Tutorial: Probability Models[1]

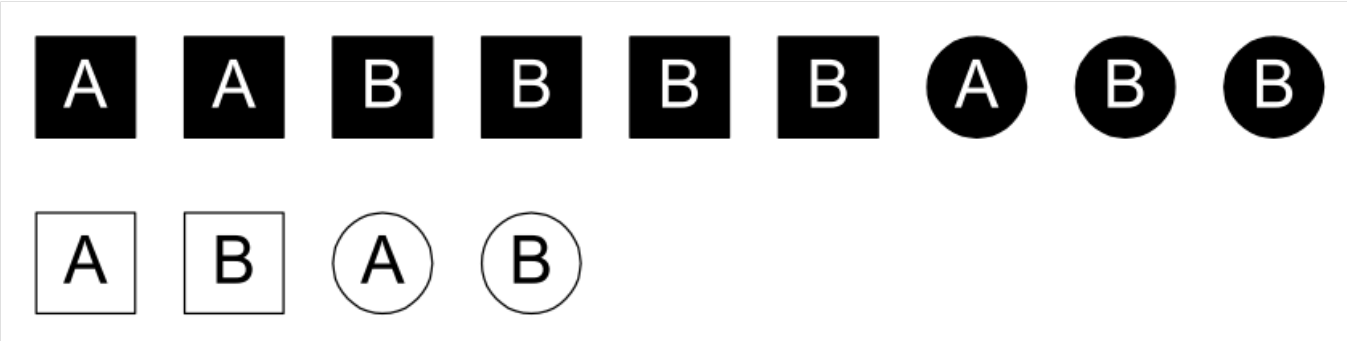
Machine Learning

September 26, 2023

[1] Questions from various sources including Neapolitan and Jiang, "Contemporary AI", CRC Press (2012)

Basic Probability I

1. You are given a set of 13 squares and circles, 9 of which are coloured black and the rest are coloured white.
   Each object also has either the letter "A" or "B" on it. There are: 2 black squares with an A, 4 black squares with a B and 1 black circle with an A. Of the remaining, there is 1 white square and 1 white circle each with an A. Here is a diagrammatic representation:



   Let *Black* denote the set of black objects, *White* denote the set of white objects, *Square* denote the set of square objects, *A* the set of objects with an "A" and so on. Assuming all

Basic Probability II

objects are equally likely (the so-called Principle of Indifference):

(a) What is $P(A)$?
(b) What is $P(A|Square)$?
(c) Are A and $Square$ independent?
(d) Are A and $Black$ independent?
(e) Are A and $Square$ conditionally independent given $Black$?
(f) Are A and $Square$ conditionally independent given $White$?
(g) The Law of Total Probability gives us: $P(A) = P(A, White) + P(A, Black)$. Verify that the law holds in this case.
(h) Using a probability-tree, calculate $P(Black|A)$.
(i) Using Bayes' Rule, calculate $P(Black|A)$

Basic Probability III

Answer. (a) $P(A) = 5/13$; (b) $P(A|Square) = 3/8$; (c) No, since $P(A) \neq P(A|Square)$; (d) $P(A|Black) = 1/3$. So, no A and $Black$ are not independent; (e) $P(A|Square, Black) = 1/3 = P(A|Black)$. So, A and $Square$ are conditionally independent, given $Black$ (f) $P(A|White) = 1/2$ and $P(A|Square, White) = 1/2$. So, A and $Square$ are conditionally independent given $White$; (g) $P(A) = P(A|White)P(White) + P(A|Black)P(Black)$. It is easy to verify that the RHS is $5/13$; (h) Left as an exercise for students; (i) $P(Black|A) = \frac{P(A|Black)P(Black)}{P(A)}$. Now $P(A) = P(A|Black)P(Black) + P(A|White)P(White)$.

That is $P(Black|A) = \frac{(1/3)(9/13)}{((1/3)(9/13)+(1/2)(4/13)} = 3/5$.

2. There are two urns (Urn1 and Urn2). Urn1 has 2 red marbles and 2 blue marbles. Urn2 has 1 red and 3 blue marbles.[2] The urn labels are now covered and a coin is flipped to select an urn. Having selected an urn, we draw a marble from the urn. The marble is red. What is the probability that the urn selected was Urn1?

   Answer. This can be solved by simply drawing the probability tree. The first branch has a binary choice with probability 0.5 of selecting Urn1 or Urn2. For Urn1 there is a probability of 0.5:0.5 of selecting red:blue marbles. The corresponding choices are 0.25:0.75 for Urn2. Conditioning on

Basic Probability V

a red marble being drawn will lead to $P(Urn1) = 0.25/(0.25 + 0.125) = 2/3$.
Now do the same with using Bayes' rule. That is, we want: $P(Urn1|Red)$. Now $P(Urn1|Red) = \alpha P(Red|Urn1)P(Urn1)$.
$P(Red|Urn1) = 0.5$ and $P(Urn1) = 0.5$.
Therefore, $P(Urn1|Red) = 0.25\alpha$. Similarly $P(\neg Urn1|Red) = \alpha P(Red|\neg Urn1)P(\neg Urn1) = 0.125\alpha$. Since $0.25\alpha + 0.125\alpha = 1$, $\alpha = 1/0.375$ and $P(Urn1|Red) = 0.25/0.375 = 2/3$

Basic Probability VI

3. In a typical English summer, the probability that the temperature falls below 10 degrees Celsius is 0.4. In that case, the English cricket team wins with probability 0.75. The probability that the temperature is between 10 and 30 degrees Celsius is 0.4, in which case the English team wins with probability 0.65. The probability that the temperature is greater than 30 degrees is 0.2 and in that case, the English team wins with probability 0.55. You have just received an SMS saying the English team has won. What is the probability that the temperature was below 10 degrees?

   Answer. Same sort of problem as the urns. Do it two ways: once with trees, and once with Bayes' rule.

[2] Some of these exercises are from M. Cargal, "Discrete Mathematics for Neophytes".

Probability Distributions I

4. The probability mass function of a discrete r.v. is as follows:

$$p(X = x) = \begin{cases} 1/3 & x = -1, 0, 1 \\ 0 & \text{otherwise} \end{cases}$$

What is $\mu_X = E(X)$?

Answer. $\mu_X = 1/3(-1 + 0 + 1) = 0$.

5. You are told $Var(X) = E[(X - \mu_X)^2]$. What is $Var(X)$ for the r.v. in the above?

Answer. $Var(X) = E[(X - \mu_X)^2] = E(X^2) - {\mu_X}^2 = E(X^2)$. This is equal to $1/3((-1)^2 + 0 + (1)^2) = 2/3$

Probability Distributions II

6. Repeat the calculations for the following mass function:

$$p(X = x) = \begin{cases} 1/3 & x = -2, 0, 2 \\ 0 & \text{otherwise} \end{cases}$$

Why does the variance increase?

Answer. The mean stays the same, but the variance changes to $8/3$, since the points are now spread out more.

7. Let X be the random variable denoting the number of dots that come up on the throw of a six-sided die. What is $E(X)$? (Are store-bought dice uniform?)

Answer. $E(X) = 1/6(1 + 2 + 3 + 4 + 5 + 6) = 3.5$

Probability Distributions III

8. Let X be a random variable denoting the number of successes in n i.i.d. Bernoulli trials, each with probability p of success. What is $E(X)$?

   Answer. The expected value for the number of successes is given by:

$$
\begin{aligned}
E(X) &= \sum_{k=0}^{n} k \binom{n}{k} p^k (1-p)^{n-k} \\
&= \sum_{k=0}^{n} k \frac{n!}{(n-k)!k!} p^k (1-p)^{n-k} \\
&= \sum_{k=1}^{n} np \frac{(n-1)!}{(n-k)!(k-1)!} p^(k-1)(1-p)^{n-k}
\end{aligned}
$$

Let $i = k - 1$. Then:

$$\begin{aligned} E(X) &= \sum_{k=1}^{n} np \frac{(n-1)!}{(n-k)!(k-1)!} p^{(}k-1)(1-p)^{n-k} \\ &= np \sum_{i=0}^{n-1} \frac{(n-1)!}{(n-1-i)!i!} p^{i}(1-p)^{n-1-i} \\ &= np \end{aligned}$$

9. Let X be an exponential random variable with pdf $f(X = x) = \lambda e^{-\lambda x}$ $(x > 0)$. What is $E(X)$? What is $E(X^2)$? Recall: integration by parts:

$$\int udv = uv - \int vdu$$

Probability Distributions V

Answer.

$$\begin{aligned} E(X) &= \int_0^\infty x\lambda e^{-\lambda x} dx \\ &= -xe^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx \\ &= \frac{1}{\lambda} \end{aligned}$$

Similarly, repeat to find $E(X^2)$.

10. A continuous real-valued variable has a power-law p.d.f. if $p(x) = Cx^{-\alpha}$ $(\alpha > 0)$. In fact, this function diverges as $x \to 0$: so how can it be a p.d.f. ?

Probability Distributions VI

Answer. It cannot. In most real-world problems, the power-law is only a good fit after some minimum value of $x = x_{min}$. So, as a first approximation, we can take the relative frequencies of such variables as being modelled by this (modified) p.d.f.

$$p(x) = \begin{cases} 0 & \text{if } x < x_{min} \\ \\ Cx^{-\alpha} & \text{otherwise} \end{cases}$$

11. Find an expression for C in the (modified) p.d.f. in the previous question.

Answer. Solving

$$\int_{x_{min}}^{\infty} Cx^{-\alpha} dx = 1$$

gives $C = (\alpha - 1){x_{min}}^{\alpha-1}$ for $\alpha > 1$.

12. Find the expected value for the random variable having the (modified) power-law p.d.f.

Answer. The expected value is:

$$\begin{aligned} E(X) &= \int_{x_{min}}^{\infty} xCx^{-\alpha} dx \\ &= C \int_{x_{min}}^{\infty} x^{-\alpha+1} dx \\ &= \frac{C}{2-\alpha} x^{2-\alpha} \Big|_{x_{min}}^{\infty} \end{aligned}$$

Probability Distributions VIII

13. Power-laws with $\alpha \leq 2$ have no finite mean. This means that as we start taking more and more samples from such populations, we will start to see the mean diverge. How can this happen?

    Answer. What must start to happen is every so often samples with a very large value of the mean must come up. That is, the fluctuation in the means must be very large.

14. Similarly show that for $\alpha \leq 3$, there is no finite variance.

    Answer. This follows straightforwardly from

$$\int_{x_{min}}^{\infty} x^2 p(x) dx$$

    There are many natural phenomena that exhibit power-law behaviours with divergent means or divergent variances (or both) like this.

Maximum Likelihood I

15. You have a sample of n observations $x_1, x_2, \ldots, x_n$ from data that appear to fit a binomial distribution
with parameters N and p. Assuming N is known, derive the maximum likelihood estimate for p in terms of N, n, and the x_i.

    Answer. The likelihood of observing x_1 successes from a binomial with parameters N, p is

$$p(x_1; N, p) = \binom{N}{x_1} p^{x_1} q^{(N-x_1)}$$

    where $q = 1 - p$

Maximum Likelihood II

So, the probability of observing $x_1, x_2, \ldots, x_n$ is:

$$\begin{aligned} L(p) &= \prod_i \binom{N}{x_i} p^{x_i} q^{(N-x_i)} \\ &= C p^{\sum_i x_i} q^{(Nn - \sum_i x_i)} \end{aligned}$$

This is the likelihood function to be maximised. Take logs of both sides, diffentiate and set to zero; and solve to find:

$$p_{ML} = \frac{1}{Nn} \sum_i x_i$$

16. Let $x_1, x_2, \ldots, x_n$ be a sample of observations from a Poisson distribution with parameter λ. Find the maximum likelihood estimate of λ in terms of the x_i and n.

Maximum Likelihood III

Answer. The likelihood function is the probability of observing $x_1, x_2, \ldots, x_n$ in n independent draws from a Poisson distribution. This is:

$$L(\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{x_i}}{x_i!}$$

That is:

$$L(\lambda) = \frac{e^{-n\lambda}\lambda^{\sum_i x_i}}{x_1! \cdots x_n}$$

Taking logs of both sides, gives:

$$\log L(\lambda) = -n\lambda + \sum x_i \log\lambda - \log c$$

Differentiate and set to zero, to find that the maximum likelihood estimate of λ is:

$$\lambda_{ML} = \frac{\sum_i x_i}{n}$$

Maximum Likelihood IV

17. Let $x_1, x_2, \ldots, x_n$ be a sample from an exponential distribution, which has a density function $f(X = x) = \lambda e^{-\lambda x}$ $(x > 0)$. Derive a maximum likelihood estimate of λ in terms of the x_i and n.

    Answer. The likelihood function for continuous random variables is the joint p.d.f. So, for an exponential p.d.f, this is:

$$L(\lambda) = \prod_{i=1}^{n} \lambda e^{-\lambda x_i}$$

    That is:

$$L(\lambda) = \lambda^n e^{-\lambda \sum x_i}$$

Maximum Likelihood V

Going through the usual process of taking logs, differentiating and setting to zero gives:

$$\lambda_{ML} = \frac{n}{\sum x_i}$$

18. Let $x_1, x_2, \ldots, x_n$ be observations from a normal distribution with parameters μ and σ^2. Derive maximum likelihood estimates of μ and σ^2.

    Answer. Go through the following steps:

    18.1 Write down the p.d.f. for a normal distribution.

    18.2 Now write down the joint p.d.f. for the x_i. This will be a product of the individual p.d.f.'s

    18.3 Simplify the product by removing out constants and translating the $\prod$ into $\sum$

    18.4 Now take logs of both sides

Maximum Likelihood VI

18.5 Since there are now two variables, we have to find separately the partial derivatives of the log likelihood w.r.t. μ and w.r.t σ.

18.6 Set each of these partial derivatives to 0 and solve the equations to get pressions for the maximum likelihood estimates.

If you follow this procedure, you should find the m.l.e. for μ is:

$$\mu_{ML} = \frac{\sum x_i}{n}$$

and for σ is:

$$\sigma_{ML} = \frac{\sum (x_i - \mu_{ML})^2}{n}$$

Logistic Regression I

Simple limear regression deals with the problem of fitting a line $Y = a + bX$ for a set of points $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The least-squares estimates of b and a are:

$$b = \sum(x_i y_i) / \sum {x_i}^2$$

where $x_i = (X_i - \overline{X})$ and $y_i = (Y_i - \overline{Y})$; and

$$a = \overline{Y} - b\overline{X}$$

This extends naturally to *weighted* simple linear regression, in which each point has a weight w_i. The least-square estimates of b and a are then:

$$b = \sum w_i(x_i y_i) / \sum w_i {x_i}^2$$

Logistic Regression II

where $x_i = (X_i - \overline{X}_w)$ and $y_i = (Y_i - \overline{Y}_w)$; and

$$a = \overline{Y_w} - b\overline{X_w}$$

The means are now weighted averages:

$$\overline{X}_w = \frac{\sum w_i x_i}{\sum w_i} \qquad \overline{Y}_w = \frac{\sum w_i y_i}{\sum w_i}$$

Clearly, if the $w_i = 1$, the ordinary linear regression results.

19. We will use weighted linear regression to build a linear model for the log-odds of Y when Y takes on one of two values: 0 and 1. For any value of $X = X_i$, the *odds* of Y (actually the odds of $Y|X = X_i$) is the ratio $P(Y = 1|X = X_i)/P(Y = 0|X = X_i)$. It is therefore simply the ratio of the number of $Y = 1$ entries for $X = X_i$ to the

Logistic Regression III

number of $Y = 0$ entries for $X = X_i$. This procedure is *simple logistic regression*.
Here is a partially completed table about a dataset:

i	ii	iii	iv	v	vi	vii
X	Y		Total	$P(Y = 1\|X)$	$Odds(Y)$	$LogOdds(Y)$
	0	1				
28	4	2				
29	3	2				
30	2	7				
31	2	7				
32	4	16				
33	1	14				

(a) Complete the table. (b) Using the total for each X_i, Y_i as the weight w_i, obtain the weighted linear regression line $LogOdds(Y) = a + bX$. (c) What is the predicted probability for $X = 31$?

Logistic Regression IV

Answer. This problem is from *Vassar Stats*.
(a) Here is the completed table:

i	ii	iii	iv	v	vi	vii	viii
X	Y		Total	$P(Y=1\|X)$	$O(Y)$	$LO(Y)$	Wt.
	0	1					
28	4	2	6	0.3333	0.5000	-0.6931	6
29	3	2	5	0.4000	0.6667	0.4055	5
30	2	7	9	0.7778	3.5000	1.2528	9
31	2	7	9	0.7778	3.5000	1.2528	9
32	4	16	20	0.8000	4.0000	1.3863	20
33	1	14	15	0.9333	14.0000	2.6391	15

(b) Calculation using the formulae for weighted linear regression yields $a = -17.2086$ and $b = 0.5934$
(c) For $X = 31$, $LogOdds(Y) = -17.2086 + (0.5934 \times 31) = 1.1868$. The corresponding $Odds(Y) = exp(LogOdds(Y)) =$

Logistic Regression V

3.2766. Solving for the probability gives $P(Y = 1|X = 31) = 0.7662$.

Simple Bayes I

20. The following table represents data collected by some machine-learning researchers at Wimbledon.

Day	Outlook	Temperature	Humidity	Wind	Play
1	Sunny	Hot	High	Weak	No
2	Sunny	Hot	High	Strong	No
3	Overcast	Hot	High	Weak	Yes
4	Rain	Mild	High	Weak	Yes
5	Rain	Cool	Normal	Weak	Yes
6	Rain	Cool	Normal	Strong	No
7	Overcast	Cool	Normal	Strong	Yes
8	Sunny	Mild	High	Weak	No
9	Sunny	Cool	Normal	Weak	Yes
10	Rain	Mild	Normal	Weak	Yes
11	Sunny	Mild	Normal	Strong	Yes
12	Overcast	Mild	High	Strong	Yes
13	Overcast	Hot	Normal	Weak	Yes
14	Rain	Mild	High	Strong	No

Simple Bayes II

From Bayes' Rule (with some simplification of notation):

$$
\begin{aligned}
P(Yes|Sunny, Cool, High, Strong) &= \frac{P(Yes)P(Sunny, Cool, High, Strong|Yes)}{P(Sunny, Cool, High, Strong)} \\
&\propto P(Yes)P(Sunny, Cool, High, Strong|Yes) \\
P(No|Sunny, Cool, High, Strong) &= \frac{P(No)P(Sunny, Cool, High, Strong|No)}{P(Sunny, Cool, High, Strong)} \\
&\propto P(No)P(Sunny, Cool, High, Strong|No)
\end{aligned}
$$

Assume that the attributes Outlook, Temperature, Humidity and Wind are conditionally independent of each other given the value of the target attribute Play.
Using the data recorded, estimate the probability of play on Day 15, which has the following forecast:

$$\langle Outlook = Sunny, Temperature = Cool, Humidity = High, Wind = Strong \rangle$$

Simple Bayes III

Answer. With the conditional independence assumption the conditional expression for *Play* becomes:

$$\begin{aligned} P(Yes|Su, C, H, St) &\propto P(Yes)P(Su|Yes)P(C|Yes)P(H|Yes)P(St|Yes) \\ P(No|Su, C, H, St) &\propto P(No)P(Su|No)P(C|No)P(H|No)P(St|No) \end{aligned}$$

Each of the probabilities can now be estimated from the frequencies observed in the table:

$$\begin{aligned} P(Yes) &= 9/14 = 0.64 \\ P(No) &= 5/14 = 0.36 \\ &\dots \\ P(St|Yes) &= 3/9 = 0.33 \\ P(St|No) &= 3/5 = 0.60 \\ &\dots \end{aligned}$$

Simple Bayes IV

yielding:

$$\begin{aligned} P(Yes|Su, C, H, St) &\propto 0.0053 \\ P(No|Su, C, H, St) &\propto 0.0206 \end{aligned}$$

Normalising, we obtain
$P(Yes|Su, C, H, St) = 0.0053/(0.0053 + 0.0206)$
$= 0.205$. Therefore, there is a 20% chance of play on Day 15.