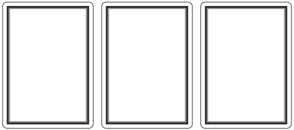
Chapter 6: Probability-based Learning

MATH2319

Big Idea

(a)



(b)

Figure: A game of *find the lady*: two Aces, one Queen

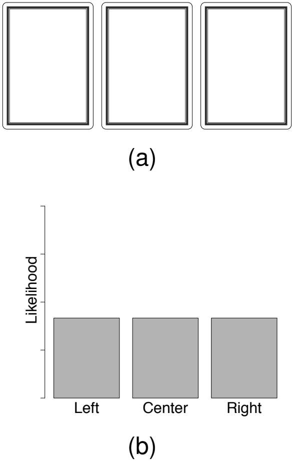
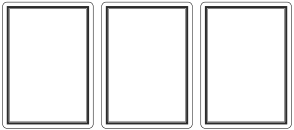
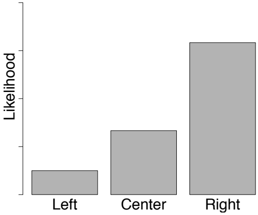


Figure: A game of *find the lady*: (a) the cards dealt face down on a table; and (b) the initial likelihoods of the queen ending up in each position.

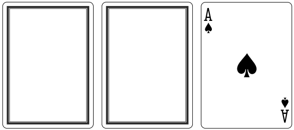


(a)

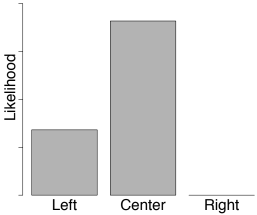


(b)

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



(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.

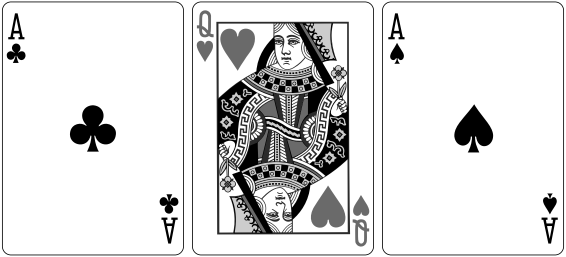


Figure: A game of *find the lady*: The final positions of the cards in the game.

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.

- 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**.

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

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

- Examples of “Summing Out”:
  1. P(not f) = sum of all cells where “not f” holds.
  2. P(h and not f) = sum of all cells where “h and not f” holds.

- Notation: P(X and Y) = P(XY)
- Definition of conditional probability: $P(X|Y) = \frac{P(XY)}{P(Y)}$
- Reorganising terms, we get:
- P(XY) = P(X|Y) P(Y) = P(Y|X) P(X)
- The above implies that it doesn't matter how you name the events!
- Moving P(Y) to the RHS (right-hand side) yields the Bayes' Theorem:

Bayes' Theorem

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

Example

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

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

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

$$\begin{aligned} P(t) &= P(t|d)P(d) + P(t|\neg d)P(\neg d) \quad \text{(Theorem of Total Probability)} \\ &= (0.99 \times 0.0001) + (0.01 \times 0.9999) = 0.0101 \end{aligned}$$

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

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

Generalized Bayes' Theorem

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

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:

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

- We can do the calculation only for $P(m|\text{evidence})$. The probability for $P(\neg m|\text{evidence})$ will be $1 - P(m|\text{evidence})$.
- 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 two of the required probabilities directly from the data: $P(m)$ and $P(h, \neg f, v)$:

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

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

- How about $P(h, \neg f, v \mid m)$?
- This conditional probability can be calculated by the "Chain Rule" for probabilities, but this usually results in a zero probability due to "curse of dimensionality".
- So, we will make a simplifying assumption to avoid this curse of dimensionality: we will assume **conditional independence**.

- 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 conditional probability is:

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

- 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 $t = l$, we have:

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

- Recall Generalised Bayes' Theorem:

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

- Under the assumption of conditional independence between the descriptive features given the level l of the target feature, Bayes' Theorem can be simplified as follows:

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

- X, Y, Z: descriptive features (evidence)
- T: target feature

$$P(X, Y, Z, T) = P(X, Y, Z|T)P(T)$$

- With conditional independence:

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

- The joint probability distribution for the meningitis dataset.

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

- We assume the descriptive features are conditionally independent of each other **given MENINGITIS**.
- Likewise, we assume the descriptive features are conditionally independent of each other **given NON-MENINGITIS**.
- Under this conditional independence assumption, 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.

$$Factor_1 : < P(M) >$$

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

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

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

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

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

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

- 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	?

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

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

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

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

- The prediction would be MENINGITIS $=$ *'false'*
- In practice, we need to compute only one of these probabilities.

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'*.

$P(fr)$	$=$	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	?

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

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

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

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMODATION	FRAUDULENT
paid	none	rent	?

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

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

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

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'*

Bernoulli Naive Bayes

- The previous examples work with categorical descriptive features without any one-hot encoding to binary features.
- Let's now assume we only have categorical descriptive features and they are all encoded via one-hot-encoding.
- One variant of Naive Bayes in this case is the **Bernoulli Naive Bayes:** it assumes all of the descriptive features are binary, that is, either 0 or 1.
- This will clearly be the case if we only have categorical descriptive features are they are all encoded via one-hot-encoding!
- You can use the **BernoulliNB()** classifier in Scikit-Learn for this type of learning.

- Bernoulli Naive Bayes has only one hyper-parameter to fine tune: **"alpha"**, which denotes the value of Laplace smoothing (default value is 1.0).
- See Section 6.4.1 in the textbook on Laplace smoothing.
- If there are any descriptive features that are not binary in your training data (e.g., a numerical feature), **BernoulliNB()** will make them binary by setting a threshold of 0 (see the Scikit-Learn documentation on how to change this threshold).

Gaussian Naive Bayes

- Let's now assume all descriptive features are numerical.
- One variant of Naive Bayes in this case is the **Gaussian Naive Bayes:** it assumes all of these numerical features follow a Gaussian, that is, a normal distribution.
- This is highly unlikely in practice, but we can perform a **power transformation** on each feature to make it more or less normally distributed.
- We will let Python perform this transformation for us.
- Gaussian Naive Bayes has only one hyper-parameter to fine tune: **"var-smoothing"**, which specifies the portion of the largest variance of all features to be added to variances for numerical stability.

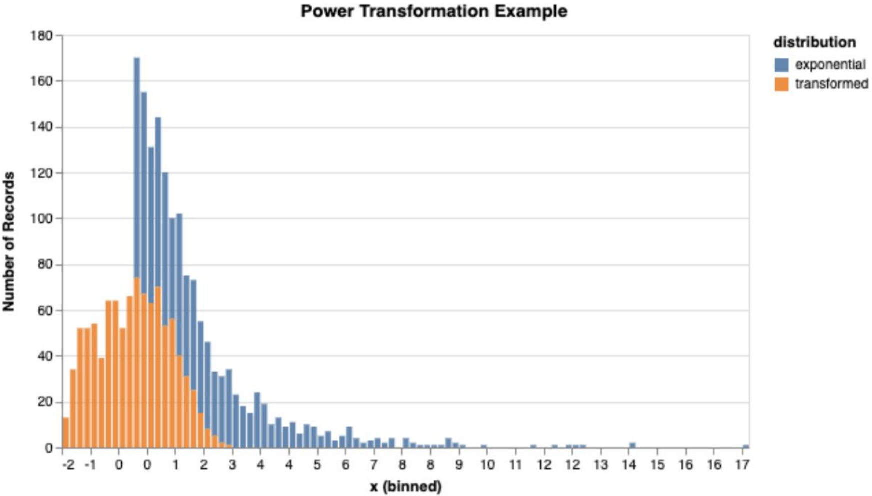


Figure: Transforming an exponential distribution (with mean 2.0) to a normal distribution.

Summary

$$P(t|\mathbf{d}) = \frac{P(\mathbf{d}|t) \times P(t)}{P(\mathbf{d})} \tag{1}$$

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