Bayesian Learning I

Data Science Specialization Spring 2025

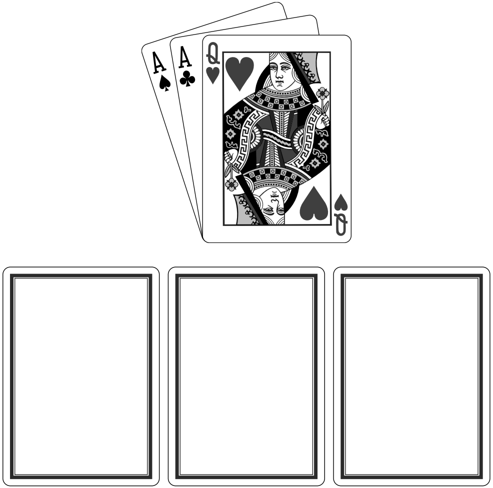
Jens Classen

Roskilde University

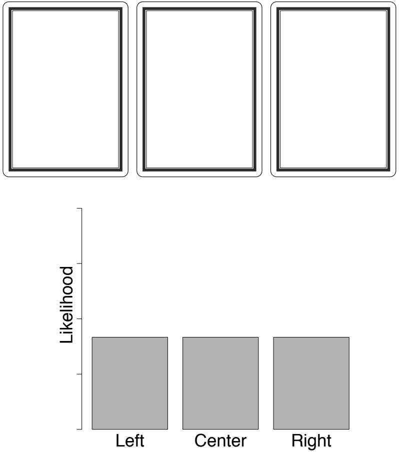
26.03.2025

Motivation

A Game of “Find the Lady”

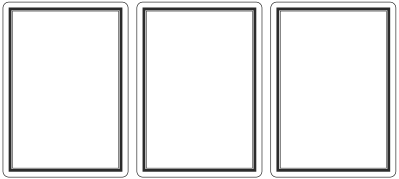


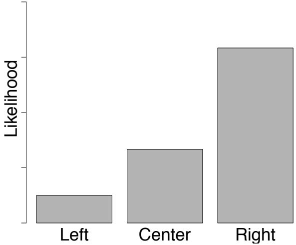
A Game of “Find the Lady”



Initial likelihoods with cards facing down.

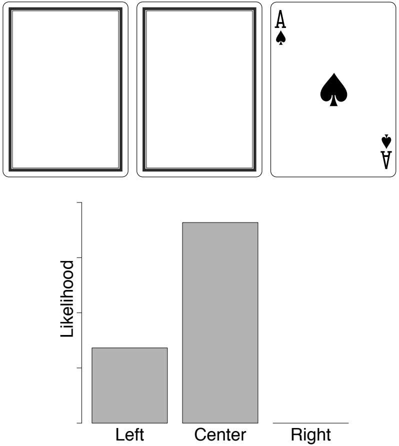
A Game of “Find the Lady”





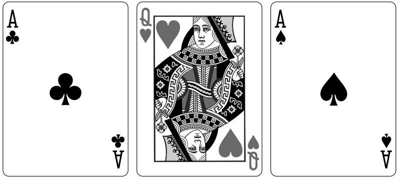
Observation: Dealer prefers right position (19x) over center (8x) and left (3x).

A Game of "Find the Lady"



Evidence: Wind blows over card on the right.

A Game of “Find the Lady”



Final positions of cards.

Big Idea

- use estimates of likelihoods to determine most likely prediction
- revise predictions based additional data/evidence

Fundamentals

Probability Distributions

$P(A)$ is the probability that A holds. P is called **probability function**.

We write $X = n$ to say that **random variable** (feature) X takes value n, taken from a discrete domain.

Example

Random variable *Weather* with values $\langle$ sunny, rain, cloudy, snow $\rangle$.

$$\begin{aligned} \text{P(Weather} = \text{sunny)} &= 0.7 \\ \text{P(Weather} = \text{rain)} &= 0.2 \\ \text{P(Weather} = \text{cloudy)} &= 0.08 \\ \text{P(Weather} = \text{snow)} &= 0.02 \end{aligned}$$

$\mathbf{P}(\text{Weather}) = (0.7; 0.2; 0.08; 0.02)$ is the **probability distribution** of Weather.

The sum of a probability distribution must equal 1.0.

Binary features: Instead of using true/false, we write $H = h$ and $H = \neg h$.

Multiple features: P(Weather = sunny, Sprinkler = on, Wet = wet) denotes **joint probability**.

Conditional Probabilities

Conditional probability: probability of a feature taking a specific value given the value of another feature.

Example: Rolling Dice

P(Roll=3) = 1/6. Let E = "Roll is divisible by 3."

Then we obtain the **conditional probability**:

$$P(\text{Roll}{=}3 \mid E) = 1/2.$$

E is also called the evidence and may represent background knowledge or observations.

Prior:	Probability before evidence.
Posterior:	Probability after the evidence.

Conditional Probabilities

$P(C \mid T) = 0.8$ is read as "the probability of C given T is 0.8."

Formal Definition: Conditional Probability

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \quad \text{or} \quad P(A, B) = P(A \mid B) \cdot P(B)$$

(Product Rule)

Dice Roll Example

$P(Roll = 3) = 1/6$. Let $E =$ "Roll is divisible by 3."

Then $P(E) = 2/6$ and $P(Roll = 3, E) = 1/6$. Thus $P(Roll = 3 \mid E) = \frac{1/6}{2/6} = 1/2$.

$\mathbf{P}(X, Y) = \mathbf{P}(X \mid Y) \cdot \mathbf{P}(Y)$ stands for a system of equations of the form:

$$P(X = x_i, Y = y_j) = P(X = x_i \mid Y = y_j) \cdot P(Y = y_j)$$

for all values x_i, y_j from the domains of X and Y.

Joint Distributions

The **joint probability distribution** $\mathbf{P}(X_1, X_2, \ldots X_n)$ assigns a probability to every combination of values for the random variables $X_1, X_2, \ldots X_n$.

Toothache-Cavity Example:

	$T = t$	$T = \neg t$
$C = c$	0.04	0.06
$C = \neg c$	0.01	0.89

- Values must add up to 1.0.
- From the table one can calculate all probabilities ("summing out"):
  - $P(t) = .04 + .01 = .05$
  - $P(c \mid t) = P(c, t)/P(t) = .04/.05 = .8$
- Note: Table grows exponentially in the number of variables.

Table 1: 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

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

Bayes Theorem

$$P(A, B) = P(A \mid B) \cdot P(B)$$
$$P(B, A) = P(B \mid A) \cdot P(A)$$

Since $P(A, B) = P(B, A)$, have $P(A \mid B) \cdot P(B) = P(B \mid A) \cdot P(A)$.

Bayes Theorem

$$P(B \mid A) = \frac{P(A \mid B) \cdot P(B)}{P(A)}$$

Again can write system of equations as

$$\mathbf{P}(\mathrm{Y} \mid \mathrm{X}) = \frac{\mathbf{P}(\mathrm{X} \mid \mathrm{Y}) \cdot \mathbf{P}(\mathrm{Y})}{\mathbf{P}(\mathrm{X})},$$

Often there is additional evidence E:

$$\mathbf{P}(\mathrm{Y} \mid \mathrm{X},\mathrm{E}) = \frac{\mathbf{P}(\mathrm{X} \mid \mathrm{Y},\mathrm{E}) \cdot \mathbf{P}(\mathrm{Y} \mid \mathrm{E})}{\mathbf{P}(\mathrm{X} \mid \mathrm{E})}$$

Example

Bayes Theorem is particular useful for doing diagnosis:
We want to determine $P(Cause \mid Effect)$, but $P(Effect \mid Cause)$ is easier to assess.

Patient Scenario

A doctor informs their patient that they have bad news and good news.

- bad news: patient tested positive for a serious disease, test is 99% accurate
  - probability of testing positive when having disease is 0.99
  - probability of testing negative when *not* having disease is 0.99
- good news: 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?

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

Have: $P(t \mid d) = 0.99$ and $P(d) = 0.0001$. But what about $P(t)$ (prior probability of the evidence)?

Normalisation

In the formula

$$\mathbf{P}(Y \mid X) = \frac{\mathbf{P}(X \mid Y) \cdot \mathbf{P}(Y)}{\mathbf{P}(X)}$$

the factor $1/\mathbf{P}(X)$ is only a normalising constant so that the right-hand side sums to 1.0 over all values of Y.

In the literature, one often uses the following form:

$$\mathbf{P}(Y \mid X) = \alpha \cdot \mathbf{P}(X \mid Y) \cdot \mathbf{P}(Y).$$

In practice one usually calculates the unnormalised case first, and then looks for an appropriate α.

Normalisation

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.

$$P(Y) = \sum_i P(Y \mid X_i) P(X_i)$$

This is also called "summing out" or "marginalization".

Patient Scenario, continued

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

$$P(d \mid t) = \frac{0.99 \times 0.0001}{0.0101} = 0.0098$$

☞ The probability to have the disease given a positive test is still very small ($< 1\%$)!

The Paradox of the False Positive

Possible mistake: **forgetting** to factor in the **prior**.

The Paradox of the False Positive

To make predictions about a rare event, the **model** has to be as **accurate** as the **prior** of the event is **rare**!

Otherwise: significant chance of **false positive** predictions!

Combining Evidence

How does one combine evidence consisting of several variables/features?

Generalized Bayes' Theorem

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

Here, $\mathbf{q}[1], \ldots, \mathbf{q}[m]$ is the query (evidence) in terms of descripitive features, and $t = l$ is the value of the target feature.

But how to compute $P(\mathbf{q}[1], \ldots, \mathbf{q}[m] \mid t = l)P(t = l)$ and $P(\mathbf{q}[1], \ldots, \mathbf{q}[m])$?

$$P(\mathbf{q}[2], \mathbf{q}[1]) = P(\mathbf{q}[2] \mid \mathbf{q}[1]) \times P(\mathbf{q}[1])$$
$$P(\mathbf{q}[3], \mathbf{q}[2], \mathbf{q}[1]) = P(\mathbf{q}[3] \mid \mathbf{q}[2], \mathbf{q}[1]) \times P(\mathbf{q}[2], \mathbf{q}[1])$$
$$P(\mathbf{q}[4], \mathbf{q}[3], \mathbf{q}[2], \mathbf{q}[1]) = P(\mathbf{q}[4] \mid \mathbf{q}[3], \mathbf{q}[2], \mathbf{q}[1]) \times P(\mathbf{q}[3], \mathbf{q}[2], \mathbf{q}[1])$$
$$\ldots$$

Chain Rule

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

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

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

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	?

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

Bayes Rule:

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

Reading off values from the dataset gives:

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

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

ID	H	F	V	M
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 chain rule (as exercise) gives:

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

Hence $P(m \mid h, \neg f, v) = 0.3333$.
Also, $P(\neg m \mid h, \neg f, v) = 1 - P(m \mid h, \neg f, v) = 0.6667$.
☞ Twice as probable to have **no** meningitis than to have it, despite headache and vomiting!

Bayesian Prediction

Task: Given query $\mathbf{q} = (\mathbf{q}[1], \ldots, \mathbf{q}[m])$ over descriptive features, predict value l for target feature t.

Bayesian MAP Prediction Model

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

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

Note: To determine most likely value for t, normalization is not needed.

Bayesian MAP Prediction Model (without normalization)

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

MAP = *maximum a posteriori*

Example

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

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

$$= \frac{0.6666 \times 0 \times 0 \times 0.3}{0.1} = 0$$

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

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

There is something odd about these results!

ID	H	F	V	M
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

Curse of Dimensionality

Curse of Dimensionality

- The more descriptive features, the more potential conditioning events.
- Every new feature requires exponentially many more examples to ensure there are enough instances matching the conditions!

- Probability of a patient with a headache and 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.

Stochastic 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:

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

Recall rules for **dependent** variables

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

Conditional Independence

- **Full independence**: rare!
- **Conditional independence**: more common!

X and Y are *conditionally independent given* Z iff:

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

Example

If MENINGITIS is given, then FEVER and HEADACHE are independent from each other:

$$P(F \mid H, M) = P(F \mid M)$$
$$P(F, H \mid M) = P(F \mid M) \times P(H \mid M)$$

In general, when a **cause** (disease) is known, then its **effects** (symptoms) can often be assumed to be 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

$$P(X, Y, Z \mid W) = P(X \mid W) \times P(Y \mid X, W) \times P(Z \mid Y, X, W) \times P(W)$$

With conditional independence

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

- 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:

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

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

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

- Calculate the factors from the data:

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

$$Factor_1 : < P(m) = 0.3 >$$
$$Factor_2 : < P(h \mid m) = 0.6666, P(h \mid \neg m) = 0.7413 >$$
$$Factor_3 : < P(f \mid m) = 0.3333, P(f \mid \neg m) = 0.4286 >$$
$$Factor_4 : < P(v \mid m) = 0.6666, P(v \mid \neg m) = 0.5714 >$$

$$Factor_1 : < P(m) = 0.3 >$$
$$Factor_2 : < P(h \mid m) = 0.6666, P(h \mid \neg m) = 0.7413 >$$
$$Factor_3 : < P(f \mid m) = 0.3333, P(f \mid \neg m) = 0.4286 >$$
$$Factor_4 : < P(v \mid m) = 0.6666, P(v \mid \neg m) = 0.5714 >$$

- Using the factors above calculate the probability of MENINGITIS=*'true'* for the following query.

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

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

$$\frac{0.6666 \times 0.3333 \times 0.3333 \times 0.3}{(0.6666 \times 0.3333 \times 0.3333 \times 0.3) + (0.7143 \times 0.4286 \times 0.4286 \times 0.7)} = 0.1948$$

$$Factor_1 : < P(m) = 0.3 >$$
$$Factor_2 : < P(h \mid m) = 0.6666, P(h \mid \neg m) = 0.7413 >$$
$$Factor_3 : < P(f \mid m) = 0.3333, P(f \mid \neg m) = 0.4286 >$$
$$Factor_4 : < P(v \mid m) = 0.6666, P(v \mid \neg m) = 0.5714 >$$

- Using the factors above calculate the probability of MENINGITIS=*'false'* for the same query.

HEADACHE	FEVER	VOMITING	MENINGITIS
true	true	false	?

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

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

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

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

- As before, the MAP prediction would be $\textsc{Meningitis}$ = *'false'*
- The posterior probabilities are not as extreme!

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 2: 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
$P(\mathrm{CH} = \textit{'none'} \mid fr)$	$=$	0.1666	$P(\mathrm{CH} = \textit{'none'} \mid \neg fr)$	$=$	0
$P(\mathrm{CH} = \textit{'paid'} \mid fr)$	$=$	0.1666	$P(\mathrm{CH} = \textit{'paid'} \mid \neg fr)$	$=$	0.2857
$P(\mathrm{CH} = \textit{'current'} \mid fr)$	$=$	0.5	$P(\mathrm{CH} = \textit{'current'} \mid \neg fr)$	$=$	0.2857
$P(\mathrm{CH} = \textit{'arrears'} \mid fr)$	$=$	0.1666	$P(\mathrm{CH} = \textit{'arrears'} \mid \neg fr)$	$=$	0.4286
$P(\mathrm{GC} = \textit{'none'} \mid fr)$	$=$	0.8334	$P(\mathrm{GC} = \textit{'none'} \mid \neg fr)$	$=$	0.8571
$P(\mathrm{GC} = \textit{'guarantor'} \mid fr)$	$=$	0.1666	$P(\mathrm{GC} = \textit{'guarantor'} \mid \neg fr)$	$=$	0
$P(\mathrm{GC} = \textit{'coapplicant'} \mid fr)$	$=$	0	$P(\mathrm{GC} = \textit{'coapplicant'} \mid \neg fr)$	$=$	0.1429
$P(\mathrm{ACC} = \textit{'own'} \mid fr)$	$=$	0.6666	$P(\mathrm{ACC} = \textit{'own'} \mid \neg fr)$	$=$	0.7857
$P(\mathrm{ACC} = \textit{'rent'} \mid fr)$	$=$	0.3333	$P(\mathrm{ACC} = \textit{'rent'} \mid \neg fr)$	$=$	0.1429
$P(\mathrm{ACC} = \textit{'free'} \mid fr)$	$=$	0	$P(\mathrm{ACC} = \textit{'free'} \mid \neg fr)$	$=$	0.0714

Table 3: 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	?

$P(fr)$	$=$	0.3	$P(\neg fr)$	$=$	0.7
$P(\mathrm{CH} = \textit{'paid'} \mid fr)$	$=$	0.1666	$P(\mathrm{CH} = \textit{'paid'} \mid \neg fr)$	$=$	0.2857
$P(\mathrm{GC} = \textit{'none'} \mid fr)$	$=$	0.8334	$P(\mathrm{GC} = \textit{'none'} \mid \neg fr)$	$=$	0.8571
$P(\mathrm{ACC} = \textit{'rent'} \mid fr)$	$=$	0.3333	$P(\mathrm{ACC} = \textit{'rent'} \mid \neg fr)$	$=$	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	?

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

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

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

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

The model is generalizing beyond the dataset!

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

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

Summary

Summary

$$P(t \mid \mathbf{d}) = \frac{P(\mathbf{d} \mid t) \times P(t)}{P(\mathbf{d})}$$

- Naive Bayes' classifier assumes that all descriptive features are conditionally independent from one another, given the target feature.
- Although often wrong, the assumption enables to maximally factorise the representation.
- Surprisingly, Naive Bayes' models often perform reasonably well, despite their "naivety".
- Naive Bayes' models are often used as baseline classifier to compare other methods against.

References

Reading

- Mitchell, T. M. (1997). *Machine Learning* (Vol. 1). McGraw-Hill New York. Chapter 6.
- Russell S. J. & Norvig P. (2020). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson. Chapter 13.
- Kelleher, Mac Namee, B., & D'Arcy, A. (2020). *Fundamentals of Machine Learning for Predictive Data Analytics Algorithms, Worked Examples, and Case Studies.* MIT Press. Chapter 6.

Acknowledgements

In addition to the authors above, some slides are adapted from or inspired by a course by Gerhard Lakemeyer (RWTH Aachen University).