

# Probability-based Learning

## Sections 6.1, 6.2, 6.3

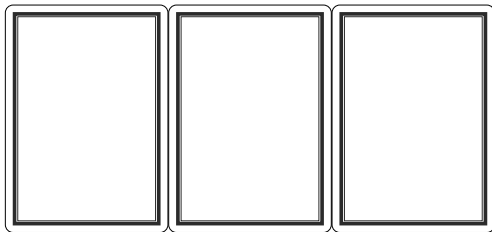
John D. Kelleher and Brian Mac Namee and Aoife D'Arcy

- 1 **Big Idea**
- 2 **Fundamentals**
  - Bayes' Theorem
  - Bayesian Prediction
  - Conditional Independence and Factorization
- 3 **Standard Approach: The Naive Bayes' Classifier**
  - A Worked Example
- 4 **Summary**

# Big Idea

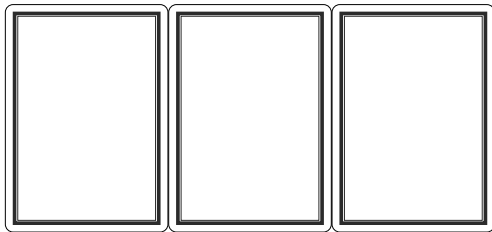


(a)

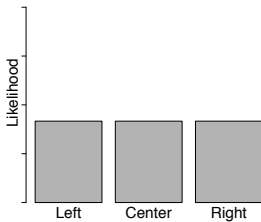


(b)

**Figure:** A game of *find the lady*

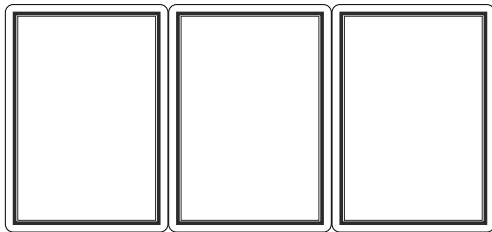


(a)

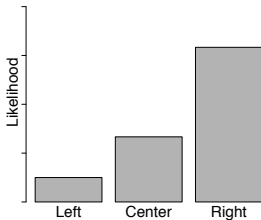


(b)

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

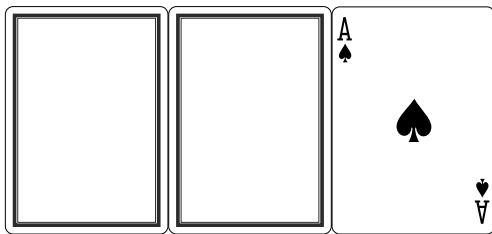


(a)

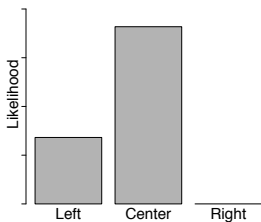


(b)

**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.

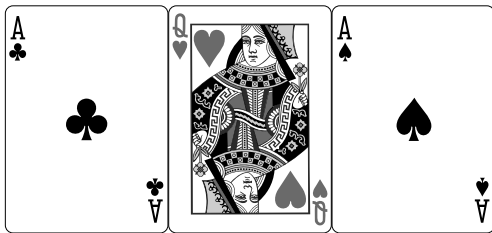


(a)



(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

oo oooooo

- 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.

oo oooooo

- 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.

oo oooooo

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

oo oooooo

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

## Bayes' Theorem

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

## Example

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

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











$$P(\mathbf{q}[1], \dots, \mathbf{q}[m]) =$$

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

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

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

$$P(\mathbf{q}[1]|t=l) \times P(\mathbf{q}[2]|\mathbf{q}[1], t=l) \times \dots$$

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





- 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:

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

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















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

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

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

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

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



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

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

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

- 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 → 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.

## 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:

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

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

- 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**.

## Conditional Independence and Factorization

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

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

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

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

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

$X$  and  $Y$  are **dependent**

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

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

$X$  and  $Y$  are **independent**

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

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

- 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:

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

## 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

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

- The joint probability distribution for the meningitis dataset.

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

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

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

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

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

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

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

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

- Calculate the factors from the data.

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

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

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

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

$$Factor_1 : \langle P(m) = 0.3 \rangle$$

$$Factor_2 : \langle P(h|m) = 0.6666, P(h|\neg m) = 0.7413 \rangle$$

$$Factor_3 : \langle P(f|m) = 0.3333, P(f|\neg m) = 0.4286 \rangle$$

$$Factor_4 : \langle P(v|m) = 0.6666, P(v|\neg m) = 0.5714 \rangle$$

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

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

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

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

- 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	?

## Conditional Independence and Factorization

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

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

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

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

- 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

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

## 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	arrear	none	own	false
6	arrear	none	own	true
7	current	none	own	false
8	arrear	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	arrear	none	own	false
16	current	none	own	false
17	arrear	coapplicant	rent	false
18	arrear	none	free	false
19	arrear	none	own	false
20	paid	none	own	false

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

**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'.

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

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMODATION	FRAUDULENT
paid	none	rent	?

## A Worked Example

$$P(fr) = 0.3$$

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

$$P(CH = 'paid' | fr) = 0.1666$$

$$P(CH = 'paid' | \neg fr) = 0.2857$$

$$P(GC = 'none' | fr) = 0.8334$$

$$P(GC = 'none' | \neg fr) = 0.8571$$

$$P(ACC = 'rent' | fr) = 0.3333$$

$$P(ACC = 'rent' | \neg fr) = 0.1429$$

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

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

CREDIT HISTORY	GUARANTOR/COAPPLICANT	ACCOMODATION	FRAUDULENT
paid	none	rent	?

## A Worked Example

$$P(fr) = 0.3$$

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

$$P(CH = 'paid' | fr) = 0.1666$$

$$P(CH = 'paid' | \neg fr) = 0.2857$$

$$P(GC = 'none' | fr) = 0.8334$$

$$P(GC = 'none' | \neg fr) = 0.8571$$

$$P(ACC = 'rent' | fr) = 0.3333$$

$$P(ACC = 'rent' | \neg fr) = 0.1429$$

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

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

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	<i>'false'</i>

# Summary

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

- 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.

- 1 **Big Idea**
- 2 **Fundamentals**
  - Bayes' Theorem
  - Bayesian Prediction
  - Conditional Independence and Factorization
- 3 **Standard Approach: The Naive Bayes' Classifier**
  - A Worked Example
- 4 **Summary**