HDDA Tutorial: Distance : Solutions

Department of Econometrics and Business Statistics, Monash University

Tutorial 3

Work in groups of two people:

1. Consider the *age* (in years) and *height* (in cm) of both you and the other person (you are allowed to lie about these). Compute the Euclidean distance between you and the other person for these two variables.

Euclidean distance between National Basketball Association (NBA) most valuable player Giannis Antentokounmpo (height 211cm, age 24 years) and footballer Leo Messi (height 170 cm, age 32) is 41.77.

2. Repeat question 1 but use the Manhattan distance.

Manhattan distance between National Basketball Association (NBA) most valuable player Giannis Antentokounmpo and footballer Leo Messi is 41+8=49.

3. Repeat question 2 but measure height in metres.

The Manhattan distance is now 0.41+8=8.41. The units of measurement influence distance. This is why data in different units are normally standardised before computing distances. If both variables are measured in the same units (e.g. height and length of hands both measured in cm) then this standardisation is not necessary.

Select from the following list the types of cuisines that you enjoy:

- Chinese food
- Indian food
- Italian food
- Japanese food
- Lebanese food
- Mexican food
- Thai food
- British food

4. Compute a Jaccard similarity between you and the other person with regards to your taste in food.

Consider someone who only likes Chinese food and someone who likes Chinese Thai, Italian and Japanese. In common they enjoy one cuisine (Chinese). Between them they enjoy four cuisines (Chinese, Thai, Italian and Japanese). The Jaccard similarity is therefore 1/4 or 0.25.

5. Compute a Jaccard distance between you and the other person with regards to your taste in food.

Jaccard distance is 1 minus Jaccard similarity. For the previous example Jaccard distance is 0.75.

6. How would you define a distance between you and the other person that takes into account height, age and food preference.

One idea may be to add the Manhattan (or Euclidean) Distance and Jaccard similarity together. This will also satisfy the axioms of a distance (non-negativity, identity of indiscernables, symmetry, triangle inequality). As an advanced exercise keen student can try to prove that if two distances respects the four axioms their sum will also respect the four axioms.

7. Load in the Beer Dataset. Using numerical variables only, find the Euclidean distance between Pabst Extra Light and Augsberger. Do NOT use the `dist` function and for now, do NOT standardise the data.

```
library(tidyverse)
Beer<-readRDS('Beer.rds')
Beer%>%
  mutate(beer=trimws(beer))%>% #The beers have many trailing spaces, trimws removes them
  filter(beer == c('Pabst Extra Light'))%>%
  select(cost,calories,alcohol)->PabstEL

PabstEL
```

```
## # A tibble: 1 x 3
##    cost calories alcohol
##   <dbl>    <dbl>   <dbl>
## 1  0.38       68     2.3
```

```
Beer%>%
  mutate(beer=trimws(beer))%>%
  filter(beer == c('Augsberger'))%>%
  select(cost,calories,alcohol)->Augs
Augs
```

```
## # A tibble: 1 x 3
##    cost calories alcohol
##   <dbl>    <dbl>   <dbl>
## 1   0.4      175     5.5
```

To calculate Euclidean distance we need to find the difference between the beers for each variable. We then square these. Doing this manually.

```
dif<-PabstEL-Augs
dif
```

```
##    cost calories alcohol
## 1 -0.02     -107    -3.2
```

```
dif2<-dif^2
dif2
```

```
##    cost calories alcohol
## 1 4e-04    11449   10.24
```

Notice that the difference in calories completely dominates the calculation. To complete the answer we sum these squares and take the square root

```
sqrt(sum(dif2))
```

```
## [1] 107.0478
```

8. Repeat the previous question but this time standardise the data.

```
Beer%>%
  select(cost,calories,alcohol)%>%
  summarise_all(mean)->means
print(means)
```

```
## # A tibble: 1 x 3
##    cost calories alcohol
##   <dbl>    <dbl>   <dbl>
## 1 0.506     140.    4.58
```

```
Beer%>%
  select(cost,calories,alcohol)%>%
  summarise_all(sd)->sds
print(sds)
```

```
## # A tibble: 1 x 3
##    cost calories alcohol
##   <dbl>    <dbl>   <dbl>
## 1 0.187     24.4   0.603
```

The standard deviation is much larger for calories. Standardise the variables by subtracting the mean and dividing by the standard deviation.

```
PabstEL_std<-(PabstEL-means)/sds

Augs_std<-(Augs-means)/sds
```

To calculate Euclidean distance we need to find the difference between the beers for each variable. We then square these. Doing this manually.

```
dif<-PabstEL_std-Augs_std
print(dif)
```

```
##         cost  calories   alcohol
## 1 -0.1067674 -4.376829 -5.307012
```

```
dif2<-dif^2
print(dif2)
```

```
##         cost calories  alcohol
## 1 0.01139929 19.15663 28.16438
```

The difference in calories is still large but does not dominate the calculation as previously. In fact the difference in alcohol is larger after standardisation. To complete the answer we sum these squares and take the square root

```
sqrt(sum(dif2))
```

```
## [1] 6.879855
```