# **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?

(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?

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.
* Assumpt that: α = 0.2, F**2001** = 0
* We have: F**t =** Ft-1 + α(At-1 – Ft-1)

F2010 = 0.2\*77 + 0.2\*0.8\*73 + 0.2\*0.8\*0.8\*72 + 0.2\*0.8\*0.8\*0.8\*73 + 0.2\*0.8\*0.8\*0.8\*0.8\*72 + 0.2\*0.8\*0.8\*0.8\*0.8\*0.8\*73 + 0.2\*0.8\*0.8\*0.8\*0.8\*0.8\*0.8\*76 + 0.2\*0.8\*0.8\*0.8\*0.8\*0.8\*0.8\*0.8\*80 + 0.2\*0.8\*0.8\*0.8\*0.8\*0.8\*0.8\*0.8\*0.8\*82 = 15.4 + 11.86 + 9.216 + 7.4752 + 5.89824 + 4.784128 + 3.985 + 3.355 + 2.751 = 64.72

F2002 = 16.4

F2003 = 29.12

F2004 = 38.49

F2005 = 45.39

F2006 = 50.72

F2007 = 55.17

F2008 = 58.54

F2009 = 61.43

F2010 = 64.55

* Compute the Mean Absolute Deviation (MAD).

MAD = **Σ**|actual – forecast| / n = 33.4017

* Do you have any viewpoint about the choice of model in this case?

With α = 0.2, the different forecasting value and actual value is big at begining years and becomes smaller overtime. Comparing with the actual trend, the forecasting trend seems 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) = 160032.2/(500.02\*320.04) ≈ 1

cos(C2,C3) = 204831.8/(320.04\*640.03) ≈ 1

cos(C1,C3) = 320044.9/(500.02\*640.03) ≈ 1

The angles between the vectors is 0.

* What are the angles between the vectors if α = 0.01 and β = 0.5?

cos(C1, C2) = 30.2008/(6.58\*4.63) = 0.99

-> angle between c1 and c2: 0.14 rad

cos(C2,C3) = 34.3056/(4.63\*7.65) = 0.96

-> angle between c2 and c3: 0.28 rad

cos(C1,C3) = 49.9352/(6.58\*7.65) = 0.99

-> angle between c1 and c3: 0.14 rad

* Do you have any viewpoint about the choice of α, β in this case?

Because the disk size value is very big comparing to processor speed and memory size, I think we should choose small α so that the disk size cannot dominate the others. I suppose that a small β is not recommended because memory size values are currently not big.