J Kohonen

1A Basic rules of probability

In this first exercise there are four class problems, and two home problems. It is probably useful to consult the Ross textbook (see link on course page) and/or the lecture slides (lecture 1A).

Class problems

1A1 (Rolling a die twice) A regular six-sided die is rolled twice. The set of all possible results (the *sample space*) is the following:

$$\begin{aligned} S &= \{(x,y) : x = 1,2,3,4,5,6 \text{ and } y = 1,2,3,4,5,6\} \\ &= \{(1,1),(1,2),(1,3),\dots,(6,6)\}, \end{aligned}$$

where x = result from the first roll, and y = result from the second roll. Thus the sample space has 36 elements, each of which is a pair of integers. For example, the element $(2,3)$ means that we first rolled a two, and then a three. Now *draw* the sample space as a grid of 6×6 cells. For each of the following events, *color* the cells corresponding to the event, and *determine* the probability of the event.

(a) $A = \{(x,y) \in S : x = 1\}$,

(b) $B = \{(x,y) \in S : y \geq 4\}$,

(c) $C = \{(x,y) \in S : x + y = 7\}$,

(d) $A \cup B$ = union of A and B,

(e) $B \cap C$ = intersection of B and C,

(f) B^c = complement of B.

To determine the probability of an event, count its cells, observe that each cell has probability 1/36, and add them up. Recall what union, difference, and complement mean in the context of sets. (Ross section 3.3, or lecture slides 1A.)

Solution. The following descriptions assume that x determines the column and y determines the row in the drawing. (The choice is arbitrary.)

(a) Event A covers a vertical column $x = 1$, which contains 6 cells. So $\mathrm{P}(A) = 6 \times \frac{1}{36} = 1/6$.

(b) Event B covers three horizontal rows, namely the rows $y = 4$, $y = 5$ and $y = 6$. Together the rows contain 18 cells, so $\mathrm{P}(B) = 18 \times \frac{1}{36} = 1/2$.

(c) Event C covers the cells $(1,6),(2,5),(3,4),(4,3),(5,2),(6,1)$. That is 6 cells, so $\mathrm{P}(C) = 6 \times \frac{1}{36} = 1/6$. Here it is useful to think *how* one can find the cells easily. Of course, one could check all 36 cells separately, to see where $x + y = 7$ happens to be true. Much better is this: if x has some value, then what must y be in order for the equation $x + y = 7$ to be true? Indeed, y must be $7 - x$, so we need not try any other values for y.

(d) Event $A \cup B$ is obtained by coloring the one vertical column for A, and the three horizontal rows for B, with the same color. We end up coloring 21 cells, so $\mathrm{P}(A \cup B) = 21 \times \frac{1}{36} = 21/36$.

(e) Event $B \cap C$ is obtained by coloring only those cells that belong to both B and C; that is the cells $(3,4)$, $(2,5)$ and $(1,6)$. Thus $\mathrm{P}(B \cap C) = 3 \times \frac{1}{36} = 1/12$.

(f) Event B^c is the complement of B, so it covers the horiztonal rows $y = 1$, $y = 2$ and $y = 3$. Thus $\mathrm{P}(B^c) = 18 \times \frac{1}{36} = \frac{1}{2}$.

1A2 (Soil samples) We analyze a soil sample from a waste dump. With probability 0.40 we find arsenic. With probability 0.30 we find lead. With probability 0.10 we find both.

Note that “finding arsenic” does not mean “finding arsenic only”. If we find arsenic, we may or may not find *also* lead.

(a) What is the probability that we find lead but not arsenic?

(b) What is the probability that we find arsenic but not lead?

(c) What is the probability that we find at least one of them?

(d) What is the probability that we find neither?

Solution. Let us name some events:

$$\begin{aligned} A &= \text{“we find arsenic”}, \\ L &= \text{“we find lead”}. \end{aligned}$$

Now there are four (mutually exclusive) possibilities of what we may find: both substances ($A \cap L$), arsenic only ($A \cap L^c$), lead only ($A^c \cap L$), or neither ($A^c \cap L^c$). It may be useful to draw these as a 2×2 grid, so that A and A^c are the rows, and L and L^c are the columns. From the problem statement, we have $\mathrm{P}(A) = 0.40$, $\mathrm{P}(L) = 0.30$ ja $\mathrm{P}(A \cap L) = 0.10$.

(a) We observe that the event L is the union of two disjoint events $L \cap A$ and $L \cap A^c$. That is, finding lead is equivalently to finding *either* “lead and arsenic” *or* “lead but not arsenic”. By the addition rule (e.g. Ross section 3.4),

$$\mathrm{P}(L) = \mathrm{P}(L \cap A) + \mathrm{P}(L \cap A^c),$$

from which can solve

$$\begin{aligned} \mathrm{P}(L \cap A^c) &= \mathrm{P}(L) - \mathrm{P}(L \cap A) \\ &= 0.30 - 0.10 = 0.20. \end{aligned}$$

(Note that $L \cap A$ is the same event as $A \cap L$.)

(b) In the same way as in the previous item, we observe that

$$\mathrm{P}(A) = \mathrm{P}(A \cap L) + \mathrm{P}(A \cap L^c),$$

from which can solve

$$\begin{aligned} \mathrm{P}(A \cap L) &= \mathrm{P}(A) - \mathrm{P}(A \cap L^c) \\ &= 0.40 - 0.10 = 0.30. \end{aligned}$$

(c) This is the union event $A \cup L$. We can do this in many different ways. For example, by the general addition rule (Ross Proposition 3.4.2),

$$\begin{aligned} P(A \cup L) &= P(A) + P(L) - P(A \cap L) \\ &= 0.40 + 0.30 - 0.10 = 0.60. \end{aligned}$$

Or, we may add up the three (mutually exclusive) possibilities "lead only" (0.20 from item a), "arsenic only" (0.30 from item b), and "lead and arsenic" (0.10 given in problem statemnt), thus $0.20 + 0.30 + 0.10 = 0.60$.

(d) Finding neither substance is the complement of finding at least one, so it has the probability $1 - 0.60 = 0.40$.

1A3 (Blue cab) In a certain town there are 100 taxicabs: one is blue, and 99 are green. One night, a cab collides with a bicycle, and flees the scene. An eye witness says that the colliding cab was *blue*. However, from previous research we know that in similar situations, a blue car is seen as blue with probability 90%; and a *green* car is seen as blue with probability 8%.

What is the probability that the colliding cab was indeed blue? Do you think that the driver of the only blue cab in town should be convicted, based on this evidence alone?

Solution. Let us denote

- T_0 = "the colliding cab is green",
- T_1 = "the colliding cab is blue",
- H_0 = "the witness sees the cab as green,
- H_1 = "the witness sees the cab as blue".

The prior probabilities of the events T_0 ja T_1 (before taking the witness report into account) are $P(T_0) = 0.99$ and $P(T_1) = 0.01$. We also know that $P(H_1 \mid T_0) = 0.08$ and $P(H_1 \mid T_1) = 0.90$. Applying the law of total probability,

$$\begin{aligned} P(H_1) &= P(H_1 \mid T_0)\,P(T_0) + P(H_1 \mid T_1)\,P(T_1) \\ &= 0.08 \times 0.99 + 0.90 \times 0.01 \\ &= 0.0882. \end{aligned}$$

Thus, according to Bayes's rule, the posterior probability of T_1 (after accounting for the witness report) is

$$P(T_1 \mid H_1) = \frac{P(H_1 \mid T_1)\,P(T_1)}{P(H_1)} = \frac{0.90 \times 0.01}{0.0882} \approx 0.102.$$

From this evidence, it is *more probable* that the colliding cab was green ($\approx 89.8\%$) than blue ($\approx 10.2\%$). Thus there does not seem to be reason to convict the driver of the blue cab!

1A4 (Sampling) A small village has 120 inhabitants, and 20 of them landowners. For convenience we label the inhabitants with integers $1, 2, \ldots, 120$ so that $1, 2, \ldots, 20$ are the landowners.

(a) One inhabitant is picked at random. What is the probability that (s)he is a landowner? Call this number p.

(b) **Sampling with replacement.** Three times we do this: Pick an inhabitant at random from the 120 inhabitants. (We can pick the same inhabitant again.) Let the inhabitants, in the order we picked them, be (i, j, k). How many different choices are there? Call this number N.

(c) In the previous item's scenario, how many different choices there are where all three inhabitants are landowners? Call this number A.

(d) Calculate A/N to five decimals. This is the probability that we pick three landowners. Compare to p^3 and explain. Hint: Think of rolling a die.

(e) **Sampling without replacement.** We pick three inhabitants (i, j, k) thus: The first is chosen at random from all inhabitants. The second is chosen from the remaining ones, and then the third from the remaining ones. How many different choices are there? Call this number M.

(f) In the scenario of the previous item, how many different choices are there where all three inhabitants are landowners? Call this number B.

(g) Calculate B/M to five decimals. This is the probability that we pick three landowners. Compare to previous results and explain.

Hint: Ordered sequences (lecture 1A), principle of counting (Ross §3.5).

Solution.

(a) Each inhabitant has the same probability $1/120$ of being picked, so by additivity, we pick a landowner with probability $20 \cdot (1/120) = 20/120 = 1/6$.

(b) For i there are 120 choices. For j again 120 and for k again 120, so $N = 120^3 = 1728000$ choices.

(c) For i there are 20 choices (the landowners). For j again 20, and for k again 20, so $A = 20^3 = 8000$ choices.

(d) $A/N = 8000/1728000 \approx 0.00463$. This equals p^3, because picking an inhabitant is like rolling a die, with $1/6$ chance of picking a landowner.

(e) For i there are 120 choices, *then* for j 119 choices, then for k 118 choices. Note that the remaining *choices* for j are different sets of people, depending on who was picked as i, but their *number* is always the same 119. So $M = 120 \cdot 119 \cdot 118 = 1685040$ choices.

(f) For i there are 20 choices. Then for j there are 19 choices (the remaining landowners), and then for k 18 choices. So $B = 20 \cdot 19 \cdot 18 = 6840$ choices.

(g) $B/M = 6840/1685040 \approx 0.00406 = 0.406\%$. This is *smaller* than p^3 because after having picked some landowners, the remaining choices are limited and there are *relatively fewer* landowners left. Indeed we could have calculated B/M also successively as

$$\frac{20}{120} \cdot \frac{19}{119} \cdot \frac{18}{118}.$$

Sampling "with" and "without" replacement is statistical jargon for taking a sample from a population. We think of the population as a box containing balls. After picking the first ball, either we *put it back to the box* ("re-place" it there), so we pick the next ball from the same population. Or we do not put the ball back ("without re-placement") so the next ball is picked from a smaller population.

Home problems

1A5 (Slot machines) At a casino there are two kinds of slot machines, which look quite indentical from the outside. In machines of type A, the winning probability is 5%. In machines of type B, the winning probability is 7%. We also know that 70% of the machines are of type A, and 30% are of type B.

(a) A gambler picks a slot machine at random, and plays once. What is the probability that he wins?

(b) If the gambler wins, what is the probability that the machine is of type B?

Grading. 1 point from each item. No penalty for small errors in calculation or rounding.

Solution.

(a) Let us denote

$$\begin{aligned} A_A &= \text{“slot machine is of type A”}, \\ A_B &= \text{“slot machine is of type B”}, \\ W_1 &= \text{“gambler wins”}, \\ W_0 &= \text{“gambler does not win”}. \end{aligned}$$

From the problem statement we find

$$\begin{aligned} \mathrm{P}(W_1 \mid A_A) &= 0.05, \\ \mathrm{P}(W_1 \mid A_B) &= 0.07, \\ \mathrm{P}(A_A) &= 0.70, \\ \mathrm{P}(A_B) &= 0.30. \end{aligned}$$

There are two ways of winning: either the machine is of type A and the player wins, or the machine is of type B and the player wins. These cases are clearly mutually exclusive, so we can apply additivity.

$$\begin{aligned} \mathrm{P}(W_1) &= \mathrm{P}(W_1 \mid A_A)\,\mathrm{P}(A_A) + \mathrm{P}(W_1 \mid A_B)\,\mathrm{P}(A_B) \\ &= 0.05 \times 0.70 + 0.07 \times 0.30 \\ &= 0.056 \\ &= 5.6\%. \end{aligned}$$

(b) Using the Bayes' formula and the result from (a), we have

$$\mathrm{P}(A_B \mid W_1) = \frac{\mathrm{P}(A_B)\,\mathrm{P}(W_1 \mid A_B)}{\mathrm{P}(W_1)} = \frac{0.30 \times 0.07}{0.056} = 0.375 \approx 38\%.$$

1A6 (Strings of letters) On a certain planet, there are 4^{10} inhabitants, each of which has a unique 10-letter identifier. Each letter is either A, B, C or D, so there are 4^{10} different identifiers, and every identifier is currently in use. Identifiers look like CBAADAABBB or AADABCAACC.

(a) How many of the identifiers are palindromes, that is, the same when read left-to-right and right-to-left? (AABCAACBAA is an example of a palindrome.)

(b) How many identifiers are such that no two consecutive letters are identical? (For example, ABADABACAD is like this, but AACBDBABCB is not.)

An inhabitant of the planet is now chosen uniformly at random (that is, each inhabitant has the same probability of being chosen). What is the probability that the identifier of the chosen inhabitant

(c) is a palindrome?

(d) has no two consecutive identical letters?

It is probably useful to apply the *generalized basic principle of counting* (Ross section 3.5, or lecture slides 1A).

Grading. 0.5 points per item if correct. Total points are rounded up to nearest integer.

Solution.

(a) If an identifier is a palindrome, then its left half (5 letters) determines the second half (5 letters). For example, if a palindrome starts AADAC, then it must end with CADAA.

We also observe that the first 5 letters can be chosen freely. Then we can always set the second half as the reverse of the first half, and we indeed get a palindromic identifier. The first 5 letters can be chosen in 4^5 ways out of the four letter ABCD, so there are 4^5 different first halves. Each one of them gives us a 10-letter palindrome.

We are not quite done yet. We must also observe that two different first halves will always give two *different* palindromes. Thus there number of different palindromes is *equal* to the number of different first parts.

(In fact, we have created a *bijective* mapping between the two sets — the set of first parts, and the set of palindromes. A bijective mapping proves that the two sets have equally many elements.)

So there are $4^5 = 1\ 024$ different 5-letter first halves, and also, 1 024 different 10-letter palindromes.

(b) We apply the rule of multiplication, or "generalized basic principle of counting", creating an identifier one letter at a time.

- Choose the 1st letter. There are *four* choices here.
- Choose the 2nd letter. Whatever the first letter was, there are *three* remaining choices for the second.

- Choose the 3rd letter. Whatever the first and second letters were, the third letter has *three* choices (those that are different from second).
- ...
- Choose the 10th letter. Whatever the previous choices were, there are *three* choices here, namely those that are different from the 9th letter.

This process has $4 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 \cdot 3 = 4 \cdot 3^9 = 78\ 732$ different outcomes, so there are 78 732 identifiers that do not contain consecutive identical letters.

(c) There are $4^{10} = 1\ 048\ 576$ inhabitants, and 1 048 have palindromic identifiers by item (a), so a random inhabitant has a palindrome with probability

$$\frac{1\ 024}{1\ 048\ 576} \approx 0.098\ \%.$$

(d) In item (b) we counted such identifiers. The probability that a random inhabitant has such an identifier is

$$\frac{78\ 732}{1\ 048\ 576} \approx 7.51\ \%.$$