INFORMATION THEORY
University of Amsterdam, 2018
Team Name: Latin square

Homework set 1

Contents

Question 1

(a)
- $P_X(x) = \begin{cases} \frac{1}{4} & \text{if } x \in \{0, 2\} \\ \frac{2}{4} = \frac{1}{4} & \text{if } x = 1 \\ 0 & \text{otherwise} \end{cases}$
- $P_Y(y) = \begin{cases} \frac{1}{4} & \text{if } y \in \{1, 2, 3, 4\} \\ 0 & \text{otherwise} \end{cases}$
- For example, the probability of $z = 2.5$ is computed by taking the sum of the different probabilities of z being 2.5, e.g. we can get $z = 2.5$ by $x = 4$ and $y = 1$ which has probability $\frac{1}{4} * \frac{2}{4} = \frac{2}{16}$, or by $x = 3$ and $y = 2$ which has probability $\frac{1}{4} * \frac{1}{4} = \frac{1}{16}$. Therefore the total probability would be $\frac{3}{16}$. This results in:

$$P_Z(z) = \begin{cases} \frac{1}{16} & \text{if } z \in \{0.5, 3\} \\ \frac{3}{16} & \text{if } z \in \{1, 2.5\} \\ \frac{4}{16} = \frac{1}{4} & \text{if } z \in \{1.5, 2\} \\ 0 & \text{otherwise} \end{cases}$$

(b)
- $mean(x) = 0 * \frac{1}{4} + 1 * \frac{2}{4} + 2 * \frac{1}{4} = 1$
  $variance(x) = (0-1)^2 * \frac{1}{4} + (1-1)^2 * \frac{2}{4} + (2-1)^2 * \frac{1}{4} = \frac{1}{2}$
- $mean(y) = 1 * \frac{1}{4} + 2 * \frac{1}{4} + 3 * \frac{1}{4} + 4 * \frac{1}{4} = 2.5$
  $variance(y) = (1-2.5)^2 * \frac{1}{4} + (2-2.5)^2 * \frac{1}{4} + (3-2.5)^2 * \frac{1}{4} + (4-2.5)^2 * \frac{1}{4} = 1\frac{1}{4}$
- $mean(z) = 0.5 * \frac{1}{16} + 1 * \frac{3}{16} + 1.5 * \frac{4}{16} + 2 * \frac{4}{16} + 2.5 * \frac{3}{16} + 3 * \frac{1}{16} = 1.75$
  $variance(y) = (0.5-1.75)^2 * \frac{1}{16} + (1-1.75)^2 * \frac{3}{16} + (1.5-1.75)^2 * \frac{4}{16} + (2-1.75)^2 * \frac{4}{16} + (2.5-2.5)^2 * \frac{3}{16} + (3-2.5)^2 * \frac{1}{16} = \frac{1}{4}$

(c)
- Probability of winning with $x = 1$ (and $y \leq 2$) : $\frac{2}{4} * \frac{2}{4} = \frac{1}{4}$
- Probability of winning with $x = 2$ (and $y \leq 4$) : $\frac{1}{4} * 1 = \frac{1}{4}$
- So probability of winning is $\frac{2}{4}$, therefore the probability of losing is $1 - \frac{2}{4} = \frac{2}{4}$

In 40 games, you will on average win $\frac{1}{4} * 40 * 1 = 10$ euro (for $x = 1$), win $\frac{1}{4} * 40 * 4 = 40$ euro for $x = 2$ and loose $\frac{2}{4} * 40 * 1 = 30$ euro (for $x = 0$). So on average you will win $10 + 40 - 20 = 30$ euros out of 40 games.

Question 2

(a)

$$\mathbb{E}[aX + bY] = \sum_x \sum_y [(ax + by) \cdot P(X = x, Y = y)]$$

By the distributive law this becomes:

$$= \sum_x \sum_y [ax \cdot P(X = x, Y = y)] + \sum_x \sum_y [by \cdot P(X = x, Y = y)]$$

$$= a \sum_x \sum_y [x \cdot P(X = x, Y = y)] + b \sum_x \sum_y [y \cdot P(X = x, Y = y)]$$

$$= a \sum_x \sum_y [x \cdot P(X = x, Y = y)] + b \sum_y \sum_x [y \cdot P(X = x, Y = y)]$$

$$= a \sum_x x \sum_y [P(X = x, Y = y)] + b \sum_y y \sum_x [P(X = x, Y = y)]$$

By the marginal probability mass function we get:

$$= a \sum_x x \cdot P(X = x) + b \sum_y y \cdot (Y = y)$$

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

Since we did not make use of independence, and $x, y \in \mathbb{R}$ were chosen arbitrarily and X, Y were arbitrary, then we have shown for any random variable that the above equality holds. Q.E.D.

(b)

$$\begin{aligned} var[aX] &= \mathbb{E}[(aX - \mathbb{E}[aX])^2] \\ &= \mathbb{E}[(aX - a\mathbb{E}[X])^2] \\ &= \mathbb{E}[a^2(X - \mathbb{E}[X])^2] \\ &= a^2 var(X) \end{aligned}$$

This follows arithmetically from the definition of variance. Q.E.D.

(c)

$$var[X + Y] = \mathbb{E}[((X + Y) - \mathbb{E}[X + Y])^2]$$

which is equivalent to:

$$= \mathbb{E}[(X + Y)^2] - \mathbb{E}[X + Y]^2$$

$$= \mathbb{E}[X^2 + 2XY + Y^2] - (\mathbb{E}[X]^2 + 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[Y]^2)$$

it was derived in (2a) that expectation is linear for any variable, hence:

$$= \mathbb{E}[X^2] + \mathbb{E}[2XY] + \mathbb{E}[Y^2] - \mathbb{E}[X]^2 - 2\mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[Y]^2$$

$$= \mathbb{E}[X^2] - \mathbb{E}[X]^2 + \mathbb{E}[Y^2] - \mathbb{E}[Y]^2 + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y]$$

$$= var[X] + var[Y] + 2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y]$$

However, since X and Y are independent we have that

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x \in supp(X)} \sum_{y \in supp(Y)} P(X = x, Y = y)xy \\ &= \sum_{x \in supp(X)} \sum_{y \in supp(Y)} P(X = x)P(Y = y)xy \\ &= \sum_{x \in supp(X)} P(X = x)x \sum_{y \in supp(Y)} P(Y = y)y \\ &= \mathbb{E}[X]\mathbb{E}[Y] \end{aligned}$$

Hence we see that $2\mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y]$ cancel, and therefore:

$$var[X + Y] = var[X] + var[Y].$$

Of course if X and Y are dependent then the terms do not cancel. Since this is exactly how covariance is defined: $cov(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]$ Consider this counterexample:
The variable X takes on values $(X = 2), (X = 3.3), (X = 3.8), (X = 5.1)$ and the variable Y takes on values $(Y = 4), (Y = 6), (Y = 8), (Y = 10)$. Now without having to compute covariance, we see that there will be some positive correlation between X and Y which tells us that X and Y are not mutually independent hence their covariance is not equal to 0. Thus:

$$var[X + Y] = var[X] + var[Y] + 2cov[X, Y].$$

Question 3

(a) The two quantities we want to compare are $E[X] = \int_0^\infty xp(x)dx$, and $P(X \geq t) = \int_t^\infty p(x)dx$. Note that p indicates the probability density function of the random variable X. We can write down the definition of $E[X]$ and obtain

$$E[X] = \int_0^\infty xp(x)dx = \int_0^t xp(x)dx \int_t^\infty xp(x)dx \geq \int_t^\infty xp(x)dx \geq \int_t^\infty tp(x)dx$$
$$= t\int_t^\infty p(x)dx = tp(X \geq t) \implies p(X \geq t) \leq \frac{E[X]}{t}$$

(b) We can find a simple distribution by looking at the inequalities in the above exercise. In order for the first inequality to be an equality, we need to have $\int_0^t xp(x)dx = 0$, so no probability mass has to be assigned to values less than t.

For the second inequality, we can obtain an equality by having a probability distribution which takes value t with probability one, so that $E[X] = 1$ and both sides of the markov inequality become equal to one.

Notice that with this solution, we can only design the desired probability distribution once t is known.

(c) First, we use the markov inequality with random variable $x := (Y - \mu)^2$ and constant $t = \epsilon^2$. From the definition of exercise 3.a, we have then

$$p[(Y - \mu)^2 \geq \epsilon^2] \leq \frac{E[(Y - \mu)^2]}{\epsilon^2} = \frac{\sigma^2}{\epsilon^2}$$

We can now note that for any random variable Z, we have that $p(|Z| \geq a) = p(Z^2 \geq a^2)$. We can then rewrite the previous equation and obtain that

$$p[|Y - \mu| \geq \epsilon] = p[(Y - \mu)^2 \geq \epsilon^2] \leq \frac{\sigma^2}{\epsilon^2}$$

(d) We know that the following identity holds $\text{VAR}(aX) = a^2\text{VAR}(X)$. Also, we know that for independent random variables X and Y, $\text{VAR}(X + Y) = \text{VAR}(X) + \text{VAR}(Y)$. We write down the variance for S_n and, as Z_i are independent to each other, we obtain

$$\text{VAR}(S_n) = \text{VAR}(\frac{1}{n}\sum_{i=1}^n Z_i) = \sum_{i=1}^n \text{VAR}(\frac{Z_i}{n}) = \sum_{i=1}^n \frac{\sigma^2}{n^2} = n\frac{\sigma^2}{n^2} = \frac{\sigma^2}{n}$$

Thus, by plugging in S_n in the Chebychev inequality, we obtain

$$p[|S_n - \mu| \geq \epsilon] \leq \frac{E[(S_n - \mu)^2]}{\epsilon^2} = \frac{\text{VAR}(S_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

Question 4

Let k denote that our student knows the answer to a question, and c denote that the answer is correct. Our goal is to determine what the probability is that knew the answer, given that she answers correctly. Or rather,

$$P(k|c) = \frac{P(k \cap c)}{p(c)}$$

We know that $P(k \cap c)$ is 0.5, so let's determine the probability of a correct answer. We know that the student knows half, or 0.5, of the answers, and thus answers them correctly. Since the student chooses the correct answers with a probability of 0.25 and 0.33 in 75% and 25% of the cases, respectively, the probability of answering correctly is as follows.

$$P(c) = 0.5 + 0.75 \times 0.25 + 0.25 \times 0.33$$

The final calculation is as follows.

$$P(k|c) = \frac{0.5}{0.771} = 0.649$$

Question 5

(a) Based on the guesses of our five group members:

- Programmers worldwide (millions): $(7 + 100 + 100 + 50 + 150)/5 = 81.4$
- Probability for programmers of using git: $(\frac{4}{5} + \frac{1}{10})/5 = 0.18$
- Probability of using git on any given day: $(\frac{1}{7} + \frac{2}{7} + \frac{2}{7} + \frac{2}{7} + \frac{3}{7})/5 = 0.286$
- Number of commits per day per git user: $(2 + 4 + 3 + 2 + 1)/5 = 2.4$

Thus, amount of git-commits per day worldwide: $81.4 * 0.18 * 0.286 * 3.6 = 15.07$ million.

(b) Github commits in October 2016: 450 million. Amount of github users in October 2016: 6 million. So average commit per github user in 2016 was 75 per month, so on average 2.5 a day. Amount of github users in october 2018 is 31 million. So on average there will be $31 * 2.5 = 77.5$ million commits a day.

Question 6

(a) **Compute all pairwise variational distances**

- $||P_{eng} - P_{ita}|| = 0.15514534928687276$
- $||P_{eng} - P_{esp}|| = 0.20937742236938414$
- $||P_{eng} - P_{fin}|| = 0.2615411711708111$
- $||P_{eng} - P_{ger}|| = 0.14376680719328008$
- $||P_{ita} - P_{esp}|| = 0.10602051098152043$
- $||P_{ita} - P_{fin}|| = 0.23652833937220297$
- $||P_{ita} - P_{ger}|| = 0.1919016369551589$
- $||P_{esp} - P_{fin}|| = 0.15958260860191312$
- $||P_{esp} - P_{ger}|| = 0.23740749351219056$
- $||P_{fin} - P_{ger}|| = 0.3019905450860071$

The closest languages are Italian and Spanish by **total variation distance**. This make sense, because these languages are similar. In the other hand, Finnish and German languages are furthest apart by **total variation distance**.

(b) **Compute the five collision probabilities**

- $Coll(P_{eng}) = 6.554726160032556 * 10^{-2}$
- $Coll(P_{ita}) = 7.128790076508662 * 10^{-2}$
- $Coll(P_{esp}) = 6.994154478831635 * 10^{-2}$
- $Coll(P_{fin}) = 7.674919239984539 * 10^{-2}$
- $Coll(P_{ger}) = 7.249092263220734 * 10^{-2}$

We can clearly see that numbers of Spanish an Italian language are close. These is also caused by their similarities.

(c) **Why is it called collision probability?** The term is called collision probability because it reflects the probability that two independent random outcomes from Ω will be the same.

(d) **Which language the original text was? (permuted_cipher.txt)**

- $||P_{cipher} - P_{eng}|| = 0.2117031983975982$
- $||P_{cipher} - P_{ita}|| = 9.858283372956739 * 10^{-2}$
- $||P_{cipher} - P_{esp}|| = 3.982758841882112 * 10^{-2}$
- $||P_{cipher} - P_{fin}|| = 0.1771581803080346$
- $||P_{cipher} - P_{ger}|| = 0.2433105311008943$

The Cipher text has the lowest **total variation distance** with Spanish. (We can also see that the Cipher text is also close to the Italian language. This is also caused by similarities of Spanish and Italian languages.)

(e) **Collision probability and permuted_cipher.txt**

- $Coll(P_{cipher}) = 7.008295280317044 * 10^{-2}$ This number is closest to the $Coll(P_{esp}) = 6.994154478831635 * 10^{-2}$. So using this method, we will also choose Spanish language.

Description of the code:

We choose haskell language, because its function background and its purity. (We can easily implement **collision probability** and **total variation distance** in this language.)

We firstly parsed the files. We count the lower and upper ASCII characters and upper characters were lowered and counted as lower characters. We didn't count spaces and special characters. In the end, our charset consists only this characters: "abcdefghijklmnopqrstuvwxyz" (lower English letters).

Sometimes can occur that the language doesn't use often letter from English lower ASCII set. This could happen during our parsing. (For example, Slovak language doesn't use x letter often, so there is high probability that in Slovak version Alice story we could not find any letter x.) That's why, we put 0 count for each missing letter, which results in the zero probabilities. (This is actually not the best solution, because there are maybe some words, which contain the missing letter, but were not used in the text. That's why some probability smoothing method should be better option.)

To check the correctness of the code, we run our solution on different text files, to check if we get similar numbers. Moreover, the correctness of **total variation distance** can be also checked. We calculated the $||P - P||$, which equals to 0.