# 1 Assignment 4

## 1.1 Selection and building nominal features

In our dataset we have several binary features, such as weekdays (`weekday_is_monday`, `weekday_is_tuesday` and so on) and belonging to one of the channels (`data_channel_is_lifestyle`, `data_channel_is_entertainment` and so on). Therefore we built two nominal features:

- channel: integer values ranging between 1 and 6 ('Lifestyle', 'Entertainment', 'Business', 'Social Media', 'Tech', 'World')
- weekday: integer values ranging between 1 and 7

To obtain the third nominal feature we divide into four parts the feature `timedelta`: days between the article publication and the dataset acquisition.

```
timegroup <- cut(data$timedelta, breaks = 4)
```

And we break range of values of `timedelta` into intervals: $(7.28, 189]$, $(189, 370]$, $(370, 550]$, $(550, 732]$

## 1.2 Contingency tables over features

Conditional frequency tables over introdused nominal fetures are obtained with R-function `table` as showned below and results are presented at Tables 1 – ??.

```
table(data$channel, data$timegroup)
table(data$channel, data$weekday)
```

Table 1: Conditional frequency table over `channel` and `timegroup`

| | (7.28,189] | (189,370] | (370,550] | (550,732] |
|---|---|---|---|---|
| 0 | 412 | 327 | 361 | 381 |
| 1 | 120 | 93 | 130 | 183 |
| 2 | 573 | 473 | 341 | 348 |
| 3 | 392 | 408 | 410 | 446 |
| 4 | 81 | 154 | 163 | 192 |
| 5 | 401 | 469 | 491 | 492 |
| 6 | 901 | 564 | 373 | 321 |

Table 2: Conditional frequency table over `channel` and `weekday`

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 0 | 224 | 256 | 253 | 257 | 238 | 112 | 141 |
| 1 | 80 | 95 | 92 | 85 | 69 | 46 | 59 |
| 2 | 317 | 332 | 311 | 287 | 241 | 95 | 152 |
| 3 | 277 | 293 | 371 | 319 | 245 | 57 | 94 |
| 4 | 88 | 116 | 105 | 115 | 88 | 44 | 34 |
| 5 | 316 | 372 | 358 | 344 | 244 | 125 | 94 |
| 6 | 360 | 379 | 399 | 392 | 342 | 136 | 151 |

Quetelet relative index tables over our nominal fetures we obtain with the function:

```
getQueteletIndex <- function(v1, v2)
  size <- length(v1)
  cont.table  <- table(v1, v2)
  row.sums  <- rowSums(cont.table)
  col.sums  <- colSums(cont.table)
  norm.cont.table  <-  cont.table / size
  norm.row.sums  <- row.sums / size
  norm.col.sums  <- col.sums / size
  list(Quetelet = norm.cont.table / (norm.row.sums %*% t(norm.col.sums)) - 1,
       PearsonIndexMatrix = (-norm.row.sums %*% t(norm.col.sums) + norm.cont.table) /
          sqrt(norm.row.sums %*% t(norm.col.sums)))
```

Table 3: Quetelet relative index table over `channel` and `timegroup`

| | (7.28,189] | (189,370] | (370,550] | (550,732] |
|---|---|---|---|---|
| 0 | -3.41 | -11.26 | 7.43 | 8.87 |
| 1 | -20.79 | -28.94 | 8.92 | **47.23** |
| 2 | 14.67 | **9.57** | -13.38 | -15.12 |
| 3 | -17.81 | -0.97 | 9.12 | 13.98 |
| 4 | -52.33 | 4.91 | **21.76** | 37.72 |
| 5 | -24.86 | 1.73 | 16.78 | 12.36 |
| 6 | **44.90** | 5.00 | -23.86 | -37.08 |

Table 4: Quetelet relative index table over `channel` and `weekday`

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|
| 0 | -9.00 | -6.21 | -9.57 | -3.54 | 9.54 | 22.97 | 31.32 |
| 1 | -8.49 | -2.00 | -7.41 | -10.17 | -10.58 | **42.20** | **54.71** |
| 2 | **9.93** | 3.83 | -5.11 | -8.05 | -5.31 | -10.97 | 20.84 |
| 3 | 0.64 | -4.00 | **18.60** | 7.08 | 0.85 | -44.03 | -21.71 |
| 4 | -10.26 | 6.68 | -5.79 | **8.35** | 1.67 | 21.26 | -20.51 |
| 5 | 2.61 | **8.93** | 2.28 | 3.19 | -10.24 | 9.69 | -30.03 |
| 6 | 0.33 | -4.75 | -2.17 | 0.93 | **7.98** | 2.43 | -3.53 |

The results (in percent) are presented at Tables 3, 6.

As we can see from the Table 3, `timegroup` is dependent with `channel` in some values. For example, we observe rather big Quetelet relative index between Entartanment channel and 4th time-group. [1] In addition, we can't reject a dependence between World channel and 1st time-group. It can be caused by not random sampling or by some extra-ordinary events with great response in the world.

Table 6 provides us less surprising and more predictable results: all channels are almost independent with weekdays inspite of Entartanment channel and Weekend. This result is easy to understand: users visit Mashable at weekends to amuse themselves. [2]

## 1.3 $\chi^2$–summary Quetelet index

| | (7.28,189] | (189,370] | (370,550] | (550,732] | Sum |
|---|---|---|---|---|---|
| 0 | -0.007 (0.00005) | -0.02161 (0.0004) | 0.01362 (0.0002) | 0.01659 (0.0003) | (0.00098) |
| 1 | -0.02558 (0.0006) | -0.03310 (0.001) | 0.00975 (0.0001) | 0.05266 (0.003) | (0.0046) |
| 2 | 0.03280 (0.001) | 0.01989 (0.0004) | -0.02655 (0.0007) | -0.03061 (0.0009) | (0.003) |
| 3 | -0.03889 (0.0015) | -0.00198 (0) | 0.01767 (0.0003) | 0.02765 (0.0008) | (0.0026) |
| 4 | -0.06821 (0.0047) | 0.00595 (0.00004) | 0.02518 (0.0006) | 0.04453 (0.002) | (0.0073) |
| 5 | -0.05743 (0.0033) | 0.00371 (0.00001) | 0.03441 (0.0012) | 0.02587 (0.0006) | (0.005) |
| 6 | 0.11197 (0.013) | 0.01158 (0.00013) | -0.05281 (0.00279) | -0.08375 (0.007) | (0.0225) |
| Sum | (0.0237) | (0.002) | (0.006) | (0.0144) | (0.04624) |

Table 5: $\chi^2$–summary Quetelet index over `channel` and `timegroup`

| | 1 | 2 | 3 | 4 | 5 | 6 | 7 | Sum |
|---|---|---|---|---|---|---|---|---|
| 0 | -0.014 (0) | -0.01 (0) | -0.016 (0) | -0.006 (0) | 0.014 (0) | 0.022 (0) | 0.032 (0.001) | (0.002) |
| 1 | -0.008 (0) | -0.002 (0) | -0.007 (0) | -0.01 (0) | -0.009 (0) | 0.024 (0.001) | 0.034 (0.001) | (0.002) |
| 2 | 0.017 (0) | 0.007 (0) | -0.009 (0) | -0.014 (0) | -0.008 (0) | -0.011 (0) | 0.023 (0.001) | (0.001) |
| 3 | 0.001 (0) | -0.007 (0) | 0.033 (0.001) | 0.012 (0) | 0.001 (0) | -0.044 (0.002) | -0.024 (0.001) | (0.004) |
| 4 | -0.01 (0) | 0.007 (0) | -0.006 (0) | 0.009 (0) | 0.002 (0) | 0.013 (0) | -0.013 (0) | (0.001) |
| 5 | 0.005 (0) | 0.017 (0) | 0.004 (0) | 0.006 (0) | -0.017 (0) | 0.01 (0) | -0.035 (0.001) | (0.002) |
| 6 | 0.001 (0) | -0.009 (0) | -0.004 (0) | 0.002 (0) | 0.014 (0) | 0.003 (0) | -0.004 (0) | (0) |
| Sum | (0.001) | (0.001) | (0.002) | (0.001) | (0.001) | (0.003) | (0.005) | (0.0124) |

Table 6: $\chi^2$–summary Quetelet index over `channel` and `weekday`

[1] WHY????

[2] To fix.

### 1.4 Sufficient sample size for significant result

Supposing the probabilities $p_{i+}$, $p_{+j}$, $p_{ij}$ are constant and sample size $n$ is varying, we can get $\chi^2$-statistics from

$$nX^2 = \sum_{k=1}^{K} \sum_{l=1}^{L} \frac{(p_{kl} - p_{k+}p_{+l})^2}{p_{k+}p_{+l}} \overset{n\to\infty}{\Rightarrow} \chi^2((K-1)(L-1))$$

We know $X^2$ and $L$ for pairs `channel−timegroup` and `channel−weekday`, so, we can get sufficient $K$ for significant results.