Predictive Mining for Time-Series Data

Ali Ridho Barakbah, Entin Martiana

Knowledge Engineering Research Group
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya

Predictive Mining

Non-time series data with categorical/ordinal value prediction

Classification algorithms can be used

Non-time series data with numerical value prediction

Some classification algorithms can be used (NN)

Time series data

- Statistical approaches for forecasting
- Some machine learning algorithms (RNN, LSTM)

Discussion today: Methods for Time-Series Data Prediction

- Simple Moving Average
- Weighted Moving Average
- Linear Regression
- Exponential Smoothing

Simple Moving Average

- Used for smoothing
- Attempts to find a local mean
- This can be done simply by taking the average of the points around the time of interest
- For example, if we are interested in a window of width k, we simply take a series of data and compute their average

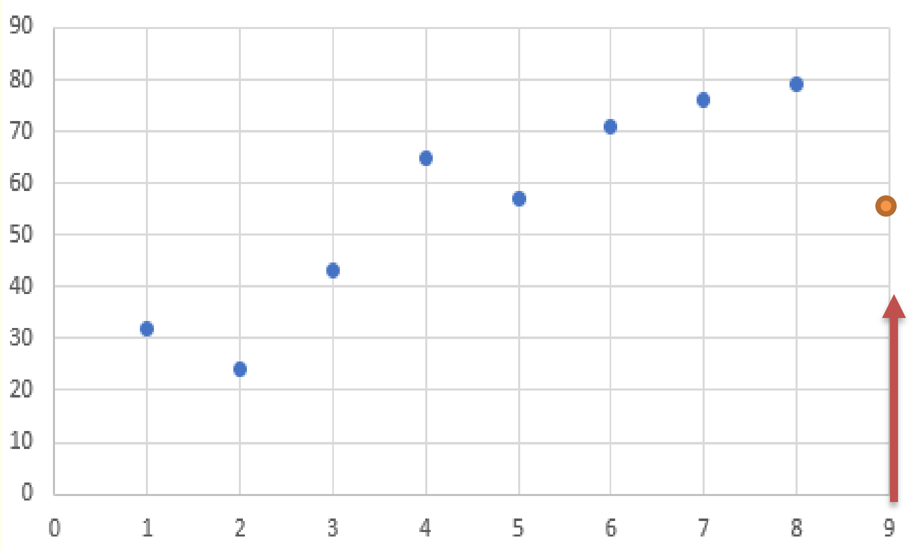
$$y_{t+1} = \frac{1}{k}\sum_{i=1}^{k} y_i$$

X	Y
1	32
2	24
3	43
4	65
5	57
6	71
7	76
8	79
Mean	55.875

k = 8

x = 9

y = 55.875



Weighted Moving Average

- To give different orientation in a window of width k
- Set a series of weights for each window of the data
- Commonly, it assigns greater weighting to recent data points and less weighting on past data points
- The weighted moving average is calculated by multiplying each observation in the data set by a predetermined weighting factor.

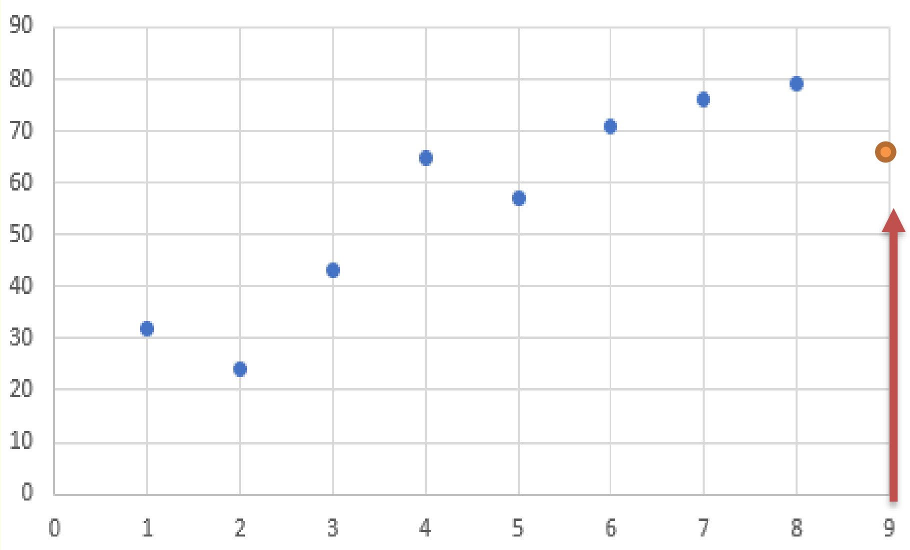
$$y_{t+1} = \sum_{i=1}^{k} w_i y_i$$

X	w	Y	w * Y
1	1 / 36	32	0.89
2	2 / 36	24	1.33
3	3 / 36	43	3.58
4	4 / 36	65	7.22
5	5 / 36	57	7.92
6	6 / 36	71	11.83
7	7 / 36	76	14.78
8	8 / 36	79	17.56
Sum 36			65.11

$k = 8$

x = 9

y = 65.11



Linear Regression

- Regression is a measuring tool used to determine whether there is a correlation between variables
- Regression analysis is more accurate in correlation analysis because the rate of change of a variable against other variables can be determined. So in regression, forecasting or estimating the value of the dependent variable on the independent variable is more accurate
- Linear regression is a regression where the independent variable (variable X) has the highest rank of one. For simple regression, i.e. linear regression which only involves 2 variables (variables X and Y)

Linear Regression from Y to X

$$Y = a + b * X$$

where:

Y = dependent variable

X = independent variable

a = intercept

b = slope (regression coefficient)

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

$$b = \frac{(n)(\Sigma XY) - (\Sigma X)(\Sigma Y)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

	X	Y	X^2	XY
	1	32	1	32
	2	24	4	48
	3	43	9	129
	4	65	16	260
	5	57	25	285
	6	71	36	426
	7	76	49	532
	8	79	64	632
Sum	36	447	204	2344

$n = 8$

$$a = \frac{(\Sigma Y)(\Sigma X^2) - (\Sigma X)(\Sigma XY)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

$$= \frac{(447 * 204) - (36 * 2344)}{(8 * 204) - (36 * 36)}$$

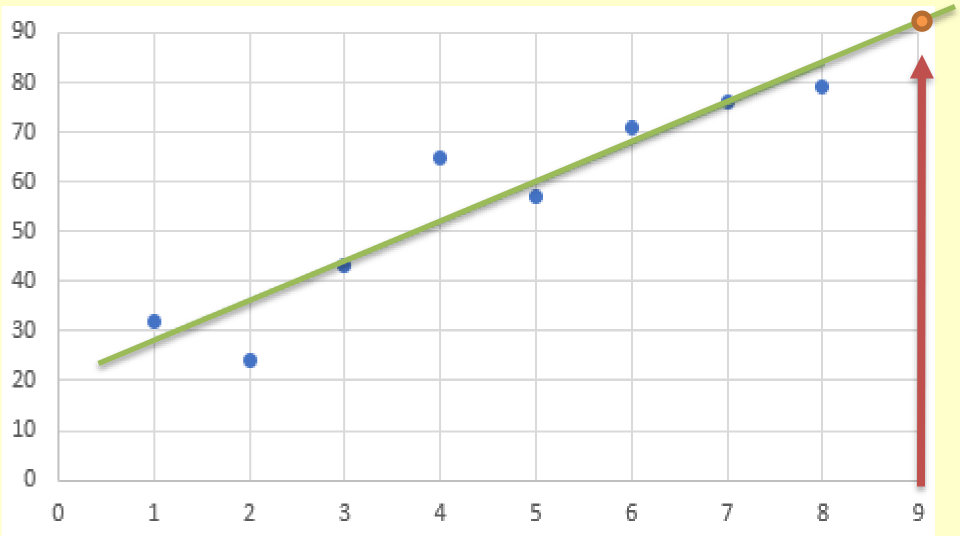
= 20.25

$$b = \frac{(n)(\Sigma XY) - (\Sigma X)(\Sigma Y)}{(n)(\Sigma X^2) - (\Sigma X)^2}$$

$$= \frac{(8 * 2344) - (36 * 447)}{(8 * 204) - (36 * 36)}$$

= 7.9167

$$Y = a + b\,X$$
$$= 20.25 + 7.9167 * X$$



$x = 9$

$$y = 20.25 + 7.9167 * 9$$
$$= 20.25 + 71.2503$$
$$= 91.5003$$

Exponential Smoothing

- Forecasting is based on past forecasting errors that are used for subsequent forecasting corrections
- Calculated based on the results of forecasting + previous forecasting errors
- Do iteratively over time k

$$F_{t+1} = \alpha\, D_t + (1-\alpha)\, F_t$$

where:

F_{t+1} = Prediction value in period t+1

F_t = Prediction value in period t

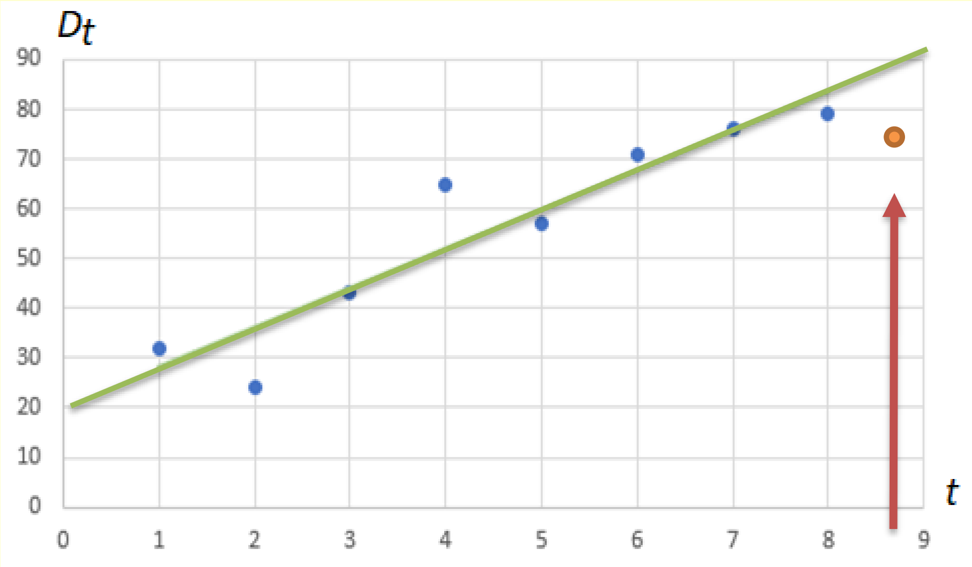
D_t = Actual value in period t

α = weight

α = 70%

$t = 9 \rightarrow D_t = 77.27$

t	D_t	F_t	αD_t	$(1-\alpha) F_t$	F_{t+1}
1	32	-	-	-	32
2	24	32	16.80	9.60	26.40
3	43	26.40	30.10	7.92	38.02
4	65	38.02	45.50	11.41	56.91
5	57	56.91	39.90	17.07	56.97
6	71	56.97	49.70	17.09	66.79
7	76	66.79	53.20	20.04	73.24
8	79	73.24	55.30	21.97	77.27

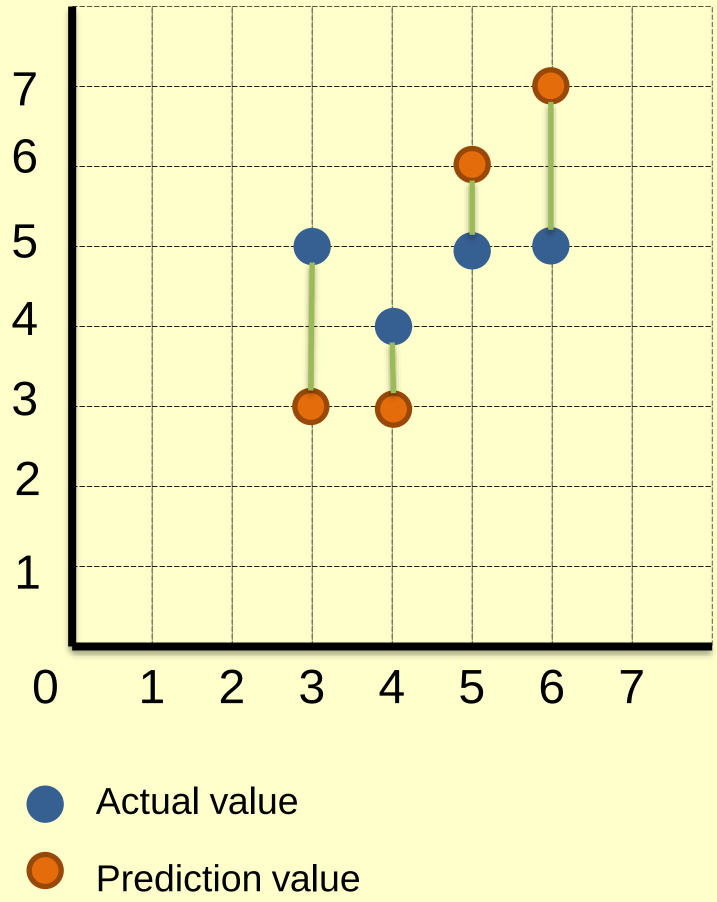


Prediction Evaluation

$$\text{Mean Absolute Error (MAE)} = \frac{\sum_{t=1}^{N} \left| d_t - d_t' \right|}{N}$$

$$\text{Mean Squared Error (MSE)} = \frac{\sum_{t=1}^{N} \left(d_t - d_t' \right)^2}{N}$$

$$\text{Mean Absolute Percent Error (MAPE)} = \frac{100}{N} \sum_{t=1}^{N} \left[\left| \frac{d_t - d_t'}{d_t} \right| \right]$$



$$MAE = \frac{2 + 1 + 1 + 2}{4} = 1.5$$

$$MSE = \frac{4 + 1 + 1 + 4}{4} = 2.5$$

$$MAPE = \frac{100}{4}\left[\frac{2}{5} + \frac{1}{4} + \frac{1}{5} + \frac{2}{5}\right]$$

$$= 31.25$$