CS 5135/6035 Learning Probabilistic Models

Lecture 3: Introduction to Probability[1]

Gowtham Atluri

August 28, 2018

[1] These slides are adapted from those that accompany the book *Bayesian Reasoning and Machine Learning*. The book and demos can be downloaded from www.cs.ucl.ac.uk/staff/D.Barber/brml

Probabilistic reasoning

- We have intuition about how uncertainty works in simple cases
- To deal with complicated situations (many events and many outcomes), we need a formal 'calculus'
- The foundational probability concepts, mathematical language, and rules of probability give us a formal framework

Part a: Essential Elements of a Probability Distribution

- Random variables
- Domain
- Probability & Axioms
- Probability Distrbution
- Interpretation

Random experiment, variable, domain

Random experiment

A phenomenon whose outcome is not predictable with certainty, but the set of all possible outcomes is known. (e.g., coin toss, roll of a dice)

Random experiment, variable, domain

Random experiment

A phenomenon whose outcome is not predictable with certainty, but the set of all possible outcomes is known. (e.g., coin toss, roll of a dice)

Random variable

A variable whose possible values/states are outcomes of a random phenomenon/experiment.

- For example, x is a random variable that capture outcome of a coin toss.

Random experiment, variable, domain

Random experiment

A phenomenon whose outcome is not predictable with certainty, but the set of all possible outcomes is known. (e.g., coin toss, roll of a dice)

Random variable

A variable whose possible values/states are outcomes of a random phenomenon/experiment.

- For example, x is a random variable that capture outcome of a coin toss.

Domain of a random variable

$\mathrm{dom}(x)$ denotes the set of possible states variable x can take.

- For example, in the case of a coin toss $\mathrm{dom}(c) = \{\text{heads}, \text{tails}\}$.
- For a roll of a dice, $\mathrm{dom}(d) = \{1, 2, 3, 4, 5, 6\}$.

Probability

$p(x = \text{s})$: the probability of variable x being in state s.

$$p(x = \text{s}) = \begin{cases} 1 & \text{we are certain } x \text{ is in state s} \\ 0 & \text{we are certain } x \text{ is not in state s} \end{cases}$$

Values between 0 and 1 represent the degree of certainty of state occupancy.

For example, in the case of a coin toss,

- $p(c = \text{heads}) = 0.5$

Kolmogorov axioms

First axiom

Probability of variable x in state s lies between 0 and 1.

$$0 \leq p(x = \text{s}) \leq 1$$

Kolmogorov axioms

First axiom

Probability of variable x in state s lies between 0 and 1.

$$0 \leq p(x = \text{s}) \leq 1$$

Second axiom

The summation of the probability over all the states is 1:

$$\sum_{x \in \text{dom}(x)} p(x = \text{x}) = 1 \qquad (or) \qquad \sum_{x} p(x) = 1$$

Kolmogorov axioms

First axiom

Probability of variable x in state s lies between 0 and 1.

$$0 \leq p(x = \text{s}) \leq 1$$

Second axiom

The summation of the probability over all the states is 1:

$$\sum_{x \in \text{dom}(x)} p(x = \text{x}) = 1 \qquad (or) \qquad \sum_{x} p(x) = 1$$

Third axiom

For mutually exclusive events s1 and s2,

$$p(x = \text{s1} \quad \cup \quad x = \text{s2}) = P(x = \text{s1}) \quad + \quad p(x = \text{s2})$$

Distribution

Given

- x is a random variable
- its domain is $\text{dom}(x) = \{\text{s1}, \text{s2}, \ldots, \text{sn}\}$

A full specification of the probability values for each of the variable states, $p(x)$, is a probability distribution.

For example, in the case of a coin toss,

- $p(c = \text{heads}) = 0.5$
- $p(c = \text{tails}) = 0.5$

Probability tables: Example

The a priori probability that a randomly selected Great British person would live in England, Scotland or Wales, is 0.88, 0.08 and 0.04 respectively.

- We can write this as a vector (or probability table) :

$$\begin{matrix} p(Cntry = \text{E}) \\ p(Cntry = \text{S}) \\ p(Cntry = \text{W}) \end{matrix} = \begin{matrix} 0.88 \\ 0.08 \\ 0.04 \end{matrix}$$

  whose component values sum to 1.
- The ordering of the components in this vector is arbitrary, as long as it is consistently applied.

Interpreting probability

Frequentist version - *empirical*

Probability is defined w.r.t a potentially infinite repetition of experiments.

E.g. If I were to repeat the experiment of flipping a coin, the limit of number of heads that occurred over the number of tosses is the probability of head occurring.

Bayesian version - *subjective*

It is a *degree of belief* that a certain event may occur.

E.g. Based on expressing a few films a user likes and dislikes, an online company tries to estimate the probability that the user will like each of the 10,000 films in their database.

Part b: Multiple variables

- Joint Probability
- Marginalization
- Conditional Probability

Operations - AND/OR

AND / Joint Probability

- p(x=a and y=b)
- Use the shorthand $p(x, y) \equiv p(x \cap y)$ for $p(x \text{ and } y)$.
- Note that $p(y, x) = p(x, y)$.

Operations - AND/OR

AND / Joint Probability

- p(x=a and y=b)
- Use the shorthand $p(x, y) \equiv p(x \cap y)$ for $p(x \text{ and } y)$.
- Note that $p(y, x) = p(x, y)$.

OR

- For specific states, we write

$$p(x = \text{a or } y = \text{b}) = p(x = \text{a}) + p(y = \text{b}) - p(x = \text{a and } y = \text{b})$$

- More generally, we can write

$$p(x \text{ or } y) \equiv p(x \cup y) = p(x) + p(y) - p(x \text{ and } y)$$

- Note that $p(x \text{ or } y) = p(y \text{ or } x)$.

Operations - Marginalization

- Marginal probability distribution is a probabilisty distribution of a subset of variables
- Given a joint distr. $p(x, y)$ the marginal distr. of x is computed as

$$p(x) = \sum_y p(x, y)$$

- E.g.: The distribution $p(MT, Cntry)$ (rows: Mother Tongue, columns: Country)

	E	S	W
Eng	0.84	0.06	0.02
$Scot$	0.03	0.02	0
Wel	0.01	0	0.02

- By summing column-wise and row-wise (separately), we have marginals

$$p(Cntry) = \begin{matrix} 0.88 \\ 0.08 \\ 0.04 \end{matrix} \qquad p(MT) = \begin{matrix} 0.92 \\ 0.05 \\ 0.03 \end{matrix}$$

Operations - Marginalization - II

- More generally,

$$p(x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n) = \sum_{x_i} p(x_1, \ldots, x_n)$$

- x_i is **marginalized out**
- This process is **marginalizing**
- $x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_n$ are **marginal variables**

Conditional Probability and Bayes' Rule

The probability of event x conditioned on knowing event y (or more shortly, the probability of x given y) is defined as

$$p(x|y) \equiv \frac{p(x,y)}{p(y)} = \frac{p(y|x)p(x)}{p(y)} \quad \text{(Bayes' rule)} \qquad \text{How is it useful?}$$

Throwing darts

$$p(\text{region 5}|\text{not region 20}) = \frac{p(\text{region 5, not region 20})}{p(\text{not region 20})} = \frac{p(\text{region 5})}{p(\text{not region 20})} = \frac{1/20}{19/20} = \frac{1}{19}$$

Applying conditional probability principle: Example

Let us assume that only three Mother Tongue languages exist : English (Eng), Scottish ($Scot$) and Welsh (Wel), with conditional probabilities given the country of residence, England (E), Scotland (S) and Wales (W).

Using the state ordering:

$$MT = [Eng, Scot, Wel]; \qquad Cntry = [E, S, W]$$

We write a (fictitious) conditional probability table and a marginal table

$$p(MT|Cntry) = \begin{pmatrix} & E & S & W \\ Eng & 0.95 & 0.7 & 0.6 \\ Scot & 0.04 & 0.3 & 0.0 \\ Wel & 0.01 & 0.0 & 0.4 \end{pmatrix} \qquad p(Cntry) = \begin{pmatrix} 0.88 \\ 0.08 \\ 0.04 \end{pmatrix}$$

What is the joint probability $p(Cntry, MT)$?

Applying conditional probability principle: Example

Given

$$p(MT|Cntry) = \begin{pmatrix} & E & S & W \\ Eng & 0.95 & 0.7 & 0.6 \\ Scot & 0.04 & 0.3 & 0.0 \\ Wel & 0.01 & 0.0 & 0.4 \end{pmatrix} \qquad p(Cntry) = \begin{pmatrix} 0.88 \\ 0.08 \\ 0.04 \end{pmatrix}$$

Applying conditional probability, $p(MT, Cntry) = p(MT|Cntry)p(Cntry)$

$$\begin{pmatrix} 0.95 \times 0.88 & 0.7 \times 0.08 & 0.6 \times 0.04 \\ 0.04 \times 0.88 & 0.3 \times 0.08 & 0.0 \times 0.04 \\ 0.01 \times 0.88 & 0.0 \times 0.08 & 0.4 \times 0.04 \end{pmatrix} = \begin{pmatrix} & E & S & W \\ Eng & 0.84 & 0.06 & 0.02 \\ Scot & 0.03 & 0.02 & 0 \\ Wel & 0.01 & 0 & 0.02 \end{pmatrix}$$

Probability tables

Large numbers of variables

- For joint distributions over a larger number of variables, $x_i, i = 1, \ldots, D$, with each variable x_i taking K_i states
  - the table for joint distribution is an array with $\prod_{i=1}^{D} K_i$ entries.
- Explicitly storing tables therefore requires space exponential in the number of variables, which rapidly becomes impractical for a large number of variables.

Probability tables

Large numbers of variables

- For joint distributions over a larger number of variables, $x_i, i = 1, \ldots, D$, with each variable x_i taking K_i states
  - the table for joint distribution is an array with $\prod_{i=1}^{D} K_i$ entries.
- Explicitly storing tables therefore requires space exponential in the number of variables, which rapidly becomes impractical for a large number of variables.

Exchangeability

A probability distribution assigns a value to each of the joint states of the variables. For this reason, $p(T, J, R, S) \equiv p(J, S, R, T)$

- In each case the joint setting of the variables is simply a different index to the same probability table.
- Not to be confused with functions $f(x, y)$

Part c: Independence

- Independence
- Conditional Independence

Independence

- Variables x and y are independent if knowing one event gives no extra information about the other event.
- Mathematically, this is expressed by

$$p(x, y) = p(x)p(y)$$ Where is it useful?

- Independence of x and y is equivalent to

$$p(x|y) = p(x) \Leftrightarrow p(y|x) = p(y)$$

If $p(x|y) = p(x)$ for all states of x and y, then the variables x and y are said to be independent. We write then $x \perp\!\!\!\perp y$.

Independence

Interpretation

- Note that $x \perp\!\!\!\perp y$ doesn't mean that, given y, we have no information about x.
- It means the only information we have about x is contained in $p(x)$.

Factorisation

If

$$p(x, y) = kf(x)g(y)$$

for some constant k, and positive functions $f(\cdot)$ and $g(\cdot)$ then x and y are independent.

Conditional Independence

$$\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \mathcal{Z}$$

denotes that the two sets of variables $\mathcal{X}$ and $\mathcal{Y}$ are independent of each other given the state of the set of variables $\mathcal{Z}$.

- This means that

$$p(\mathcal{X}, \mathcal{Y}|\mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})p(\mathcal{Y}|\mathcal{Z}) \text{ and } p(\mathcal{X}|\mathcal{Y}, \mathcal{Z}) = p(\mathcal{X}|\mathcal{Z})$$

for all states of $\mathcal{X}, \mathcal{Y}, \mathcal{Z}$.

- In case the conditioning set is empty we write $\mathcal{X} \perp\!\!\!\perp \mathcal{Y}$ for $\mathcal{X} \perp\!\!\!\perp \mathcal{Y} | \emptyset$
  - in which case $\mathcal{X}$ is (unconditionally) independent of $\mathcal{Y}$.

Conditional Independence

Conditional independence does not imply marginal independence

Given

- three variables x, y, z
- their joint probability $p(x, y, z)$
- and $x \perp\!\!\!\perp y | z$

$$p(x, y) = \sum_z \underbrace{p(x|z)p(y|z)}_{\text{cond. indep.}} p(z) \neq \underbrace{\sum_z p(x|z)p(z)}_{p(x)} \underbrace{\sum_z p(y|z)p(z)}_{p(y)}$$

Conditional dependence

If $\mathcal{X}$ and $\mathcal{Y}$ are not conditionally independent, they are conditionally dependent.
This is written as

$$\mathcal{X} \top\!\!\!\top \mathcal{Y} | \mathcal{Z}$$

Conditional Independence example

Based on a survey of households in which husband and wife each own a car:

wife's car type $\perp\!\!\!\perp$ husband's car type | family income

There are 4 car types, the first two being *cheap* and the last two being *expensive*. Using w for the wife's car type and h for the husband's:

$$p(w|inc = \text{low}) = \begin{pmatrix} 0.7 \\ 0.3 \\ 0 \\ 0 \end{pmatrix}, \quad p(w|inc = \text{high}) = \begin{pmatrix} 0.2 \\ 0.1 \\ 0.4 \\ 0.3 \end{pmatrix}$$

$$p(h|inc = \text{low}) = \begin{pmatrix} 0.2 \\ 0.8 \\ 0 \\ 0 \end{pmatrix}, \quad p(h|inc = \text{high}) = \begin{pmatrix} 0 \\ 0 \\ 0.3 \\ 0.7 \end{pmatrix}$$

$$p(inc = \text{low}) = 0.9$$ Compute $p(w, h)$

Conditional Independence example

Then the marginal distribution $p(w, h)$ is

$$p(w, h) = \sum_{inc} p(w, h|inc)p(inc) = \sum_{inc} p(w|inc)p(h|inc)p(inc)$$

resulting in

$$p(w, h) = \begin{pmatrix} 0.126 & 0.504 & 0.006 & 0.014 \\ 0.054 & 0.216 & 0.003 & 0.007 \\ 0 & 0 & 0.012 & 0.028 \\ 0 & 0 & 0.009 & 0.021 \end{pmatrix}$$

Conditional Independence example

Then the marginal distribution $p(w, h)$ is

$$p(w,h) = \sum_{inc} p(w,h|inc)p(inc) = \sum_{inc} p(w|inc)p(h|inc)p(inc)$$

resulting in

$$p(w,h) = \begin{pmatrix} 0.126 & 0.504 & 0.006 & 0.014 \\ 0.054 & 0.216 & 0.003 & 0.007 \\ 0 & 0 & 0.012 & 0.028 \\ 0 & 0 & 0.009 & 0.021 \end{pmatrix}$$

From this we can find the marginals and calculate

$$p(w)p(h) = \begin{pmatrix} 0.117 & 0.468 & 0.0195 & 0.0455 \\ 0.0504 & 0.2016 & 0.0084 & 0.0196 \\ 0.0072 & 0.0288 & 0.0012 & 0.0028 \\ 0.0054 & 0.0216 & 0.0009 & 0.0021 \end{pmatrix}$$

This shows that while $w \perp\!\!\!\perp h | \, inc$, it is not true that $w \perp\!\!\!\perp h$. E.g., even if we don't know the family income, if we know that the husband has a cheap car then his wife must also have a cheap car.

Part d: Reasoning

- Probabilistic Reasoning
- Scientific Inference

Inspector Clouseau

Inspector Clouseau arrives at the scene of a crime. The Butler (B) and Maid (M) are his main suspects. The inspector has a prior belief of 0.6 that the Butler is the murderer, and a prior belief of 0.2 that the Maid is the murderer. These probabilities are independent in the sense that $p(B, M) = p(B)p(M)$. (It is possible that both the Butler and the Maid murdered the victim or neither). The inspector's *prior* criminal knowledge can be formulated mathematically as follows:

$$\text{dom}(B) = \text{dom}(M) = \{\text{murderer, not murderer}\}$$
$$\text{dom}(K) = \{\text{knife used, knife not used}\}$$
$$p(B = \text{murderer}) = 0.6, \qquad p(M = \text{murderer}) = 0.2$$
$$\begin{aligned} &p(\text{knife used}|B = \text{not murderer}, \quad M = \text{not murderer}) &&= 0.3 \\ &p(\text{knife used}|B = \text{not murderer}, \quad M = \text{murderer}) &&= 0.2 \\ &p(\text{knife used}|B = \text{murderer}, \quad M = \text{not murderer}) &&= 0.6 \\ &p(\text{knife used}|B = \text{murderer}, \quad M = \text{murderer}) &&= 0.1 \end{aligned}$$

The victim lies dead in the room and the inspector quickly finds the murder weapon, a Knife (K). What is the probability that the Butler is the murderer? (Remember that it might be that neither is the murderer).

Inspector Clouseau

Using b for the two states of B and m for the two states of M,

$$p(B|K) = \sum_m p(B, m|K) = \sum_m \frac{p(B,m,K)}{p(K)} = \frac{p(B)\sum_m p(K|B,m)p(m)}{\sum_b p(b)\sum_m p(K|b,m)p(m)}$$

Plugging in the values we have

$$\begin{aligned} p(B = \text{murderer}|\text{knife used}) &= \frac{\frac{6}{10}\left(\frac{2}{10}\times\frac{1}{10} + \frac{8}{10}\times\frac{6}{10}\right)}{\frac{6}{10}\left(\frac{2}{10}\times\frac{1}{10} + \frac{8}{10}\times\frac{6}{10}\right) + \frac{4}{10}\left(\frac{2}{10}\times\frac{2}{10} + \frac{8}{10}\times\frac{3}{10}\right)} \\ &= \frac{300}{412} \approx 0.73 \end{aligned}$$

Hence knowing that the knife was the murder weapon strengthens our belief that the butler did it.

Inspector Clouseau

The role of $p(\text{knife used})$ in the Inspector Clouseau example can cause some confusion. In the above,

$$p(\text{knife used}) = \sum_b p(b) \sum_m p(\text{knife used}|b, m)p(m)$$

is computed to be 0.456. But surely, $p(\text{knife used}) = 1$, since this is given in the question! Note that the quantity $p(\text{knife used})$ relates to the *prior* probability the model assigns to the knife being used (in the absence of any other information). If we know that the knife is used, then the *posterior* is

$$p(\text{knife used}|\text{knife used}) = \frac{p(\text{knife used, knife used})}{p(\text{knife used})} = \frac{p(\text{knife used})}{p(\text{knife used})} = 1$$

which, naturally, must be the case.

Scientific Inference

Much of science deals with problems of the form: *tell me something about the variable* θ given that I have observed data $\mathcal{D}$ and have some knowledge of the underlying data generating mechanism.

Our interest is then the quantity

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_\theta p(\mathcal{D}|\theta)p(\theta)}$$

- $p(\mathcal{D}|\theta)$ is a *generative model* of the dataset
- $p(\theta)$ is a *prior* belief about which variable values are appropriate
- $p(\theta|\mathcal{D})$ is the inferred *posterior* distribution of the variable in light of the observed data.

Generative models in science

- This use of a generative model sits well with physical models of the world which typically postulate how to generate observed phenomena, assuming we know the model.
- For example, one might postulate how to generate a time-series of displacements for a swinging pendulum but with unknown mass, length and damping constant.
- Using this generative model, and given only the displacements, we could infer the unknown physical properties of the pendulum, such as its mass, length and friction damping constant.

Prior, Likelihood and Posterior

For data $\mathcal{D}$ and variable θ, Bayes rule tells us how to update our prior beliefs about the variable θ in light of the data to a posterior belief:

$$\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{D}|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathcal{D})}_{\text{evidence}}}$$

The evidence is also called the marginal likelihood.

The term likelihood is used for the probability that a model generates observed data.

Prior, Likelihood and Posterior

More fully, if we condition on the model M, we have

$$p(\theta|\mathcal{D}, M) = \frac{p(\mathcal{D}|\theta, M)p(\theta|M)}{p(\mathcal{D}|M)}$$

where we see the role of the likelihood $p(\mathcal{D}|\theta, M)$ and marginal likelihood $p(\mathcal{D}|M)$.

The marginal likelihood is also called the model likelihood.