# **Week 5 –** **Predictive Data Analysis**

# **Exercise 01: Weighted Moving Everage**

A firm has the following order history over the last 6 months.

January 120

February 95

March 100

April 75

May 100

June 50

* What would be a 3-month moving average forecast for July?
* July = (April + May + June)/3 = (75 + 100 + 50) /3 = 75
* What would be a 3-month weighted moving average forecast for July, using weights of 40% for the most recent month, 30% for the month preceding the most recent month, and 30% for the month preceding that one?
* July = 0.4\*June + 0.3\*May + 0.3\*April = 0.4\*50+0.3\*100+0.3\*75= 72.5

# **Exercise 02: Exponential smoothing**

The mean price for rubber during 10 years is shown in the Table below:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Year | 2001 | 2002 | 2003 | 2004 | 2005 | 2006 | 2007 | 2008 | 2009 | 2010 |
| Price | 82 | 80 | 76 | 73 | 72 | 73 | 72 | 73 | 77 | 74 |

* Give a forecast for the price of schnaps in 2010 based on simple exponential smoothing.
* Give α=0.2
* F(2010) = α\*A(2009) + α\*(1-α)\*A(2008) + α\*(1-α)^2\*A(2007) + … + α\*(1-α)^9\*A(2001) = 0.2\*77 + 0.2\*0.8\*73 + … + 0.2\*0.8^9\*82 = 15.4 + 11.68 + 9.216 + 7.4752 + 5.98016 + 4.718592 + 3.985 + 3.355 + 2.751 = 64.72
* Compute the Mean Absolute Deviation (MAD).
* F(2001) = 0
* F(2002) = 16.4
* F(2003) = 29.12
* F(2004) = 38.496
* F(2005) = 45.397
* F(2006) = 50.71
* F(2007) = 55.17
* F(2008) = 56.539
* F(2009) = 61.431
* F(2010) = 64.72
* MAD = ((A(2010) – F(2010)) + (A(2009) – F(2009)) + … + (At(2001) – F(2001)) / 10 = (9.28+15.569+…82)/10=33.4017
* Do you have any viewpoint about the choice of model in this case?
* When we choose α = 0.2, the different between the forecasting values and the actual values is pretty big at the beginning and smaller at the end. When we observe the trend of the actual values and compare it with the trend of calculated forecasting values, it’s not appropriate.

# **Exercise 03: Items-based Recommender**

Three computers, C1, C2, and C3, have the numerical features listed below:

|  |  |  |  |
| --- | --- | --- | --- |
| **Feature** | **Processor Speed** | **Disk Size** | **Memory Size** |
| C1 | 3.06 | 500 | 6 |
| C2 | 2.68 | 320 | 4 |
| C3 | 2.92 | 640 | 6 |

We may imagine these values as defining a vector for each computer; for instance, C1’s vector is [3.06, 500, 6]. We can compute the cosine distance between any two of the vectors, but if we do not scale the components, then the disk size will dominate and make differences in the other components essentially invisible. Let us use 1 as the scale factor for processor speed, α for the disk size, and β for the main memory size. In terms of α and β, compute the cosines of the angles between the vectors for each pair of the three computers in two following senerios:

* What are the angles between the vectors if α = β = 1?
* cos(C1, C2) = (8.2008 + 160000 + 24) / ( ) = 1 🡪 the angel between C1 and C2 = 0
* cos(C2, C3) = (7.8256 + 204800 + 24) / ( \* ) = 1 🡪 the angel between C2 and C3 = 0
* cos(C1, C3) = (8.9352 + 320000 + 36) / ( \* ) = 1 🡪 the angel between C1 and C3 = 0
* What are the angles between the vectors if α = 0.01 and β = 0.5?
* cos(C1, C2) = (8.2008 + 16 + 6) / ( ) = 0.99 🡪 the angel between C1 and C2 = 8.1
* cos(C2, C3) = (7.8256 + 20.48 + 6) / ( \* ) = 0.97 🡪 the angel between C2 and C3 = 14
* cos(C1, C3) = (8.9352+32+9) / ( \* ) = 0.99 🡪 the angel between C1 and C3 = 8.1
* Do you have any viewpoint about the choice of α, β in this case?
* According to given figures, we can see that the disk size is larger than the memory size and processor speed. We should choose small α so it won’t dominate the others. Besides that, smaller β is not suitable because the memory size is already too small.