Data Mining

Pertemuan 8
Regression

Definisi

- Serangkaian proses statistik untuk memperkirakan hubungan antara variabel dependen dan satu atau lebih variabel independen.

Dependent dan Independent

- Dependent variable = outcome variable
  - Variabel yang bergantung dari variabel lain.
- Independent variables = predictors, covariates, atau features
  - Variabel yang tidak bergantung dari variabel lain tetapi ada kemungkinan berpengaruh pada variabel lain

Prediction & Forecasting

- **Prediction** adalah proses menebak suatu nilai berdasarkan nilai lain.
- **Forecasting** atau peramalan adalah proses membuat prediksi masa depan berdasarkan data masa lalu dan sekarang dan paling umum dengan analisis tren.

Notasi Umum Regression

$$Y_i = f(X_i, \beta) + e_i$$

- beta : unknown parameters
- Y : dependent parameters
- X : independent parameters
- e : error

Distance Function

Distance functions

Euclidean

$$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan

$$\sum_{i=1}^{k}|x_i - y_i|$$

Minkowski

$$\left(\sum_{i=1}^{k}(|x_i - y_i|)^q\right)^{1/q}$$

q : konstanta (umumnya 1 atau 2)

Contoh menggunakan KNN

Age	Loan	House Price Index	Distance	
25	$40,000	135	102000	
35	$60,000	256	82000	
45	$80,000	231	62000	
20	$20,000	267	122000	
35	$120,000	139	22000	2
52	$18,000	150	124000	
23	$95,000	127	47000	
40	$62,000	216	80000	
60	$100,000	139	42000	3
48	$220,000	250	78000	
33	$150,000	264	8000	1
48	**$142,000**	**?**		

$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01

K = 3 (3 distance terdekat)

HPI = (264+139+139)/3 = 180.7

Menggunakan Normalisasi

Age	Loan	House Price Index	Distance
0.125	0.11	135	0.7652
0.375	0.21	256	0.5200
0.625	0.31	231	0.3160
0	0.01	267	0.9245
0.375	0.50	139	0.3428
0.8	0.00	150	0.6220
0.075	0.38	127	0.6669
0.5	0.22	216	0.4437
1	0.41	139	0.3650
0.7	1.00	250	0.3861
0.325	0.65	264	0.3771
0.7	**0.61**	**?**	

$$X_s = \frac{X - Min}{Max - Min}$$

Hanya independent variable saja yang di normalisasi

Mean Square Error

$$\mathrm{MSE} = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2.$$

Average dari hasil kuadrat selisih antara dependent variabel yang sebenarnya dan diprediksi

Mean Absolute Error

$$\mathrm{MAE} = \frac{\sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|}{n}$$

Average dari hasil absolut selisih antara dependent variabel yang sebenarnya dan diprediksi

Coefficient of determination

$$SS_{\text{tot}} = \sum_i (Y_i - \bar{Y})^2$$ Total sum of squares

$$SS_{\text{res}} = \sum_i (Y_i - \hat{Y}_i)^2$$ Residual sum of squares

R squared

$$R^2 = 1 - \frac{SS_{\text{res}}}{SS_{\text{tot}}}$$

Proporsi varians dalam variabel dependen yang dapat diprediksi dari variabel independen