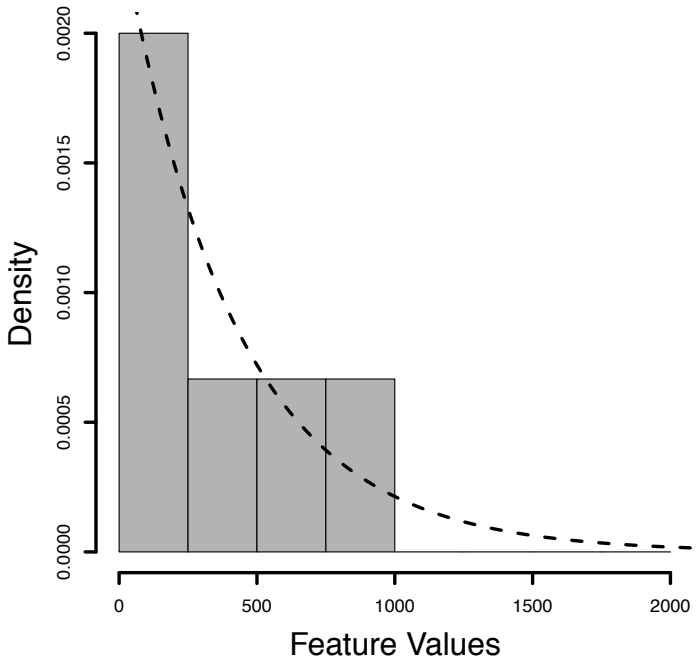
(h) (i) (j)

(h) The area under a density curve between the limits $x - \frac{\epsilon}{2}$ and $x + \frac{\epsilon}{2}$; (i) the approximation of this area computed by $PDF(x) \times \epsilon$; and (j) 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.

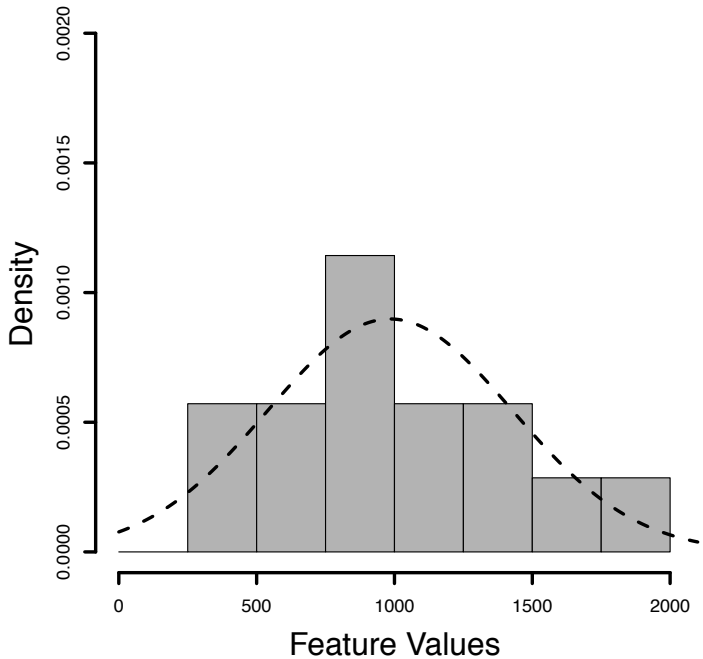
To illustrate how to use PDFs in NB, we extend loan application fraud detection query to have an ACCOUNT BALANCE feature

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

- Define two PDFs for the new feature with each PDF conditioned on a different value in the domain or the target:
  - $P(AB = X|fr) = PDF_1(AB = X|fr)$
  - $P(AB = X|\neg fr) = PDF_2(AB = X|\neg fr)$
- PDFs can be defined with different distributions.



(k)



(l)

- Histograms indicate:
  - ACCOUNT BALANCE for FRAUDULENT=*'True'* follows an exponential distribution.
  - ACCOUNT BALANCE for FRAUDULENT=*'False'* resembles a normal distribution.
- Next, fit the distributions to the data.
  - For the exponential distribution, compute the sample mean $\bar{x}$ of ACCOUNT BALANCE where FRAUDULENT=*'True'* and set λ to $1/\bar{x}$.
  - For the normal distribution, compute the sample mean and standard deviation s of ACCOUNT BALANCE where FRAUDULENT=*'False'* and use these as the distribution parameters.

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	

The Laplace smoothed probabilities extended to include ACCOUNT BALANCE

$P(fr)$	$=$	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
$P(AB = x \mid fr)$			$P(AB = x \mid \neg fr)$		
	$\approx$	$E\begin{pmatrix} x, \\ \lambda = 0.0024 \end{pmatrix}$		$\approx$	$N\begin{pmatrix} x, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix}$

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

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

$$P(CH = paid|fr) = 0.2222 \qquad P(CH = paid|\neg fr) = 0.2692$$

$$P(GC = guarantor|fr) = 0.2667 \qquad P(GC = guarantor|\neg fr) = 0.1304$$

$$P(ACC = free|fr) = 0.2 \qquad P(ACC = free|\neg fr) = 0.1739$$

$$P(AB = 759.07|fr) \approx E\begin{pmatrix} 759.07, \\ \lambda = 0.0024 \end{pmatrix} = 0.00039$$

$$P(AB = 759.07|\neg fr) \approx N\begin{pmatrix} 759.07, \\ \mu = 984.26, \\ \sigma = 460.94 \end{pmatrix} = 0.00077$$

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

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

Binning

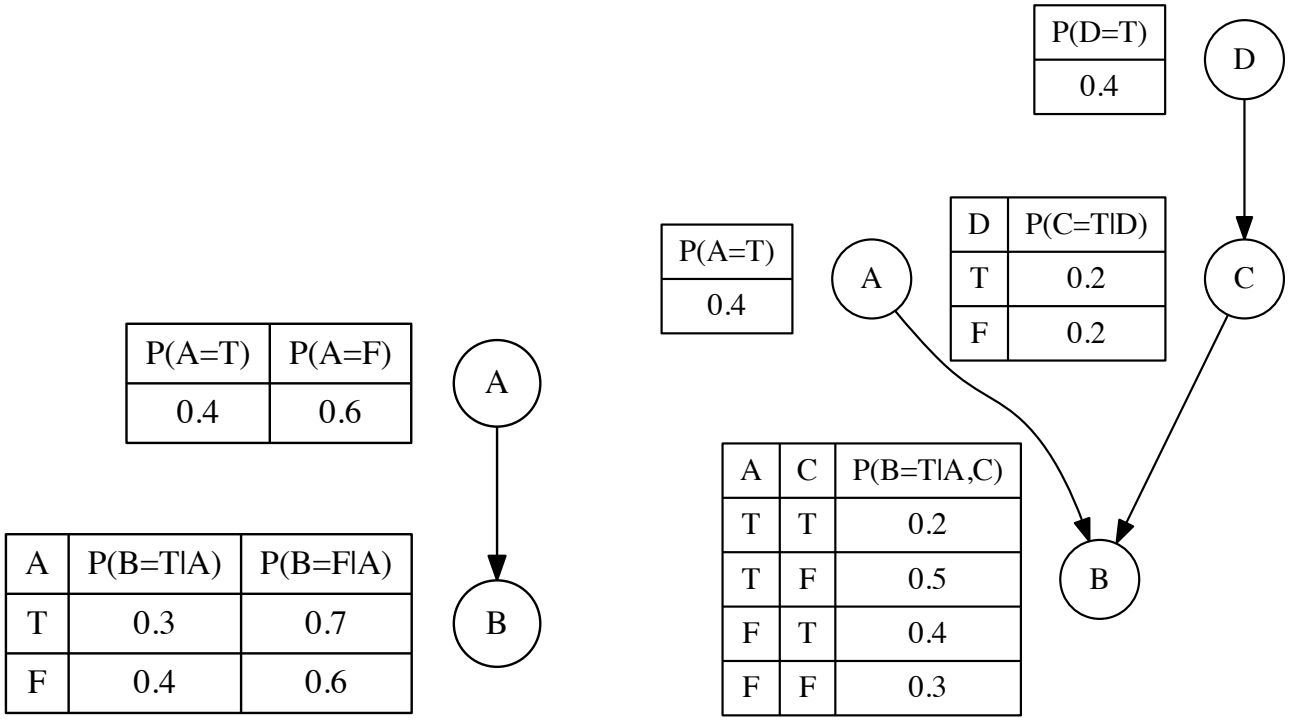
- We have explained two of the best known binning techniques **equal-width** and **equal-frequency**.
- We can use these techniques to *bin* continuous features into categorical features
- In general we recommend **equal-frequency binning**.

Bayesian Networks

- **Bayesian networks** use a graph to represent structural relationships like direct influence and conditional independence between features.
- This makes Bayesian networks more compact than a full joint distribution without assuming global conditional independence between all features.

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

- nodes
- edges
- conditional probability tables (CPT)



(m) (n)

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

- In probability terms, the directed edge from A to B states:

$$P(A, B) = P(B|A) \times P(A) \tag{2}$$

- 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 (2)[51] can be generalized as, for any network with N nodes, the probability of an event $x_1, \ldots, x_n$ is:

$$P(x_1, \ldots, x_n) = \prod_{i=1}^{n} P(x_i | Parents(x_i)) \tag{3}$$

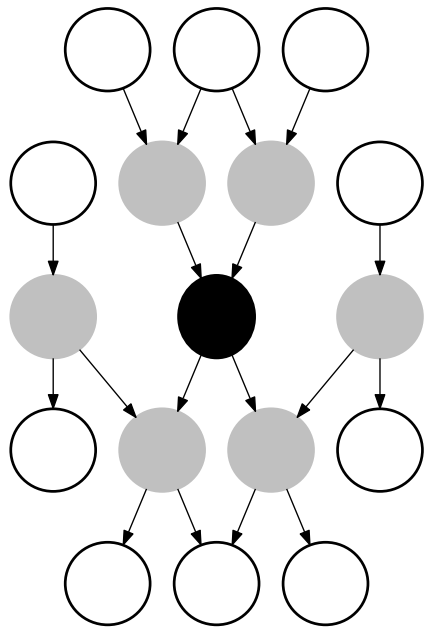
- For example, using (y) above, the probability of the joint event $P(a, \neg b, \neg c, d)$ is:

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

- 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, consider a node's parents, its children, and their parents.
- The set of nodes that make a node independent of the rest of the graph is called the **Markov blanket**.



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:

$$P(x_i|x_1,\ldots,x_{i-1},x_{i+1},\ldots,x_n) = P(x_i|Parents(x_i)) \prod_{j\in Children(x_i)} P(x_j|Parents(x_j)) \quad (4)$$

- 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.

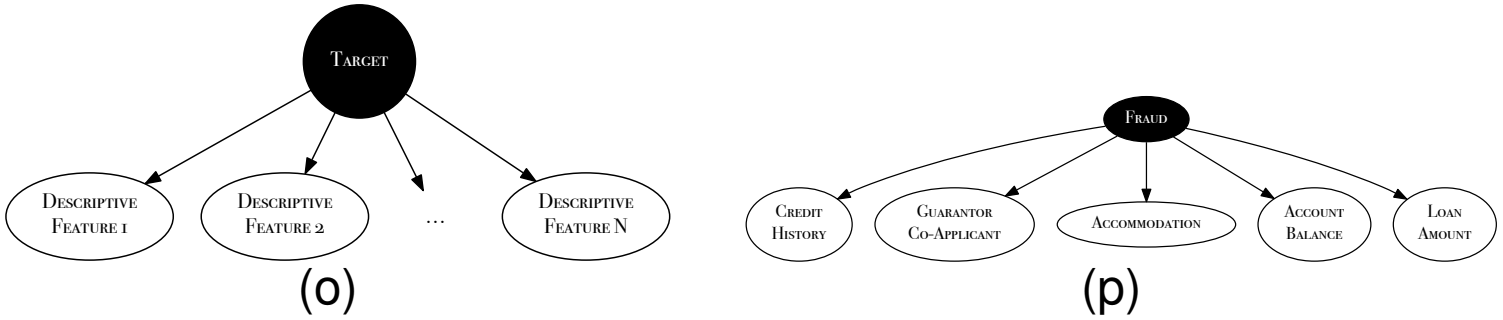


Figure: (o) A Bayesian network representation of the conditional independence asserted by a naive Bayes model between the descriptive features given knowledge of the target feature; (p) a Bayesian network representation of the conditional independence assumption for the naive Bayes model in the fraud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

- While computing a conditional probability of a target feature using a NB, we used

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

- This equation is equivalent to Equation (4)[55].

- Computing a conditional probability for a node becomes more complex if the value of one or more of the parent nodes is unknown.
- In (n), to compute $P(b|a,d)$ where the status of node C in unknown we would do:
  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: $P(b \mid a, C) = \sum_i P(b \mid a, C_i)$

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

- This illustrates the power of BN: when complete knowledge of the state of all the nodes is not available, we clamp the values of nodes that we do have knowledge of and sum out the unknown nodes.

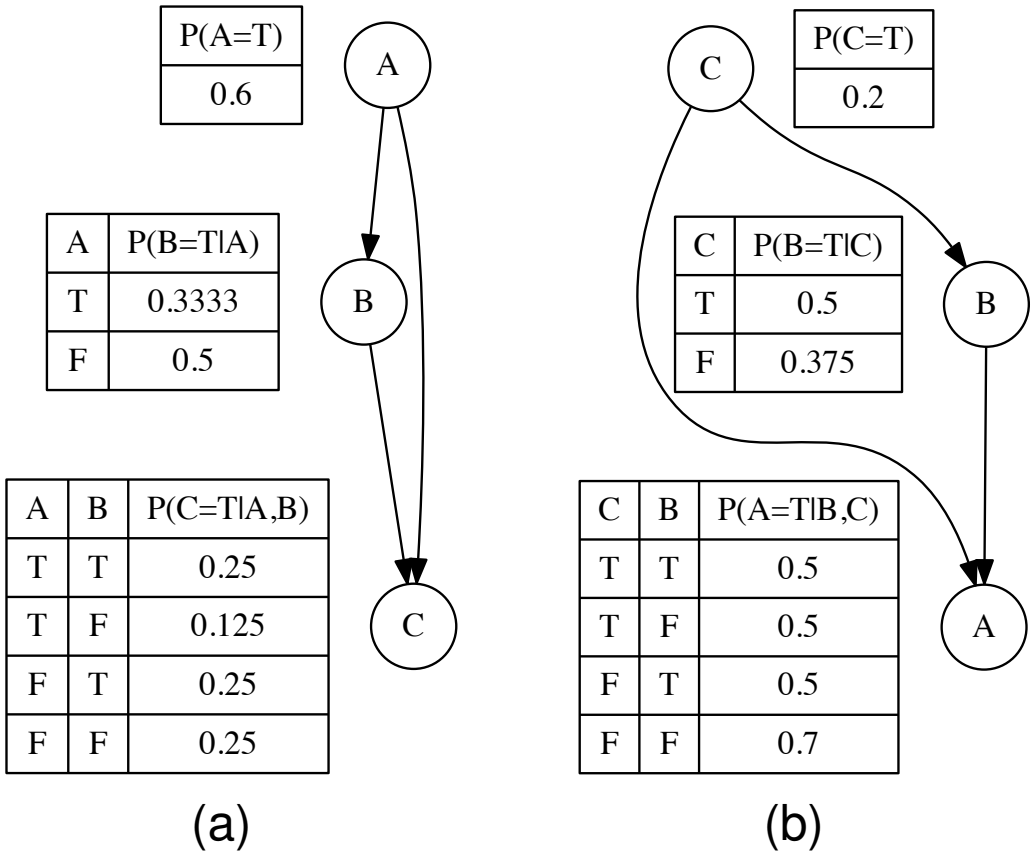


Figure: Two different Bayesian networks, each defining the same full joint probability distribution.

- 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:

$$\begin{aligned} P(\neg a, b, c) &= P(c|\neg a, b) \times P(b|\neg a) \times P(\neg a) \\ &= 0.25 \times 0.5 \times 0.4 = 0.05 \end{aligned}$$

- Using network (b) we get:

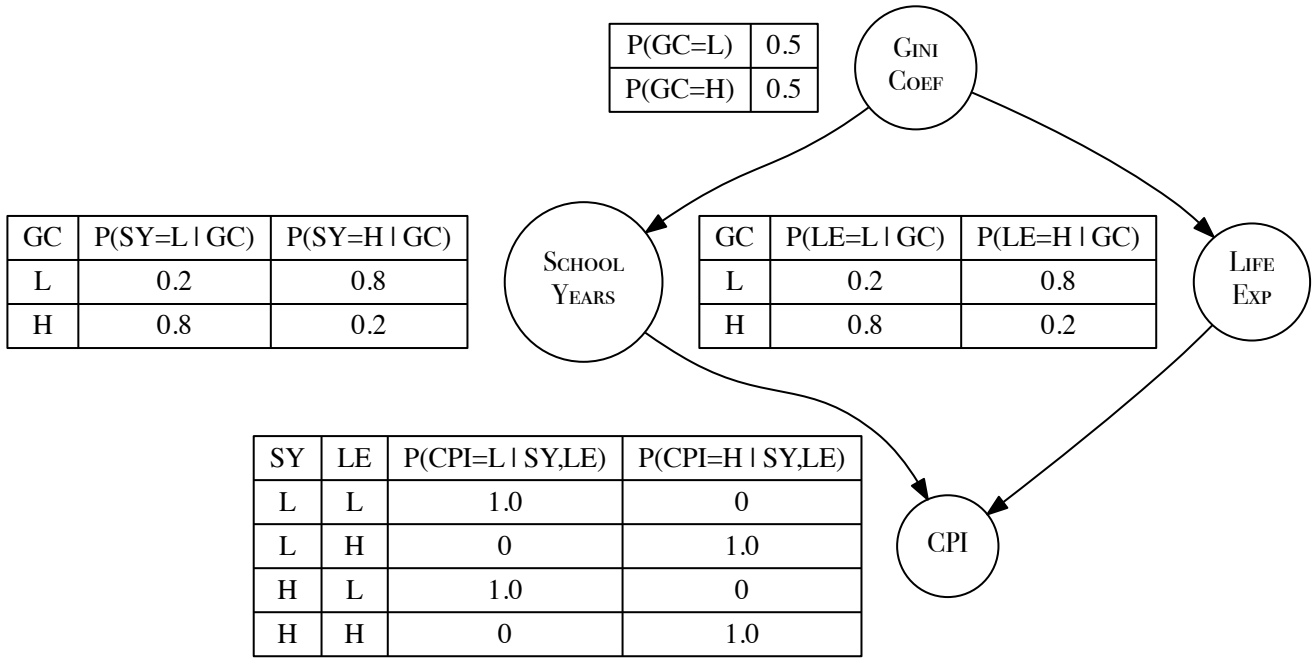
$$\begin{aligned} P(\neg a, b, c) &= P(\neg a|c, b) \times P(b|c) \times P(c) \\ &= 0.5 \times 0.5 \times 0.2 = 0.05 \end{aligned}$$

- 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)



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

Example

Predict CPI with GINI COEF = *'high'*, SCHOOL YEARS = *'high'*

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

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

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

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

$$\times P(LE = i|GC = H) \times P(GC = H)$$

$$= (1.0 \times 0.2 \times 0.2 \times 0.5) + (0 \times 0.2 \times 0.8 \times 0.5) = 0.02$$

$$P(SY = H, GC = H) = P(SY = H|GC = H) \times P(GC = H)$$

$$= 0.2 \times 0.5 = 0.1$$

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