Advanced Machine Learning Subsidary Notes

Lecture 20: Probability

Adam Prügel-Bennett

March 12, 2024

1 Keywords

- Probability, Random variable, Expectations

2 Main Points

2.1 Basic Language

- Probability is the process of modelling a world with uncertain outcomes
- The model is not perfect, but reflects what we care about
- Mathematicians model the possible outcomes of the world they are considering in terms of a set of *elementary events*
- This set is frequently called Ω (but you are free to give it any name)
- The elementary events are *mutual exclusive* so that if $\omega_i, \omega_j \in \Omega$ then $\omega_i \cap \omega_j = \emptyset$ (this just means there is no intersection between elementary events)
- The set is exhaustive so that $\bigcup_i \omega_i = \Omega$
- This means in our model of the world every possible outcome is included in Ω
- Now we are often not interested in elementary events (they may be too small)
- Instead we consider **events** $\mathcal{E} = \bigcup_{i\in\mathcal{I}} \omega_i$ where $\mathcal{I}$ is some *index set*
- Suppose we are modelling the probability of a darts match
- Ω might represent every location where a dart might land
- $\omega_i \in \Omega$ is a particular position
- I might be interested in the probability of a bull's eye $\mathcal{B} = \bigcup_{i\in\mathcal{I}} \omega_i$
- In this particular case this is a union of an uncountable infinity of points
- In this case I might represent the elementary events by a 2-d vector $\boldsymbol{x}$
- The event of getting a bull's eye corresponds to $\boldsymbol{x}$ being within the region that defines the bull's eye

2.2 Probability and Random Variables

- We can attribute a **probability**, $\mathbb{P}[\mathcal{E}]$, to an event, $\mathcal{E}$
- Probabilities are number between 0 and 1: $0 \leq \mathbb{P}[\mathcal{E}] \leq 1$
- 0 means the event never occurs, 1 means the event definitely will occur
- $\mathbb{P}[\mathcal{E}] + \mathbb{P}[\neg\mathcal{E}] = 1$ where $\neg\mathcal{E} = \Omega \setminus \mathcal{E}$
- That is, the probability of an event occurring and the event not occurring is 1
- In some cases we can interpret $\mathbb{P}[\mathcal{E}]$ as the expected frequency of occurrence of a repetitive trial
- E.g. If the event, $\mathcal{E}$ is rolling an (honest) dice and getting an even number then $\mathbb{P}[\mathcal{E}] = \frac{1}{2}$ can be seen as saying that on average this will happen half the time if you repeat the experiment often enough
- However how to we interpret $\mathbb{P}[\text{Pass COMP6208 exam}]$?
- Hopefully, this is something you only attempt once
- We can think of probability as encoding our informed belief that something will happen
- When our knowledge changes our probability will change
- Random variables, X, are numbers we assign to each elementary event (many events can have the same number assigned)
- That is we partition the set of outcomes Ω and assign a numbers to each partition
- E.g. for a dice

$$X = \begin{cases} 0 & \text{if } \omega \in \{1, 3, 5\} \\ 1 & \text{if } \omega \in \{2, 4, 6\} \end{cases}$$

- If we let $\mathcal{E}_i$ be the event that corresponding to the partition with value x_i then $\mathbb{P}[X = x_i] = \mathbb{P}[\mathcal{E}_i]$
- It is common practice (though not universal) to denote random variables with capital letters
- The correspond to events with uncertain outcomes (things that happen in the future)
- We distinguish the scalar that we assign to random variables by writing them in lower case, e.g. x_i
- When we write $\mathbb{P}[X]$ we can view this as short-hand for

$$(\mathbb{P}[X = x] \mid x \in \mathcal{X}) = (\mathbb{P}[X = x_1], \mathbb{P}[X = x_2], \dots \mathbb{P}[X = x_n])$$

  where $\mathcal{X}$ is the set of possible values that X can take
- That is, $\mathbb{P}[X]$, is the probability mass function which tells us the probability of the random variable X for all possible values it can take
- We differentiate between random variables (capitals) and scalars (lower case) because they behave very differently when we take expectations
- A function, $Y = g(X)$, of a random variable X is a random variable
- **Continuous random variables**

- When we consider random variables with continuous outcomes (e.g. the time I go to bed, T) then there is an uncountable infinity of possible outcomes and $\mathbb{P}[T = t] = 0$
- Here we can consider the probability of a random variable, $\boldsymbol{X}$ occurring in a ball, $\mathcal{B}(\boldsymbol{x}, \epsilon)$, of radius ϵ around the point $\boldsymbol{x}$
- Taking the ratio of that probability to the volume of the ball in the limit $\epsilon \to 0$ defines a **probability density**

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = \lim_{\epsilon \to 0} \frac{\mathbb{P}[\boldsymbol{X} \in \mathcal{B}(\boldsymbol{x}, \epsilon)]}{|\mathcal{B}(\boldsymbol{x}, \epsilon)|}$$

- Probability densities are not probabilities (they are positive, but not necessarily less than 1)
- Because they are densities, if I change the way I measure volumes then the numeric values of the density will change (probability density measured in units of probability per cm is different to that measured in units of probability per inch)
- A consequent is that if we do a change of variables then often the probability density will change
- Consider a region R—we can describe this using different coordinate systems x or $y = g(x)$

$$\mathbb{P}[X \in \mathcal{R}] = \int_{\mathcal{R}} f_X(x)\,\mathrm{d}x = \mathbb{P}[Y \in \mathcal{R}] = \int_{\mathcal{R}} f_Y(y)\,\mathrm{d}y$$

(that is the probability of the event occurring in some region should not depend on coordinate region we use)
- Thus $f_X(x)\,|\mathrm{d}x| = f_Y(y)\,|\mathrm{d}y|$ or

$$f_X(x) = f_Y(y) \left|\frac{\mathrm{d}y}{\mathrm{d}x}\right| = f_Y(g(x))\,|g'(x)|$$

- In high dimensions if we make a change of variables $\boldsymbol{x} \to \boldsymbol{y}(\boldsymbol{x})$ (which can be seen as a change of random variables $\boldsymbol{X} \to \boldsymbol{Y}(\boldsymbol{X})$) then

$$f_{\boldsymbol{X}}(\boldsymbol{x}) = f_{\boldsymbol{Y}}(\boldsymbol{y})\,|\det(\mathbf{J})|$$

where $\mathbf{J}$ is the Jacobian matrix

$$\mathbf{J} = \begin{pmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} & \cdots & \frac{\partial y_1}{\partial x_n} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} & \cdots & \frac{\partial y_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial y_n}{\partial x_1} & \frac{\partial y_n}{\partial x_2} & \cdots & \frac{\partial y_n}{\partial x_n} \end{pmatrix}$$

- An alternative definition of probability densities in 1-d is

$$f_X(x) = \lim_{\delta x \to 0} \frac{\mathbb{P}[x \le X < x + \delta x]}{\delta x}$$

- Thus

$$f_X(x)\,\delta x \approx \mathbb{P}[x \le X < x + \delta x]$$

so that $f_X(x)\,\delta x \le 1$ (since this is a probability)

- **Cumulative Distribution Functions (CDF)**

$$F_X(x) = \mathbb{P}[X \le x] = \begin{cases} \sum_{i: x_i \le x} \mathbb{P}[X = x_i] \\ \int_{-\infty}^{x} f_X(y)\,\mathrm{d}y \end{cases}$$

- This is a function that goes from 0 to 1 as x goes from $-\infty$ to ∞
- For continuous random variables

$$f_X(x) = \frac{\mathrm{d}F_X(x)}{\mathrm{d}x}$$

- Sometimes it is easy working with the CDF than the PDF

2.3 Expectations

- Expectations compute the average value of a quantity (it is a number)
- Mathematicians write expectations as $\mathbb{E}_{\boldsymbol{X}}[\cdots]$ (note that physicist often use $\langle \cdots \rangle$)
- We can define the expectation of $\boldsymbol{Y} = g(\boldsymbol{X})$ as

$$\mathbb{E}_{\boldsymbol{X}}[g(\boldsymbol{X})] = \begin{cases} \displaystyle\sum_{\boldsymbol{x}\in\mathcal{X}} g(\boldsymbol{x})\,\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}] \\ \displaystyle\int g(\boldsymbol{x})\,f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} \end{cases}$$

- The expectation of a constant c is c

$$\mathbb{E}_{\boldsymbol{X}}[c] = \begin{cases} \displaystyle\sum_{\boldsymbol{x}\in\mathcal{X}} c\,\mathbb{P}[\boldsymbol{X}=\boldsymbol{x}] = c\sum_{\boldsymbol{x}\in\mathcal{X}} \mathbb{P}[\boldsymbol{X}=\boldsymbol{x}] = c \\ \displaystyle\int c\,f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = c\int f_{\boldsymbol{X}}(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x} = c \end{cases}$$

- A consequence of this is that $\mathbb{E}_X[\mathbb{E}_X[g(X)]] = \mathbb{E}_X[g(X)]$
- Summations and integrals are linear operators

$$\sum_i (a\,x_i + b\,y_i) = a\left(\sum_i x_i\right) + b\left(\sum_i y_i\right)$$
$$\int (a\,f(\boldsymbol{x}) + b\,g(\boldsymbol{x}))\,\mathrm{d}\boldsymbol{x} = a\left(\int f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right) + b\left(\int g(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}\right)$$

  from which it follows that expectations are linear

$$\mathbb{E}[a\,X + b\,Y] = a\,\mathbb{E}[X] + b\,\mathbb{E}[Y]$$

- Beware usually $\mathbb{E}[X\,Y] \neq \mathbb{E}[X]\;\mathbb{E}[Y]$ (unless X and Y are independent)
- **Indicator Functions**
  - Indicator functions are functions that take on the value of 0 or 1
  - They are written in different ways. One nice form is the Iverson notation

$$\textit{predicate} = \begin{cases} 1 & \text{if } \textit{predicate} \text{ is True} \\ 0 & \text{if } \textit{predicate} \text{ is False} \end{cases}$$

  - Sometimes it is written $\boldsymbol{I}_A(x)$ where $A(x)$ is the predicate
  - Donald Kunth is a champion of the Iverson notation (it is easy to manipulate)
  - For probabilities

$$\mathbb{P}[\textit{predicate}] = \mathbb{E}[\textit{predicate}]$$

    it is not too difficult to convince yourself that this is correct as you are summing the probabilities where the predicate is true
  - As an example the CDF is given by

$$F_X(x) = \mathbb{P}[X \leq x] = \mathbb{E}[X \leq x]$$

2.4 Calculus of Probabilities

- When we model the world it is common to consider different random variables
- In this case we can consider the **joint probability**

$$p_{X,Y}(x, y) = \mathbb{P}[X = x, Y = y]$$

  i.e. the probability of the event where both $X = x$ and $Y = y$

- $\mathbb{P}[X = x, Y = y] = \mathbb{P}[Y = y, X = x]$ the order in which you write the random variables in a joint distribution function doesn't matter
- There are really only a couple of ways you can manipulate probabilities (this defines the calculus of probability)
- The first is known as *marginalisation*. Given $\mathbb{P}[X, Y]$

$$\mathbb{P}[X = x] = \sum_{y \in \mathcal{Y}} \mathbb{P}[X = x, Y = y]$$

  where $\mathcal{Y}$ is the set of values that the random variable Y takes

- Often we will just write

$$\mathbb{P}[X] = \sum_{Y} \mathbb{P}[X, Y]$$

- **Conditional Probability**
  - The other rule of probability involves introducing **conditional probabilities**
  - We can also define the probability of an event X given that $Y = y$ has occurred

$$\mathbb{P}[X \mid Y = y] = \frac{\mathbb{P}[X, Y = y]}{\mathbb{P}[Y = y]}$$

  - This is a probability over X. If we marginalise over X then

$$\sum_{x \in \mathcal{X}} \mathbb{P}[X = x \mid Y = y] = \frac{\sum_{x \in \mathcal{X}} \mathbb{P}[X = x, Y = y]}{\mathbb{P}[Y = y]} = \frac{\mathbb{P}[Y = y]}{\mathbb{P}[Y = y]} = 1$$

  - In constructing a model it is often much easier to specify conditional probabilities (because you know something) rather than joint probabilities
  - When manipulating probabilities it is often easier to work with joint probabilities because we can simplify them by marginalising out random variables we are not interested in
  - To obtain the joint probability from a conditional probability we use

$$\mathbb{P}[X, Y] = \mathbb{P}[X|Y]\ \mathbb{P}[Y] = \mathbb{P}[Y|X]\ \mathbb{P}[X]$$

  - This generalises to more random variables

$$\mathbb{P}[X, Y, Z] = \mathbb{P}[X, Y|Z]\ \mathbb{P}[Z] = \mathbb{P}[X|Y, Z]\ \mathbb{P}[Y|Z]\ \mathbb{P}[Z]$$

  - This way of breaking down probabilities in not unique. For example, we could have written

$$\mathbb{P}[X, Y, Z] = \mathbb{P}[Y, Z|X]\ \mathbb{P}[X] = \mathbb{P}[Z|Y, X]\ \mathbb{P}[Y|X]\ \mathbb{P}[X]$$

- Note that $\mathbb{P}[A, B \mid X, Y]$ means the probability of random variables A and B given that X and Y take particular values
- **Beware**: conditional probabilities, $\mathbb{P}[X \mid Y]$ are probabilities for X, but not Y

$$\sum_{x \in \mathcal{X}} \mathbb{P}[X = x \mid Y] = 1$$
$$\sum_{y \in \mathcal{Y}} \mathbb{P}[X \mid Y = y] \neq 1$$

(in general)

- Note that

$$\begin{aligned}\mathbb{E}_Y[\mathbb{P}[X \mid Y]] &= \sum_{y \in \mathcal{Y}} \mathbb{P}[Y = y]\ \mathbb{P}[X|Y = y] \\ &= \sum_{y \in \mathcal{Y}} \mathbb{P}[X, Y = y] = \mathbb{P}[X]\end{aligned}$$

that is taking the expectation of $\mathbb{P}[X \mid Y]$ with respect to Y is the same as marginalising the joint probability $\mathbb{P}[X, Y]$ with respect to Y

- Joint probabilities does not imply causality
- We might believe there is a causal relationship between getting very cold and catching a cold $\mathbb{P}[\text{Catch cold} \mid \text{freezing}] > \mathbb{P}[\text{Catch cold}]$
- However we can use Bayes' rule (see next lecture) to compute $\mathbb{P}[\text{freezing} \mid \text{Catch cold}]$
- That is, given all I know is that Bob has a cold I can deduce the probability that he was freezing shortly before catching the cold. This is a perfectly sensible thing to consider, but clearly it does not imply that freezing yesterday was caused by him catching a cold today

- **Independence**
  - Random variables X and Y are said to be *independent* if

$$\mathbb{P}[X, Y] = \mathbb{P}[X]\ \mathbb{P}[Y]$$

  - Because $\mathbb{P}[X, Y] = \mathbb{P}[X|Y]\ \mathbb{P}[Y]$ and $\mathbb{P}[X, Y] = \mathbb{P}[Y|X]\ \mathbb{P}[X]$ independence implies

$$\mathbb{P}[X|Y] = \mathbb{P}[X] \qquad\qquad \mathbb{P}[Y|X] = \mathbb{P}[Y]$$

  - Probabilistic independence only implies a mathematical co-incident not necessarily causal independence. There are examples of events that are causally dependent but nevertheless statistical independent (although such cases are rare)
  - However causal independence always implies probabilistic independence
  - If $X \in \{0, 1\}$ represents the outcome of tossing a coin and $Y \in \{1, 2, 3, 4, 5, 6\}$ the outcome of rolling a dice then X and Y are independent
  - In well conducted experiments we expect the results we obtain are independent (that is the outcome of one experiment should not affect the outcome of another experiment)
  - Let $\mathcal{D} = (X_1, X_2, \ldots, X_m)$ represents possible outcomes from a set of m well conducted experiments then

$$\mathbb{P}[\mathcal{D}] = \prod_{i=1}^{m} \mathbb{P}[X_i]$$

- Denoting a possible sentence I might say by $\mathcal{S} = (W_1, W_2, \ldots, W_m)$ then

$$\mathbb{P}[\mathcal{S}] \neq \prod_{i=1}^{m} \mathbb{P}[W_i]$$

- If this is not the case they you expect my sentences to be equally likely irrespective of the order of my words (I certainly home this is not the case)
- Independence is a very useful condition that substantially simplifies probabilistic calculations
- Independence is a strong condition that does not happen in many models. However, a far more likely condition is **conditional independence**. For example, X and Y are conditionally independent given Z if

$$\mathbb{P}[X, Y|Z] = \mathbb{P}[X|Z]\ \mathbb{P}[Y|Z]$$

- Me working today, W, and me getting stuck in a traffic jam, J, are not statistically independent events, but they are statistically independent (at least in a simple of model of the world) given the day of the week, D

$$\mathbb{P}[W, J|D] = \mathbb{P}[W|D]\ \mathbb{P}[J|D]$$

3 Exercises

1. Suppose I make an investment of capital C_0 and each day I get a random return R_t such that $C_t = (1 + r_t)\, C_{t-1}$. Taking logs

$$\log(C_t) = \log(1 + R_t) + \log C_{t-1} \approx R_r + \log(C_{t-1})$$

where we assume $|R_t| \ll 1$. Rearranging $\log(C_t) - \log(C_{t-1}) \approx R_t$ and summing

$$\sum_{t=1}^{T} (\log(C_t) - \log(C_{t-1})) = \log(C_T) - \log(C_0) = \sum_{t=1}^{T} R_t$$

Now $\log(C_T) - \log(C_0) = \log(C_T/C_0) = \log(G_T)$ where $G_T = C_T/C_0$ is the multiplicative gain (or loss) in the investment. Let $L_T = \log(G_T)$. Assuming we live in a random world so that $R_t \sim \mathcal{N}(0, \epsilon)$. That is, the return each day is normally distributed with mean 0 and variance ϵ. Furthermore, let us assume that R_t is independent of $R_{t'}$ then L_T is the sum of T normal independent variables so that $L_T \sim \mathcal{N}(0, \sqrt{T}\,\epsilon)$. That is, $f_L(\ell) = \mathcal{N}(\ell \mid 0, \sqrt{T}\,\epsilon)$. Compute the PDF of G_T using

$$f_G(g) = f_L(\ell)\, \frac{\mathrm{d}\ell}{\mathrm{d}g}$$

where $\ell = \log(g)$

4 Answers

1. This may seem a crazy model, but it is very commonly used in finance and models stock prices extremely well. I'm not going to give the answer, but the distribution is a log-normal distribution which you can look up on Wikipedia. It occurs frequently when a random variable is equal to a product of random variables (taking the logarithm and using the central limit theorem—i.e. the sum of many random variables is approximately normally distributed—gives rise to a log-normal distribution).