**4171059-0: Data Science 1**

# Probabilities (5/09/18 )

*Instructor: Guillermo Cabrera*                    *Assistant: Cristobal Donoso*

# Problem 1

My neighbor has two children. Assuming that the gender of a child is like a coin flip, it is most likely, a priori, that my neighbor has one boy and one girl, with probability 1/2. The other possibilities -two boys or two girls- have probabilities 1/4 and 1/4.
   a. Suppose I ask him whether he has any boys, and he says yes. What is the probability that one child is a girl?
   b. Suppose instead that I happen to see one of his children run by, and it is a boy. What is the probability that the other child is a girl?

## Solution

a.
My neighbor has two children then we have 4 possible combinations. We define the sample space $\Omega$ as

$$\Omega = \{BB, BG, GB, GG\} \tag{0.1}$$

where $B$ and $G$ is a boy and girl respectively. So the probability of one child being girl is $\frac{1}{2}$. If he said me that he has one any boys, then $\Omega$ changes,

$$\Omega = \{BB, BG, GB\} \tag{0.2}$$

You're left with a $\frac{2}{3}$ probability that he has a girl. Formally,

$$P(G|B) = \frac{P(G \cap B)}{P(B)} = \frac{\frac{2}{3} \times 1}{1} = \frac{2}{3} \tag{0.3}$$

b.
Now we have seen a boy. We only need to estimate the probability associated to the other child. So, the sample space is :

$$\Omega = \{B, G\} \tag{0.4}$$

The probability related to the other child is $\frac{1}{2}$

# Problem 2

Show that the variance of a sum is VAR$[X + Y]$ = VAR$[X]$ + VAR$[Y]$ + 2cov$[X,Y]$ where cov$[X, Y]$ is the covariance between $X$ and $Y$

## Solution

By definition

$$E[X + Y] = E[X] + E[Y] \tag{0.5}$$

$$VAR[Z] = E[Z - E[Z]]^2 \tag{0.6}$$

$$COV(X, Y) = E[(X - E[X])(Y - E[Y])] \tag{0.7}$$

Let $Z = X + Y$ then 0.7 becomes

$$VAR[X + Y] = E[(X + Y) - E[X + Y]]^2 \tag{0.8}$$

using 0.5 in 0.8

$$VAR[X + Y] = E[(X + Y) - (E[X] + E[Y])]^2 \tag{0.9}$$

$$VAR[X + Y] = E[(X - E[X]) + (Y - E[Y])]^2 \tag{0.10}$$

$$VAR[X + Y] = E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] \tag{0.11}$$

$$VAR[X + Y] = E[(X - E[X])^2] + E[(Y - E[Y])^2] + 2E[(X - E[X])(Y - E[Y])] \tag{0.12}$$

Finally using 0.7 properties

$$VAR[X + Y] = VAR[X] + VAR[Y] + 2COV(X, Y) \tag{0.13}$$

Notice that when we assume 0 covariance we have

$$VAR[X + Y] = VAR[X] + VAR[Y] \tag{0.14}$$

# Problem 3

Let $H \in \{1, ..., K\}$ be a discrete random variable, and let $e_1$ and $e_2$ be the observed values of two other random variables $E_1$ and $E_2$. Suppose we wish to calculate the vector

$$\overrightarrow{P}(e_1, e_2) = (P(H = 1|e_1, e_2), ..., P(H = K|e_1, e_2)) \tag{0.15}$$

Which of the following sets of numbers are sufficient for the calculation?
   a.  i.  $P(e_1, e_2), P(H), P(e_1|H), P(e_2|H)$
      ii. $P(e_1, e_2), P(H), P(e_1, e_2|H)$
      iii. $P(e_1|H), P(e_2|H), p(H)$
   b. Now suppose we now assume $E_1 \perp E_2|H$ (i.e $E_1$ and $E_2$ are conditionally independent given H), Which of the above 3 sets are sufficient now?
Show your calculations as well as giving the final results. Hint : use Bayes rule

## Solution

a.
We do not have any independence assumption between variables. Then, using Bayes rule

$$P(H|e_1, e_2) = \frac{P(e_1, e_2|H)P(H)}{P(e_1, e_2)} \tag{0.16}$$

The second set(ii.) have the numbers will we need.

b.
Now, we know that $P(E_1|H, E_2) = P(E_1, H)$. Using Bayes rule,

$$P(H|E_1, E_2) = \frac{P(H, E_1, E_2)}{P(E_1, E_2)} \tag{0.17}$$

$$= \frac{P(E_1|H, E_2)P(H, E_2)}{P(E_1, E_2)} \tag{0.18}$$

$$= \frac{P(E_1|H)P(E_2, H)}{P(E_1, E_2)} \tag{0.19}$$

$$= \frac{P(E_1|H)P(E_2|H)P(H)}{P(E_1, E_2)} \tag{0.20}$$

Thus we need to know the set i

# Problem 4

After your yearly checkup, the doctor has bad news and good news. The bad news is that you tested positive for a serious disease and that the test is 99% accurate (i.e., the probability of testing positive when you do have the disease is 0.99, as is the probability of testing negative when you don't have the disease). The good news is that it is a rare disease, striking only 1 in 10,000 people of your age. What is the probability that you actually have the disease?

## Solution

Let define $X$ as the probability that you have the disease and $T$ the probability associated to the test accuracy. We need to calculate,

$$P(X = 1|T) \tag{0.21}$$

Using the Bayes rule

$$P(X = 1|T) = \frac{P(T|X)P(X)}{P(T)} \tag{0.22}$$

Assuming the test accurate only depends on the disease's probability, we can derive $P(T)$ as the sum of all possibilities, namely

$$P(T) = P(T|X = 1)P(X = 1)P(T|X = 0)P(X = 0) \tag{0.23}$$

Then, using 0.23 in 0.22

$$P(X = 1|T) = \frac{P(T|X = 1)P(X = 1)}{P(T|X = 1)P(X = 1)P(T|X = 0)P(X = 0)} \tag{0.24}$$

$$P(X = 1|T) = \frac{0.99 \times 0.0001}{0.99 \times 0.0001 + 0.01 \times 0.9999} \tag{0.25}$$

$$\cong 0.009804 \tag{0.26}$$

# Problem 5

Verify that the Bernoulli distribution

$$Bern(x|\mu) = \mu^x(1 - \mu)^{1-x} \tag{0.27}$$

satisfies the following properties

$$\sum_{x=0}^{1} p(x|\mu) = 1 \tag{0.28}$$

$$E[x] = \mu \tag{0.29}$$

$$var[x] = \mu(1 - \mu) \tag{0.30}$$

## Solution

From 0.27 we have

$$\sum_{x \in \{0,1\}} p(x|\mu) = p(x = 0|\mu) + p(x = 1|\mu) \tag{0.31}$$

$$= (1 - \mu) + \mu = 1 \tag{0.32}$$

$$\sum_{x \in \{0,1\}} xp(x|\mu) = 0 \times p(x = 0|\mu) + 1 \times p(x = 1|\mu) = \mu \tag{0.33}$$

$$\sum_{x \in \{0,1\}} (x - \mu)^2 p(x|\mu) = \mu^2 p(x = 0|\mu) + (1 - \mu)^2 p(x = 1|\mu) \tag{0.34}$$

$$= \mu^2(1 - \mu) + (1 - \mu)^2 \mu = \mu(1 - \mu) \tag{0.35}$$

# Problem 6

Consider the generalization of the squared loss function for a single target variable $t$

$$E[L] = \int \int \{y(x) - t\}^2 p(x,t) dx dt \tag{0.36}$$

to the case of multiple target variables described by the vector **t** given by

$$E[L(\overrightarrow{t}, y(x))] = \int \int ||y(x) - \overrightarrow{t}||^2 p(x, \overrightarrow{t}) dx d\overrightarrow{t} \tag{0.37}$$

Using the calculus of variations, show that the function $y(x)$ for which this expected loss is minimized is given by $y(x) = E_t[\overrightarrow{t}|x]$. Show that this result reduces to

$$y(x) = \frac{\int t p(x,t) dt}{p(x)} = \int t p(t|x) dt = E_t[t|x] \tag{0.38}$$

for the case of a single target variable t

## Solution

Our goal is to choose $y(x)$ so as to minimize $E[L]$. We can do it using,

$$\frac{\partial E[L]}{\partial y(x)} = \int 2(y(x) - \overrightarrow{t}) p(\overrightarrow{t}, x) d\overrightarrow{t} = 0 \tag{0.39}$$

solving $y(x)$,

$$\int 2(y(x) - \overrightarrow{t}) p(\overrightarrow{t}, x) d\overrightarrow{t} = 0 \tag{0.40}$$

$$\int 2y(x) p(\overrightarrow{t}, x) - 2\overrightarrow{t} p(\overrightarrow{t}, x) d\overrightarrow{t} = 0 \tag{0.41}$$

$$2y(x) \int p(\overrightarrow{t}, x) d\overrightarrow{t} - 2 \int \overrightarrow{t} p(\overrightarrow{t}, x) d\overrightarrow{t} = 0 \tag{0.42}$$

$$y(x) = \frac{\int \overrightarrow{t} p(\overrightarrow{t}, x) d\overrightarrow{t}}{\int p(\overrightarrow{t}, x) d\overrightarrow{t}} \tag{0.43}$$

by definition,

$$P(x,y) = P(y|x)p(x) \qquad (product\ rule) \tag{0.44}$$

$$P(x) = \sum_y P(x,y) \qquad (sum\ rule) \tag{0.45}$$

using 0.44 and 0.45 on 0.43

$$y(x) = \frac{\int \overrightarrow{t} p(\overrightarrow{t}|x) p(x) d\overrightarrow{t}}{p(x)} \tag{0.46}$$

$$y(x) = \frac{p(x) \int \overrightarrow{t} p(\overrightarrow{t}|x) d\overrightarrow{t}}{p(x)} \tag{0.47}$$

$$y(x) = \int \overrightarrow{t} p(\overrightarrow{t}|x) d\overrightarrow{t} \tag{0.48}$$

0.48 represents the conditional average of $\overrightarrow{t}$ conditioned on $x$