Inference

Christos Dimitrakakis

June 5, 2024

Outline

Set theory

- First, consider some universal set Ω.
- A set A is a collection of points x in Ω.
- $\{x \in \Omega : f(x)\}$: the set of points in Ω with the property that $f(x)$ is true.

Unary operators

- $\neg A = \{x \in \Omega : x \notin A\}$.

Binary operators

- $A \cup B$ if $\{x \in \Omega : x \in A \vee x \in B\}$ - (c.f. $A \vee B$)
- $A \cap B$ if $\{x \in \Omega : x \in A \wedge x \in B\}$ - (c.f. $A \wedge B$)

Binary relations

- $A \subset B$ if $x \in A \Rightarrow x \in B$ - (c.f. $A \implies B$)
- $A = B$ if $x \in A \Leftrightarrow x \in B$ - (c.f. $A \Leftrightarrow B$)

The inference problem

- Given statements $A_1, \ldots, A_n$ we know to be true (i.e. a knowledge base), is another statement B true?

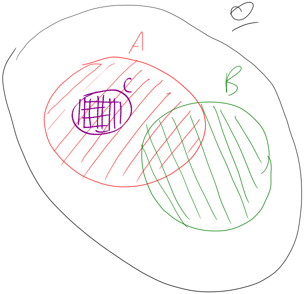
The following statements are equivalent:

- $A \implies B$ iff $(A \cap \neg B) = \emptyset$.
- $A \implies B$ iff $A \subset B$.

In addition

- If $(A \Rightarrow B) \wedge A$ then B.
- If $(A \wedge B)$ then A.

Illustration



$(A \mid C) =$

infered known

$(B \mid C) =$

$(C \mid A) =$

$(A \wedge B \mid C)$

Events as sets

The universe and random outcomes

- The Ω contains all events that can happen.
- When something happens, we observe an element $\omega \in \Omega$.

Events in the universe

- An event is true if $\omega \in A$, and false if $\omega \notin A$.
- The negative event $\neg A = \Omega \setminus A$ is the set
- The possible events are a collection of subsets Σ of Ω so that

(i) $\Omega \in \Sigma$, (ii) $A, B \in \Sigma \Rightarrow A \cup Bin\Sigma$ (iii) $A \in \Sigma \Rightarrow \neg A \in \Sigma$

Example: Traffic violation

- A car is moving with speed $\omega \in [0, \infty)$ in front of the speed camera.
- $A_0 = [0, 50]$: below the speed limit
- $A_1 = (50, 60]$: low fine
- $A_2 = (60, \infty]$: high fine
- $A_3 = (100, \infty)$: Suspension of license
- All combinations of the above events are interesting.

Probability fundamentals

Probability measure P

Probability can be seen as an area-like function assigning a likelihood to sets.

- $P : \Sigma \to [0, 1]$ gives the likelihood $P(A)$ of an event $A \in \Sigma$.
- $P(\Omega) = 1$
- For $A, B \subset \Omega$, if $A \cap B = \emptyset$ then $P(A \cup B) = P(A) + P(B)$.

Marginalisation

If $A_1, \ldots, A_n \subset \Omega$ are a partition of Ω

$$P(B) = \sum_{i=1}^{n} P(B \cap A_i).$$

Conditional probability

Definition (Conditional probability)

The conditional probability of an event A given an event B is defined as

$$P(A|B) \triangleq \frac{P(A \cap B)}{P(B)}$$

The above definition requires $P(B)$ to exist and be positive.

Conditional probabilities as a collection of probabilities

More generally, we can define conditional probabilities as simply a collection of probability distributions:

$$\{P_\theta : \theta \in \Theta\},$$

where Θ is indexing possible values of θ.

- θ is sometimes called the model or parameter

The theorem of Bayes

Theorem (Bayes's theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The theorem of Bayes

Theorem (Bayes's theorem)

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

The general case

If $A_1, \ldots, A_n$ are a partition of Ω, meaning that they are mutually exclusive events (i.e. $A_i \cap A_j = \emptyset$ for $i \neq j$) such that one of them must be true (i.e. $\bigcup_{i=1}^n A_i = \Omega$), then

$$P(B) = \sum_{i=1}^{n} P(B|A_i)P(A_i)$$

and

$$P(A_j|B) = \frac{P(B|A_j)}{\sum_{i=1}^n P(B|A_i)P(A_i)}$$

Independence

Independent events $A \perp\!\!\!\perp B$

A, B are independent iff $P(A \cap B) = P(A)P(B)$.

Conditional independence $A \perp\!\!\!\perp B \mid C$

A, B are conditionally independent given C iff
$P(A \cap B|C) = P(A|C)P(B|C)$.

Bayes's theorem

As a conditional measure

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

As a causal explanation

$$\mathbb{P}(\text{cause} \mid \text{effect}) = \frac{\mathbb{P}(\text{effect} \mid \text{cause})\,\mathbb{P}(\text{cause})}{\mathbb{P}(\text{effect})}$$

As model inference

- Prior $\beta(\theta)$
- Model class $\{P_\theta(\beta) : \theta \in \Theta\}$
- Data x

$$\beta(\theta \mid x) = \frac{P_\theta(x)\beta(\theta)}{\mathbb{P}_\beta(x)} = \frac{P_\theta(x)\beta(x)}{\sum_{\theta' \in \Theta} P_{\theta'}(x)\beta(\theta')}$$

Example: Naive Bayes models

Sometimes we observe multiple effects that have a common cause, but which are otherwise independent:

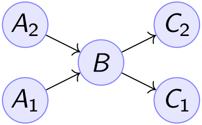
$$\mathbb{P}(\text{effect}_1, \ldots \text{effect}_n \mid \text{cause}) = \prod_{i=1}^{n} \mathbb{P}(\text{effect}_i \mid \text{cause})$$

Naive Bayes model

- Observations $(\boldsymbol{x}_t, y_t)_{t=1}^T$ with $\boldsymbol{x}_t = (x_{t,1}, \ldots, x_{t,n})$.
- Probability models $P_\mu(y \mid \boldsymbol{x}) = \prod_{i=1}^n P_\mu(y \mid x_i)$.

Conditional independence

For any set of events $A_1, A_2, A_3, \ldots$, we can write their co-occurence probability as $\prod_i P(A_i \mid \cap A_1 \cap A_2 \cap \cdots \cap A_{i-1})$. However, we can use a Bayesian network to define conditional independence structures.



If A is a parent of B and C is a child of B, and there are no other paths from A to C then the following conditional independence holds:

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

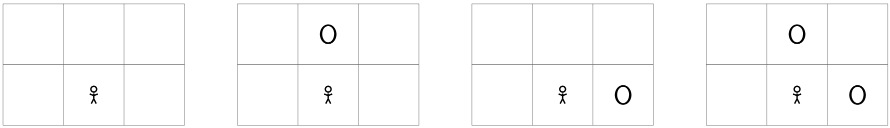
i.e. C is conditionally independent of A given B.

Conditional probability tables

We can now write the distribution of the above example as

$$P(B, C_1, C_2) = P(A_1)P(A_2)P(B|A_1 \cap A_2)P(C_1|B)P(C_2|B).$$

Example: Wumpus world



Details

- Probability of each world A_i being true: $1/4$
- Probability of each hole generating a breeze: $P(B_1|A_2 \cup A_4) = P(B_2|A_3 \cup A_4)$ with B_1, B_2 conditionally independent given A.

Questions

- What is the probability of feeling a breeze $B = B_1 \cup B_2$ in each world?
- What is the probability of a hole above if you feel a breeze?
- What is the probability of a hole above f you don't feel a breeze?

Example: The k-meteorologists problem

- A set of stations $\mathcal{M}$, with $\mu \in \mathcal{M}$ making weather predictions:

$$P_\mu(x_{t+1} \mid x_1, \ldots, x_t)$$

- A prior probability $P(\mu)$ on the stations.
- The marginal probability

$$P(x_1, \ldots, x_t) = \sum_{\mu \in \mathcal{M}} P_\mu(x_1, \ldots, x_t) P(\mu)$$

- The posterior probability

$$\begin{aligned} P(\mu \mid x_1, \ldots, x_t) &= \frac{P_\mu(x_1, \ldots, x_t) P(\mu)}{P(x_1, \ldots, x_t)} = \frac{\prod_{i=1}^{t} P_\mu(x_t \mid x_1, \ldots, x_{t-1}) P(\mu)}{P(x_1, \ldots, x_t)} \\ &= \frac{P_\mu(x_t \mid x_1, \ldots, x_{t-1}) P(\mu \mid x_1, \ldots, x_{t-1})}{P(x_t \mid x_1, \ldots, x_{t-1})} \end{aligned}$$

- The marginal posterior probability

$$P(x_{t+1} \mid x_1, \ldots, x_t) = \sum_{\mu \in \mathcal{M}} P_\mu(x_{t+1} \mid x_1, \ldots, x_t) P(\mu \mid x_1, \ldots, x_t)$$

Preferences

Types of rewards

- For e.g. a student: Tickets to concerts.
- For e.g. an investor: A basket of stocks, bonds and currency.
- For everybody: Money.

Preferences among rewards

For any rewards $x, y \in R$, we either

- (a) Prefer x at least as much as y and write $x \preceq^* y$.
- (b) Prefer x not more than y and write $x \succeq^* y$.
- (c) Prefer x about the same as y and write $x \eqsim^* y$.
- (d) Similarly define $\succ^*$ and $\prec^*$

Utility and Cost

Utility function

To make it easy, assign a utility $U(x)$ to every reward through a utility function $U : R \to \mathbb{R}$.

Utility-derived preferences

We prefer items with higher utility, i.e.

- (a) $U(x) \geq U(y) \Leftrightarrow x \succeq^* y$
- (b) $U(x) \leq U(y) \Leftrightarrow y \succeq^* x$

Cost

It is sometimes more convenient to define a cost function $C : R \to \mathbb{R}$ so that we prefer items with lower cost, i.e.

- $C(x) \geq C(y) \Leftrightarrow y \succeq^* x$

Random outcomes

Choosing among rewards

-[A] Bet 10 CHF on black -[B] Bet 10 CHF on 0 -[C] Bet nothing What is the reward here?

Choosing among trips

-[A] Taking the car to Zurich (50' without delays, 80' with delays) -[B] Taking the train to Zurich (60' without delays) What is the reward here?

Random rewards

- Each gamble gives us different rewards with different probabilities.
- These rewards are then random
- For simplicity, we assign a real-valued utility to outcomes. This is a random variable

Random variables

A random variable $f : \Omega \to \mathbb{R}$ is a real-valued function, with $\omega \sim P$.

The distribution of f

The probability that f lies in some subset $A \subset \mathbb{R}$ is

$$P_f(A) \triangleq P(\{\omega \in \Omega : f(\omega) \in A\}),$$

and we write $f \sim P_f$.

Shorthands for RV

- For RVs $f : \Omega \to \mathbb{R}$, we can write $P(f \in A)$ to mean $P_f(A)$.
- For RVs $f : \Omega \to X$, where X is a finite set e.g. $\{1, 2, \ldots, n\}$, we can write $P(f = x)$ for any $x \in X$.

Independence

Two RVs f, g are independent in the same way that events are independent:

$$P(f \in A \wedge g \in B) = P(f \in A)P(g \in B) = P_f(A)P_g(B).$$

In that sense, $f \sim P_f$ and $g \sim P_g$.

Expectation

For any real-valued random variable $f : \Omega \to \mathbb{R}$, the expectation with respect to a probability measure P is

$$\mathbb{E}_P(f) = \sum_{\omega \in \Omega} f(\omega) P(\omega).$$

When Ω is continuous, we can use a density p

$$\mathbb{E}_P(f) = \int_\Omega f(\omega) p(\omega) d\omega.$$

Linearity of expectations

For any RVs x, y:

$$\mathbb{E}_P(x + y) = \mathbb{E}_P(x) + \mathbb{E}_P(y)$$

Multiple variables

The joint distribution $P(x, y)$

For two (or more) RVs $x : \Omega \to \mathbb{R}$, and $y : \Omega \to \mathbb{R}$, this is a shorthand for the distribution of $(x(\omega), y(\omega))$ when $\omega \sim P$. We can also use $P(x = i, y = j)$ for the probability that the two variables assume the values i, j respectively.

Independence

If x, y are independent RVs then $P(x, y) = P_x(x) P_y(y)$.

Correlation

If x, y are not correlated then $\mathbb{E}_P(xy) = \mathbb{E}(x)\,\mathbb{E}(y)$.

IID (Independent and Identically Distributed) random variables

A sequence x_t of r.v.s is IID if $x_t \sim P$ so that

$$(x_1, \ldots, x_t, \ldots, x_T) \sim P^T$$

i.e. a T-length sample is drawn from the product distribution $P^T = P \times P \times \cdots \times P$.

Conditional expectation

The conditional expectation of a random variable $f : \Omega \to \mathbb{R}$, with respect to a probability measure P conditioned on some event B is simply

$$\mathbb{E}_P(f|B) = \sum_{\omega \in \Omega} f(\omega) P(\omega|B).$$

Conditional expectations are similar to conditional probabilities.

Conditional probabilities of RVs

Similarly to the notation over sets,

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

when dealing with RVs, it is common to use the notation

$$P(x, y) = P(x|y)P(y)$$

This equation works for all possible values of x, y e.g.

$$P(x = 1, y = 0) = P(x = 1|y = 0)P(y = 0)$$

which then denotes the probability msas of each

Expected utility

Actions, outcomes and utility

In this setting, we obtain random outcomes that depend on our actions.

- Actions $a \in A$
- Outcomes $\omega \in \Omega$.
- Probability of outcomes $P(\omega \mid a)$
- Utility $U : \Omega \to \mathbb{R}$

Expected utility

The expected utility of an action is:

$$\mathbb{E}_P[U \mid a] = \sum_{\omega \in \Omega} U(\omega) P(\omega \mid a).$$

The expected utility hypothesis

We prefer a to a' if and only if

$$\mathbb{E}_P[U \mid a] \geq \mathbb{E}_P[U \mid a']$$

The St-Petersburg Paradox

The game

If you give me x CHF, then I promise to (a) Throw a fair coin until it comes heads. (b) If it does so after T throws, then I will give you 2^T CHF.

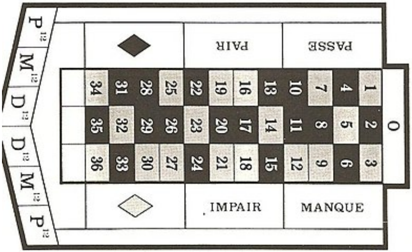
The question

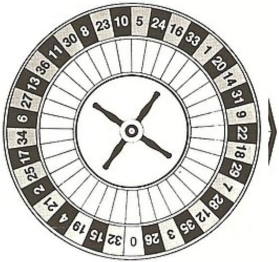
- How much x are you willing to pay to play?
- Given that the expected amount of money is infinite, why are you only willing to pay a small x?

Example: Betting

In this example, probabilities reflect actual randomness

Choice	Win Probability p	Payout w	Expected gain
Don't play	0	0	0
Black	18/37	2	
Red	18/37	2	
0	1/37	36	
1	1/37	36	





What are the expected gains for these bets?

Example: Route selection

- In this example, probabilities reflect subjective beliefs

Choice	Best time	Chance of delay	Delay amount	Expected time
Train	80	5%	5	
Car, route A	60	50%	30	
Car, route B	70	10%	10	

Example: Estimation

- In this example, probabilities are calculated starting from subjective beliefs

Mean-Square Estimation

If we want to guess $\hat{\mu}$, and we knew that $\mu \sim P$, then the guess

$$\hat{\mu} = \mathbb{E}_P(\mu) = \arg\min_{\hat{\mu}} \mathbb{E}_P[(\mu - \hat{\mu})^2]$$

Example: The k-meteorologists problem

- A set of stations $\mathcal{M}$, with $\mu \in \mathcal{M}$ making weather predictions:

$$P_\mu(x_{t+1} \mid x_1, \ldots, x_t)$$

- A prior probability $P(\mu)$ on the stations.
- The marginal probability

$$P(x_1, \ldots, x_t) = \sum_{\mu \in \mathcal{M}} P_\mu(x_1, \ldots, x_t) P(\mu)$$

- The posterior probability

$$P(\mu \mid x_1, \ldots, x_t) = \frac{P_\mu(x_1, \ldots, x_t) P(\mu)}{P(x_1, \ldots, x_t)} = \frac{\prod_{i=1}^{t} P_\mu(x_t \mid x_1, \ldots, x_{t-1}) P(\mu)}{P(x_1, \ldots, x_t)}$$

$$= \frac{P_\mu(x_t \mid x_1, \ldots, x_{t-1}) P(\mu \mid x_1, \ldots, x_{t-1})}{P(x_t \mid x_1, \ldots, x_{t-1})}$$

- The marginal posterior probability

$$P(x_{t+1} \mid x_1, \ldots, x_t) = \sum_{\mu \in \mathcal{M}} P_\mu(x_{t+1} \mid x_1, \ldots, x_t) P(\mu \mid x_1, \ldots, x_t)$$