Class 11 in-class problems, 18.05, Spring 2022

Concept questions

Concept question 1. Learning from experience

(a) *Which treatment would you choose?*

1. Treatment 1: cured 100% of patients in a trial.

2. Treatment 2: cured 95% of patients in a trial.

3. Treatment 3: cured 90% of patients in a trial.

Solution: No one correct answer.

(b) *Which treatment would you choose?*

1. Treatment 1: cured 3 out of 3 patients in a trial.

2. Treatment 2: cured 19 out of 20 patients treated in a trial.

3. Standard treatment: cured 90000 out of 100000 patients in clinical practice.

Solution: No one correct answer.

Board questions

Problem 1. Learning from data

- *A certain disease has a prevalence of 0.005.*
- *A screening test has 2% false positives an 1% false negatives.*

Suppose a random patient is screened and has a positive test.

(a) *Represent this information with a tree and use Bayes' theorem to compute the probabilities the patient does and doesn't have the disease.*

(b) *Identify the data, hypotheses, likelihoods, prior probabilities and posterior probabilities.*

(c) *Make a full likelihood table containing all hypotheses and possible test data.*

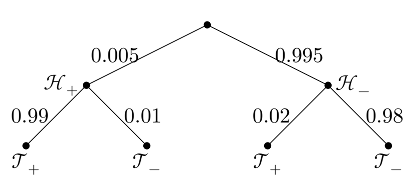
(d) *Redo the computation using a Bayesian update table. Match the terms in your table to the terms in your previous calculation.*

Solution: (a) Tree based Bayes computation

Let $\mathcal{H}_+$ mean the patient has the disease and $\mathcal{H}_-$ they don't.

Let $\mathcal{T}_+$: they test positive and $\mathcal{T}_-$ they test negative.

We can organize this in a tree:



Bayes' theorem says $P(\mathcal{H}_+ \,|\, \mathcal{T}_+) = \dfrac{P(\mathcal{T}_+ \,|\, \mathcal{H}_+)P(\mathcal{H}_+)}{P(\mathcal{T}_+)}$.

Using the tree, the total probability

$$\begin{aligned} P(\mathcal{T}_+) &= P(\mathcal{T}_+ \,|\, \mathcal{H}_+)P(\mathcal{H}_+) + P(\mathcal{T}_+ \,|\, \mathcal{H}_-)P(\mathcal{H}_-) \\ &= 0.99 \cdot 0.005 + 0.02 \cdot 0.995 = 0.02485 \end{aligned}$$

So,

$$\begin{aligned} P(\mathcal{H}_+ \,|\, \mathcal{T}_+) &= \frac{P(\mathcal{T}_+ \,|\, \mathcal{H}_+)P(\mathcal{H}_+)}{P(\mathcal{T}_+)} = \frac{0.99 \cdot 0.005}{0.02485} = 0.199 \\ P(\mathcal{H}_- \,|\, \mathcal{T}_+) &= \frac{P(\mathcal{T}_+ \,|\, \mathcal{H}_-)P(\mathcal{H}_-)}{P(\mathcal{T}_+)} = \frac{0.02 \cdot 0.995}{0.02485} = 0.801 \end{aligned}$$

The positive test greatly increases the probability of $\mathcal{H}_+$, but it is still much less probable than $\mathcal{H}_-$.

(b) Terminology

Data: The data are the results of the experiment. In this case, the positive test.

Hypotheses: The hypotheses are the possible answers to the question being asked. In this case they are $\mathcal{H}_+$ the patient has the disease; $\mathcal{H}_-$ they don't.

Likelihoods: The likelihood given a hypothesis is the probability of the data given that hypothesis. In this case there are two likelihoods, one for each hypothesis

$$P(\mathcal{T}_+ \,|\, \mathcal{H}_+) = 0.99 \qquad \text{and} \qquad P(\mathcal{T}_+ \,|\, \mathcal{H}_-) = 0.02.$$

We repeat: the likelihood is a probability **given** the hypothesis, **not** a probability of the hypothesis.

Prior probabilities of the hypotheses: The priors are the probabilities of the hypotheses prior to collecting data. In this case,

$$P(\mathcal{H}_+) = 0.005 \qquad \text{and} \qquad P(\mathcal{H}_-) = 0.995$$

Posterior probabilities of the hypotheses: The posteriors are the probabilities of the hypotheses **given** the data. In this case

$$P(\mathcal{H}_+ \,|\, \mathcal{T}_+) = 0.199 \qquad \text{and} \qquad P(\mathcal{H}_- \,|\, \mathcal{T}_+) = 0.801.$$

$$\underbrace{P(\mathcal{H}_+ \,|\, \mathcal{T}_+)}_{\text{Posterior}} = \frac{\overbrace{P(\mathcal{T}_+ \,|\, \mathcal{H}_+)}^{\text{Likelihood}} \cdot \overbrace{P(\mathcal{H}_+)}^{\text{Prior}}}{\underbrace{P(\mathcal{T}_+)}_{\text{Total probability of the data}}}$$

(c) Full likelihood table

The table holds likelihoods $P(\mathcal{D}|\mathcal{H})$ for every possible hypothesis and data combination.

hypothesis $\mathcal{H}$	likelihood $P(\mathcal{D}|\mathcal{H})$	
disease state	$P(\mathcal{T}_+|\mathcal{H})$	$P(\mathcal{T}_-|\mathcal{H})$
$\mathcal{H}_+$	0.99	0.01
$\mathcal{H}_-$	0.02	0.98

Notice in the table below that the $P(\mathcal{T}_+ \mid \mathcal{H})$ column is exactly the likelihood column in the Bayesian update table.

(d) Calculation using a Bayesian update table

$\mathcal{H}$ = hypothesis: $\mathcal{H}_+$ (patient has disease); $\mathcal{H}_-$ (they don't).

Data: $\mathcal{T}_+$ (positive screening test).

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{T}_+\mid\mathcal{H})$	$P(\mathcal{T}_+\mid\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\mid\mathcal{T}_+)$
$\mathcal{H}_+$	0.005	0.99	0.00495	0.199
$\mathcal{H}_-$	0.995	0.02	0.0199	0.801
total	1	NO SUM	$P(\mathcal{T}_+) = 0.02485$	1

Data $\mathcal{D} = \mathcal{T}_+$

Total probability: $P(\mathcal{T}_+)$ = sum of Bayes numerator column = 0.02485

Bayes' theorem: $P(\mathcal{H}\mid\mathcal{T}_+) = \dfrac{P(\mathcal{T}_+\mid\mathcal{H})P(\mathcal{H})}{P(\mathcal{T}_+)} = \dfrac{\text{likelihood} \times \text{prior}}{\text{total prob. of data}}$

Problem 2. Dice

I have five dice: 4-sided, 6-sided, 8-sided, 12-sided, 20-sided.

I pick one at random, roll it and report that the roll was a 13.

Goal: Find the probabilities the die is 4, 6, 8, 12 or 20 sided.

***(a)** Identify the hypotheses.*

***(b)** Make a likelihood table with columns for the data 'rolled a 13', 'rolled a 5' and 'rolled a 9'.*

***(c)** Make a Bayesian update table and compute the posterior probabilities that the chosen die is each of the five dice.*

***(d)** Same question if I had reported a 5.*

***(e)** Same question if I had reported a 9.*

Solution: (a) The hypotheses are: $\mathcal{H}_4$, the chosen die was 4-sided; $\mathcal{H}_6$, the chosen die was 6-sided; Likewise $\mathcal{H}_8$, $\mathcal{H}_{12}$, $\mathcal{H}_{20}$.

(b) The likelihoods for a roll of 5, 9 and 13 are

hypothesis $\mathcal{H}$	$P(5\mid\mathcal{H})$	$P(9\mid\mathcal{H})$	$P(13\mid\mathcal{H})$
$\mathcal{H}_4$	0	0	0
$\mathcal{H}_6$	1/6	0	0
$\mathcal{H}_8$	1/8	1/8	0
$\mathcal{H}_{12}$	1/12	1/12	0
$\mathcal{H}_{20}$	1/20	1/20	1/20

(c) $\mathcal{D}$ = 'rolled a 13'. So our likelihood column uses the $P(13|\mathcal{H})$ from part (b).

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
$\mathcal{H}_4$	1/5	0	0	0
$\mathcal{H}_6$	1/5	0	0	0
$\mathcal{H}_8$	1/5	0	0	0
$\mathcal{H}_{12}$	1/5	0	0	0
$\mathcal{H}_{20}$	1/5	1/20	1/100	1
total	1		1/100	1

The only possibility is the 20-sided die.

(d) $\mathcal{D}$ = 'rolled a 5'. So our likelihood column uses the $P(5|\mathcal{H})$ from part (b).

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
$\mathcal{H}_4$	1/5	0	0	0
$\mathcal{H}_6$	1/5	1/6	1/30	0.392
$\mathcal{H}_8$	1/5	1/8	1/40	0.294
$\mathcal{H}_{12}$	1/5	1/12	1/60	0.196
$\mathcal{H}_{20}$	1/5	1/20	1/100	0.118
total	1		0.085	1

$\mathcal{H}_4$ is impossible. The most probable hypothesis is $\mathcal{H}_6$.

(e) $\mathcal{D}$ = 'rolled a 9'. So our likelihood column uses the $P(9|\mathcal{H})$ from part (b).

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
$\mathcal{H}_4$	1/5	0	0	0
$\mathcal{H}_6$	1/5	0	0	0
$\mathcal{H}_8$	1/5	0	0	0
$\mathcal{H}_{12}$	1/5	1/12	1/60	0.625
$\mathcal{H}_{20}$	1/5	1/20	1/100	0.375
total	1		0.0267	1

The most probable hypothesis is $\mathcal{H}_{12}$.

Problem 3. Iterated updates

Suppose I rolled a 9 and then a 5.

(a) *Do the Bayesian update in two steps:*
Step 1: First update for the 9.
Step 2: Then update the update for the 5.

(b) *Do the Bayesian update in one step.*

That is, the data is $\mathcal{D}$ = '9 followed by 5'

Solution: **(a)** Tabular solution: two steps

$\mathcal{D}_1$ = 'rolled a 9', $\mathcal{D}_2$ = 'rolled a 5'

Bayes numerator$_1$ = likelihood$_1 \times$ prior.

Bayes numerator$_2$ = likelihood$_2 \times$ Bayes numerator$_1$

hyp.	prior	likel. 1	Bayes num. 1	likel. 2	Bayes num. 2	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}_1\|\mathcal{H})$	$* * *$	$P(\mathcal{D}_2\|\mathcal{H})$	$* * *$	$P(\mathcal{H}\|\mathcal{D}_1, \mathcal{D}_2)$
$\mathcal{H}_4$	1/5	0	0	0	0	0
$\mathcal{H}_6$	1/5	0	0	1/6	0	0
$\mathcal{H}_8$	1/5	0	0	1/8	0	0
$\mathcal{H}_{12}$	1/5	1/12	1/60	1/12	1/720	0.735
$\mathcal{H}_{20}$	1/5	1/20	1/100	1/20	1/2000	0.265
total	1				0.0019	1

(b) Tabular solution: one step

$\mathcal{D}$ = 'rolled a 9 then a 5'

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})$	$P(\mathcal{D}\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|\mathcal{D})$
$\mathcal{H}_4$	1/5	0	0	0
$\mathcal{H}_6$	1/5	0	0	0
$\mathcal{H}_8$	1/5	0	0	0
$\mathcal{H}_{12}$	1/5	1/144	1/720	0.735
$\mathcal{H}_{20}$	1/5	1/400	1/2000	0.265
total	1		0.0019	1

Problem 4. Probabilistic prediction (Probably won't get here till next time)

With the same setup as before let:

$\mathcal{D}_1$ = result of first roll, $\quad \mathcal{D}_2$ = result of second roll

(a) *Find $P(\mathcal{D}_1 = 5)$.*

(b) *Find $P(\mathcal{D}_2 = 4|\mathcal{D}_1 = 5)$.*

Solution: $\mathcal{D}_1$ = 'rolled a 5', $\quad \mathcal{D}_2$ = 'rolled a 4'

hyp.	prior	likel. 1	Bayes num. 1	post. 1	likel. 2	post. 1 × likel. 2
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}_1\|\mathcal{H})$	$* * *$	$P(\mathcal{H}\|\mathcal{D}_1)$	$P(\mathcal{D}_2\|\mathcal{H}, \mathcal{D}_1)$	$P(\mathcal{D}_2\|\mathcal{H}, \mathcal{D}_1)P(\mathcal{H}\|\mathcal{D}_1)$
$\mathcal{H}_4$	1/5	0	0	0	$*$	0
$\mathcal{H}_6$	1/5	1/6	1/30	0.392	1/6	$0.392 \cdot 1/6$
$\mathcal{H}_8$	1/5	1/8	1/40	0.294	1/8	$0.294 \cdot 1/40$
$\mathcal{H}_{12}$	1/5	1/12	1/60	0.196	1/12	$0.196 \cdot 1/12$
$\mathcal{H}_{20}$	1/5	1/20	1/100	0.118	1/20	$0.118 \cdot 1/20$
total	1		0.085	1		0.124

The law of total probability tells us $P(\mathcal{D}_1)$ is the sum of the Bayes numerator 1 column in the table: $\boxed{P(\mathcal{D}_1) = 0.085}$.

The law of total probability tells us $P(\mathcal{D}_2|\mathcal{D}_1)$ is the sum of the last column in the table:

$\boxed{P(\mathcal{D}_2|\mathcal{D}_1) = 0.124}$ **Solution:** $\mathcal{D}_1$ = 'rolled a 5', $\mathcal{D}_2$ = 'rolled a 4'

hyp.	prior	likel. 1	Bayes num. 1	post. 1	likel. 2	post. 1 × likel. 2
$\mathcal{H}$	$P(\mathcal{H})$	$P(\mathcal{D}_1\|\mathcal{H})$	$* * *$	$P(\mathcal{H}\|\mathcal{D}_1)$	$P(\mathcal{D}_2\|\mathcal{H},\mathcal{D}_1)$	$P(\mathcal{D}_2\|\mathcal{H},\mathcal{D}_1)P(\mathcal{H}\|\mathcal{D}_1)$
$\mathcal{H}_4$	1/5	0	0	0	$*$	0
$\mathcal{H}_6$	1/5	1/6	1/30	0.392	1/6	$0.392 \cdot 1/6$
$\mathcal{H}_8$	1/5	1/8	1/40	0.294	1/8	$0.294 \cdot 1/40$
$\mathcal{H}_{12}$	1/5	1/12	1/60	0.196	1/12	$0.196 \cdot 1/12$
$\mathcal{H}_{20}$	1/5	1/20	1/100	0.118	1/20	$0.118 \cdot 1/20$
total	1		0.085	1		0.124

The law of total probability tells us $P(\mathcal{D}_1)$ is the sum of the Bayes numerator 1 column in the table: $\boxed{P(\mathcal{D}_1) = 0.085}$.

The law of total probability tells us $P(\mathcal{D}_2|\mathcal{D}_1)$ is the sum of the last column in the table: $\boxed{P(\mathcal{D}_2|\mathcal{D}_1) = 0.124}$

Extra problems

Extra 1. Bayesian updating: terminology, trees, tables

I have a bag with one 4-sided die and 999 6-sided dice. I pick one at random and roll it. Suppose I get a 3.

Goal: find the probabilities the chosen die was 4-sided or 6-sided.

***(a)** Identify the hypotheses.*

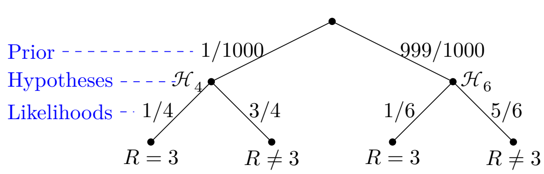
***(b)** Use Bayes' theorem to compute the posterior probabilities. Organize the computation using trees.*

***(c)** Connect all the Bayesian updating terminology with the parts of the computation.*

***(d)** Redo the computation using a Bayesian updating table.*

Solution: (a) The hypotheses are: $\mathcal{H}_4$, the chosen die was 4-sided and $\mathcal{H}_6$, the chosen die was 4-sided;

(b) Let R be the value of the roll. Here is the probability tree:



Bayes' theorem says $P(\mathcal{H}_4 \,|\, R = 3) = \dfrac{P(R = 3 \,|\, \mathcal{H}_4)P(\mathcal{H}_4)}{P(R = 3)}$. Likewise for $\mathcal{H}_6$

Using the tree, the total probability of $R = 3$.

$$P(R = 3) = P(R = 3 \,|\, \mathcal{H}_4)P(\mathcal{H}_4) + P(R = 3 \,|\, \mathcal{H}_6)P(\mathcal{H}_6) = 1/4000 + 999/6000 = 0.167$$

So,

$$P(\mathcal{H}_4 \,|\, R=3) = \frac{P(R=3 \,|\, \mathcal{H}_4)P(\mathcal{H}_4)}{P(R=3)} = \frac{1/4000}{0.167} \approx 0.0015$$
$$P(\mathcal{H}_6 \,|\, R=3) = \frac{P(R=3 \,|\, \mathcal{H}_6)P(\mathcal{H}_6)}{P(R=3)} = \frac{999/6000}{0.167} \approx 0.9985$$

The roll of 3 increases the probability of $\mathcal{H}_4$, but it is still much less probable than $\mathcal{H}_6$.

(c) Terminology

Data: The data are the results of the experiment. In this case, $R = 3$.

Hypotheses: The hypotheses are the possible answers to the question being asked. In this case they are $\mathcal{H}_4$ the die is 4-sided and $\mathcal{H}_6$ the die is 6-sided.

Likelihoods: The likelihood given a hypothesis is the probability of the data given that hypothesis. In this case there are two likelihoods, one for each hypothesis

$$P(R=3 \,|\, \mathcal{H}_4) = 1/4 \qquad \text{and} \qquad P(R=3 \,|\, \mathcal{H}_6) = 1/6.$$

We repeat: the likelihood is a probability **given** the hypothesis, **not** a probability of the hypothesis.

Prior probabilities of the hypotheses: The priors are the probabilities of the hypotheses prior to collecting data. In this case,

$$P(\mathcal{H}_4) = 1/1000 \qquad \text{and} \qquad P(\mathcal{H}_6) = 999/1000.$$

Posterior probabilities of the hypotheses: The posteriors are the probabilities of the hypotheses **given** the data. In this case

$$P(\mathcal{H}_4 \,|\, R=3) \approx 0.0015 \qquad \text{and} \qquad P(\mathcal{H}_6 \,|\, R=3) \approx 0.9985. \text{ (Computed below.)}$$

Posterior ↓ Likelihood ↓ Prior ↓

$$P(\mathcal{H}_4 \,|\, R=3) = \frac{P(R=3 \,|\, \mathcal{H}_4) \cdot P(\mathcal{H}_4)}{P(R=3)}$$

↗ Total probability of the data

(d) Calculation using a Bayesian update table

$\mathcal{H}$ = hypothesis: $\mathcal{H}_4$ (4-sided die); $\mathcal{H}_6$ (6-sided die).

Data: $R = 3$ (roll of 3).

hypothesis	prior	likelihood	Bayes numerator	posterior
$\mathcal{H}$	$P(\mathcal{H})$	$P(R=3\|\mathcal{H})$	$P(R=3\|\mathcal{H})P(\mathcal{H})$	$P(\mathcal{H}\|R=3)$
$\mathcal{H}_4$	1/1000	1/4	1/4000	$\frac{1/4000}{0.167} \approx 0.0015$
$\mathcal{H}_6$	999/1000	1/6	999/6000	$\frac{999/6000}{0.167} \approx 0.9985$
total	1	NO SUM	$P(R=3) = 0.167$	1

The Total probability of the data is $P(R = 3)$ = sum of Bayes numerator column = 0.167.

Bayes' theorem: $P(\mathcal{H}|R = 3) = \dfrac{P(R = 3|\mathcal{H})P(\mathcal{H})}{P(R = 3)} = \dfrac{\text{likelihood} \times \text{prior}}{\text{total prob. of data}}$

The posterior probabilities are identical to those from the tree based calculation.