

Outline

- Introduction
- [Bayesian Probability Theory](#)
- Sequence Prediction and Data Compression
- Decision Trees and Ensemble Learning
- Bayesian Networks

Medical Diagnosis

Suppose in a population only 8 in 1000 people have a certain cancer. We have a lab test for the cancer that returns

- a positive result in 98% of the cases in which the cancer is actually present; and
- a negative result in 97% of the cases in which the cancer is not present.

Suppose we now observe a new patient for whom the lab test returns a positive result. Should we diagnose the patient as having cancer or not?

Monty Hall 'Paradox'

Suppose you're on a game show, and you're given the choice of three doors: Behind one door is a car; behind the others, goats. You pick a door, say No. 1, and the host, who knows what's behind the doors, opens another door, say No. 3, which has a goat. He then says to you, "Do you want to pick door No. 2?" Is it to your advantage to switch your choice?

Bayesian Probability Theory

- Probability denotes degree of belief.
- Prescribes a rational approach to changing one's belief upon seeing new evidence.
- A central concept is the Bayes Rule, named after Reverend Thomas Bayes (1702-1761).

Bayes Rule

Let A and B be events/propositions, we have

$$\Pr(A|B) = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

Derivation is simple:

- By definition of conditional probability,

$$\Pr(A|B) = \frac{\Pr(A \wedge B)}{\Pr(B)} \text{ and } \Pr(B|A) = \frac{\Pr(A \wedge B)}{\Pr(A)}.$$

- Thus $\Pr(A|B)\Pr(B) = \Pr(A \wedge B) = \Pr(B|A)\Pr(A)$
- Dividing by $\Pr(B)$ yields

$$\Pr(A|B) = \frac{\Pr(A \wedge B)}{\Pr(B)} = \frac{\Pr(B|A)\Pr(A)}{\Pr(B)}.$$

Application to Medical Diagnosis

Prior: $\Pr(\text{cancer}) = 0.008$ $\Pr(\neg\text{cancer}) = 0.992$

Lab test:

$$\Pr(\text{posTest}|\text{cancer}) = 0.98 \quad \Pr(\text{negTest}|\text{cancer}) = 0.02$$

$$\Pr(\text{posTest}|\neg\text{cancer}) = 0.03 \quad \Pr(\text{negTest}|\neg\text{cancer}) = 0.97$$

Posterior probabilities:

$$\begin{aligned} \Pr(\text{cancer}|\text{posTest}) &= \Pr(\text{posTest}|\text{cancer})\Pr(\text{cancer})/\Pr(\text{posTest}) \\ &= 0.98 \cdot 0.008 / \Pr(\text{posTest}) = 0.0078 \cdot K \end{aligned}$$

Application to Medical Diagnosis

Prior: $\Pr(\text{cancer}) = 0.008$ $\Pr(\neg\text{cancer}) = 0.992$

Lab test:

$$\Pr(\text{posTest}|\text{cancer}) = 0.98 \quad \Pr(\text{negTest}|\text{cancer}) = 0.02$$

$$\Pr(\text{posTest}|\neg\text{cancer}) = 0.03 \quad \Pr(\text{negTest}|\neg\text{cancer}) = 0.97$$

Posterior probabilities:

$$\begin{aligned} \Pr(\text{cancer}|\text{posTest}) &= \Pr(\text{posTest}|\text{cancer})\Pr(\text{cancer})/\Pr(\text{posTest}) \\ &= 0.98 \cdot 0.008 / \Pr(\text{posTest}) = 0.0078 \cdot K \end{aligned}$$

$$\begin{aligned} \Pr(\neg\text{cancer}|\text{posTest}) &= \Pr(\text{posTest}|\neg\text{cancer})\Pr(\neg\text{cancer})/\Pr(\text{posTest}) \\ &= 0.03 \cdot 0.992 / \Pr(\text{posTest}) = 0.0298 \cdot K \end{aligned}$$

A Second Test

New prior after first test: $\Pr(\text{cancer}) = 0.207$ $\Pr(\neg\text{cancer}) = 0.793$

Second lab test:

$$\Pr(\text{posTest}_2|\text{cancer}) = 0.8 \quad \Pr(\text{negTest}_2|\text{cancer}) = 0.2$$

$$\Pr(\text{posTest}_2|\neg\text{cancer}) = 0.1 \quad \Pr(\text{negTest}_2|\neg\text{cancer}) = 0.9$$

Posterior probabilities after second positive test:

$$\begin{aligned} \Pr(\text{cancer}|\text{posTest}_2) &= \Pr(\text{posTest}_2|\text{cancer})\Pr(\text{cancer})/\Pr(\text{posTest}_2) \\ &= 0.8 \cdot 0.207 / \Pr(\text{posTest}_2) = 0.1656 \cdot K_2 \end{aligned}$$

$$\begin{aligned} \Pr(\neg\text{cancer}|\text{posTest}_2) &= \Pr(\text{posTest}_2|\neg\text{cancer})\Pr(\neg\text{cancer})/\Pr(\text{posTest}_2) \\ &= 0.1 \cdot 0.793 / \Pr(\text{posTest}_2) = 0.079 \cdot K_2 \end{aligned}$$

Some Lessons

- That's why doctors order multiple tests.
- Probabilistic reasoning can appear counter-intuitive (in other words, we are not very good with probabilistic reasoning).

Application to Monty Hall Problem

Let C_i denotes "car is behind door i ".

Prior probability: $\Pr(C_1) = \Pr(C_2) = \Pr(C_3) = \frac{1}{3}$

Suppose we pick door 1 and the host then reveals a goat behind door 2 (event denoted by H_2).

$$\Pr(C_1|H_2) = \frac{\Pr(H_2|C_1)\Pr(C_1)}{\Pr(H_2)} = \frac{1/2 \cdot 1/3}{\Pr(H_2)} = \frac{1}{\Pr(H_2)} \frac{1}{6}$$

Application to Monty Hall Problem

Let C_i denotes "car is behind door i ".

Prior probability: $\Pr(C_1) = \Pr(C_2) = \Pr(C_3) = \frac{1}{3}$

Suppose we pick door 1 and the host then reveals a goat behind door 2 (event denoted by H_2).

$$\Pr(C_1|H_2) = \frac{\Pr(H_2|C_1)\Pr(C_1)}{\Pr(H_2)} = \frac{1/2 \cdot 1/3}{\Pr(H_2)} = \frac{1}{\Pr(H_2)} \frac{1}{6}$$

$$\Pr(C_3|H_2) = \frac{\Pr(H_2|C_3)\Pr(C_3)}{\Pr(H_2)} = \frac{1 \cdot 1/3}{\Pr(H_2)} = \frac{1}{\Pr(H_2)} \frac{1}{3}$$

Application to Monty Hall Problem

Let C_i denotes "car is behind door i ".

Prior probability: $\Pr(C_1) = \Pr(C_2) = \Pr(C_3) = \frac{1}{3}$

Suppose we pick door 1 and the host then reveals a goat behind door 2 (event denoted by H_2).

$$\Pr(C_1|H_2) = \frac{\Pr(H_2|C_1)\Pr(C_1)}{\Pr(H_2)} = \frac{1/2 \cdot 1/3}{\Pr(H_2)} = \frac{1}{\Pr(H_2)} \frac{1}{6}$$

$$\Pr(C_3|H_2) = \frac{\Pr(H_2|C_3)\Pr(C_3)}{\Pr(H_2)} = \frac{1 \cdot 1/3}{\Pr(H_2)} = \frac{1}{\Pr(H_2)} \frac{1}{3}$$

Since $\Pr(C_3|H_2) > \Pr(C_1|H_2)$, we should switch!

Bayes Rule Applied to Learning

- Let H be a class of functions.
- Suppose we have seen data D .
- Given $h \in H$, what is $\Pr(h|D)$, i.e. the posterior probability of h given that we have seen the data D ?

Bayes Rule Applied to Learning

- Let H be a class of functions.
- Suppose we have seen data D .
- Given $h \in H$, what is $\Pr(h|D)$, i.e. the posterior probability of h given that we have seen the data D ?

By Bayes rule,

$$\Pr(h|D) = \frac{\Pr(D|h) \Pr(h)}{\Pr(D)}$$

Bayes Rule Applied to Learning

- Let H be a class of functions.
- Suppose we have seen data D .
- Given $h \in H$, what is $\Pr(h|D)$, i.e. the posterior probability of h given that we have seen the data D ?

By Bayes rule,

$$\Pr(h|D) = \frac{\Pr(D|h) \Pr(h)}{\Pr(D)}$$

$\Pr(h)$ is the prior probability of h

$\Pr(D|h)$ is the likelihood of h

Bayes Rule Applied to Learning

By Bayes rule,

$$\Pr(h|D) = \frac{\Pr(D|h) \Pr(h)}{\Pr(D)}$$

Bayes rule says posterior probability of h is proportional to prior probability of h times the likelihood of h .

$\Pr(h)$ incorporates our background knowledge

$\Pr(D|h)$ can be simplified by making assumptions on the data-generation process.

For example, if $D = [(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)]$ are generated independently at random, then

$$\Pr(D|h) = \prod_{i=1}^n \Pr((x_i, y_i)|h) = \prod_{i=1}^n \Pr(h(x_i) = y_i).$$

MAP Estimator

A learning algorithm:

Input: H , D , and prior probability $\Pr(h)$

1. For each function $h \in H$, calculate the posterior

$$\Pr(h|D) = \frac{\Pr(D|h)\Pr(h)}{\Pr(D)}.$$

2. Output the function h_{MAP} with the highest posterior

$$h_{MAP} = \arg \max_{h \in H} \Pr(h|D).$$

MAP = Maximum a posteriori

Relationship to MDL Principle

- MDL = Minimum Description Length
- A formal version of Occam's razor: choose the shortest explanation for the observed data.
- We have

$$\begin{aligned} h_{MAP} &= \arg \max_{h \in H} \Pr(h)\Pr(D|h) \\ &= \arg \max_{h \in H} \log_2 \Pr(h) + \log_2 \Pr(D|h) \\ &= \arg \min_{h \in H} -\log_2 \Pr(h) - \log_2 \Pr(D|h) \end{aligned}$$

- The last line can be interpreted as short hypotheses are preferred.

A Short Detour on Information Theory

- Consider the problem of designing a code to transmit messages drawn at random, where the probability of getting message i is $\Pr(i)$.
- We want a code that minimises the expected number of bits we need to transmit.
- This implies assigning shorter codes to messages that are more probable.
- Shannon and Weaver (1949) showed that the optimal code assigns $-\log_2 \Pr(i)$ bits to encode message i .

Relationship to MDL Principle

We have

$$h_{MAP} = \arg \min_{h \in H} -\log_2 \Pr(h) - \log_2 \Pr(D|h)$$

Information-theoretic interpretation:

- $-\log_2 \Pr(h)$ is the code length of h under the optimal code (wrt the prior) for hypothesis space H ;
- $-\log_2 \Pr(D|h)$ is the code length of the observed data D under the optimal code wrt conditional probability of data given hypothesis h ,

Bayes Optimal Estimator

- The hypothesis h_{MAP} is not the best estimator!

Bayes Optimal Estimator

- The hypothesis h_{MAP} is not the best estimator!
- Consider $H = \{h_1, h_2, h_3\}$.
- Suppose

$$\Pr(h_1|D) = 0.4 \quad \Pr(h_2|D) = 0.3 \quad \Pr(h_3|D) = 0.3.$$

- Given x , suppose further that

$$h_1(x) = \text{pos} \quad h_2(x) = \text{neg} \quad h_3(x) = \text{neg}.$$

- We have $h_{MAP}(x) = h_1(x) = \text{pos}$.
- But taking h_1, h_2, h_3 into account, x is positive with probability 0.4 and negative with probability 0.6.

Bayes Optimal Estimator

The best estimate of the label of x given data D and function class H is

$$\arg \max_{y \in Y} \sum_{h_i \in H} \Pr(h_i|D) \mathbb{I}(h_i(x) = y) \quad (3)$$

- Any system that predicts using (3) is called a Bayes optimal estimator.
- On average, no other prediction method that uses the same function class H and same prior can outperform a Bayes optimal estimator.
- A Bayes optimal estimator may not be in H itself.

Bayes, Occam, Epicurus

The best estimate of the label of x given data D and function class H is

$$\arg \max_{y \in Y} \sum_{h_i \in H} \Pr(h_i|D) \mathbb{I}(h_i(x) = y) \quad (4)$$

- Epicurus' (342 BC – 270 BC) principle of multiple explanations:
keep all hypotheses that are consistent with the data.
- Occam's razor (1285 – 1349):
entities should not be multiplied beyond necessity.
Widely understood to mean: Among all hypotheses consistent with the observations, choose the simplest.
- The Bayes optimal estimator strikes a balance between the two principles.

Problems with Bayesian Analysis

- Prescribed solution is often computationally intractable.
- Choosing a good prior $\Pr(h)$ is not always straightforward.

Another Problem (Homework)

You have been called to jury duty in a town where there are two taxi companies, Green Cabs Ltd. and Blue Taxi Inc. Blue Taxi uses cars painted blue; Green Cabs uses green cars. Green Cabs dominates the market, with 85% of the taxies on the road. On a misty winter night a taxi sideswiped another car and drove off. A witness says it was a blue cab.

The witness is tested under conditions like those on the night of the accident, and 80% of the time she correctly reports the colour of the cab that is seen.

What is the probability that the taxi that caused the accident is blue?

Lecture Review

- Bayes rule is a useful tool in probability-type problems.
- The maximum a posteriori (MAP) estimator can be interpreted as the shortest description of the data.
- A weighted average of the models (Bayesian optimal estimator) is better than the MAP estimator.

