Nama : Akmal Zuhdy Prasetya
NIM : H071191035
Kelas : Machine Learning

Resume Pertemuan 7

A. Linear Regression

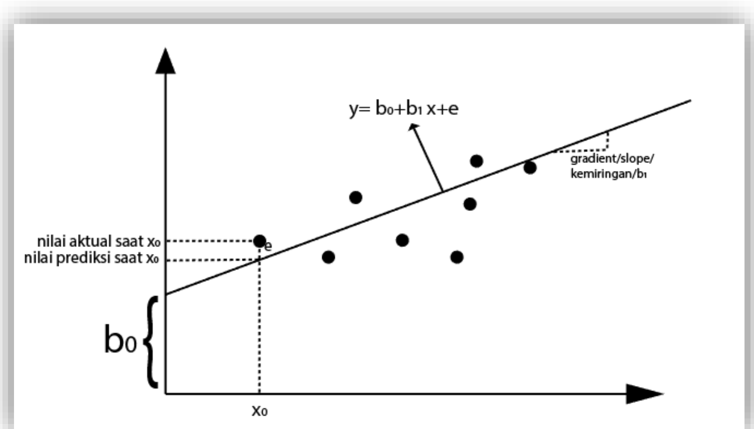
Regresi adalah teknik statistik untuk pemodelan dan investigasi hubungan dua atau lebih variabel. Regresi adalah membangun model untuk memprediksi nilai dari data masukan yang diberikan, dimana prediksi berbeda dengan klasifikasi. Yang sering dipakai dan paling sederhana adalah regresi linear sederhana.

Metode utama untuk melakukan prediksi, yaitu membangun model regresi dengan mencari hubungan antara satu atau lebih variable independent atua predictor (x) dengan variable dependen atau respon (y).

Model umum regresi linear dapat dituliskan seperti persamaan dibawah ini:

$$y = b_0 + b_1 x + e$$

Persamaan diatas dapat diilustrasikan sebagai berikut:



Pada gambar di bawah, tepatnya di sebelah kiri, slope/gradien b_1 bernilai positif, sedangkan disebelah kanan itu bernilai negatif. Prediksi nilai dengan pendekatan regresi linier sederhana, di dapatkan dengan:

$$\hat{y} = b_0 + b_1 x$$

Dalam analisis regresi linier sederhana, kita ingin menemukan nilai optimal dari b_0 dan b_1. Untuk mencari nilai koefisien b_0 dan b_1 tersebut, kita pergunakan pendekatan Least Squares Estimation (LSE). Pendekatan ini berusaha meminimalkan fungsi objektif $j(b)$, yakni jumlah error kuadrat dari semua data yang digunakan untuk menemukan estimasi garis regresi. Fungsi objektifnya sebagai berikut.

$$J(b) = \sum_{i=1}^{m} (\hat{y} - y_i)^2$$

Diberikan data training $(x_1, y_1), (x_2, y_2), (x_3, y_3), \dots (x_m, y_m)$, tujuannya dari regregi linear sederhana adalah mencari parameter b_0 dan b_1 yang meminimalkan nilai error yang telah kita definisikan dengan fungsi objektif diatas. Untuk menemukan parameter-parameter tersebut, kita dapat melakukannya dengan menurunkan fungsi objektif dengan parameter terkait.

$$\frac{\partial J(b)}{\partial b_0} = 0 \qquad \frac{\partial J(b)}{\partial b_0} = 0$$

Dengan menyelesaikan penurunan pertama diatas, kita bisa mendapatkan nilai optimal dari b_0 dan b_1 masing-masing sebagai berikut.

$$b_1 = \frac{\sum_{i=0}^{m} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=0}^{m} (x_i - \bar{x})^2} \text{ atau } b_1 = \frac{m \sum x_i y_i - (\sum x_i)(\sum y_i)}{m \sum x_i 2 - (\sum x_i) 2}$$

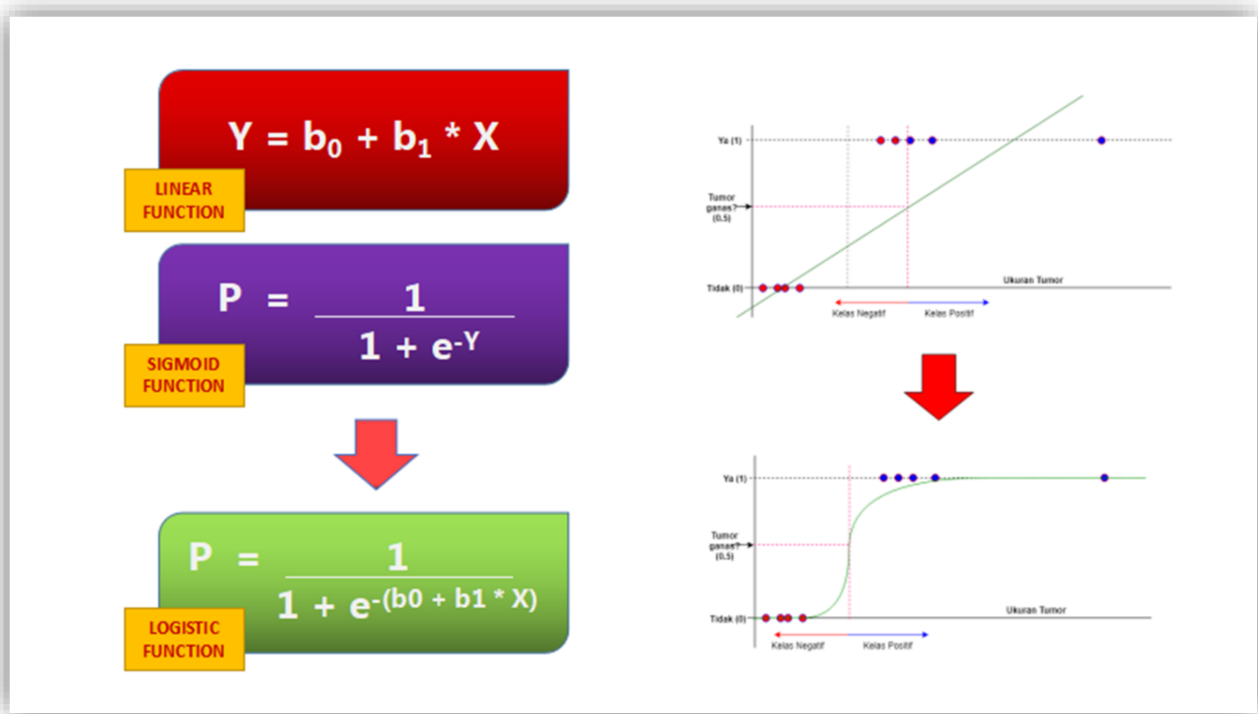
$$b_0 = \bar{y} - b_1 \bar{x} \text{ atau } b_0 = \frac{(\sum y_i - b_1 \sum x_i)}{n}$$

B. Logistic Regression

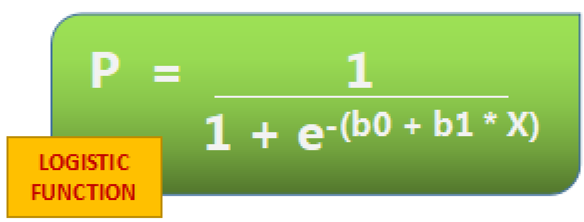
Logistic Regression adalah sebuah algoritma klasifikasi untuk mencari hubungan antara fitur (input) diskrit/kontinu dengan probabilitas hasil output diskrit tertentu.

Berikut merupakan tipe-tipe regresi logistik:

- Binary Logistic Regression. Adalah Logistic Regression yang hanya memiliki 2 output saja (mengklasifikasi kedalam 2 kelas berbeda). Contoh: Positif-Negatif, Obesitas-Tidak Obesitas.
- Multinomial Logistic Regression. Adalah Logistic Regression yang memiliki 2 output atau lebih (mengklasifikasi kedalam 2 kelas berbeda). Contohnya kelas Sentiment Analysis kalimat positif, negatif, dan netral.
- Ordinal Logistic Regression: Adalah Logistic Regression yang memiliki 2 output atau lebih dengan memperhatikan urutan. (mengklasifikasi kedalam 2 kelas berbeda dengan memperhatikan urutan). Contohnya adalah membagi kelas mahasiswa dalam range Index Prestasi Kumulatif 1.xx, 2.xx, 3.xx, dan 4.00.



Dalam regresi ini juga terdapat istilah Maximum Likelihood, yaitu cara untuk menentukan posisi Sigmoid yang menjadi model terbaik yang dapat dibentuk dari data-data yang tersedia. Bagaimana caranya? perhatikan persamaan fungsi Logistik berikut:



Dari fungsi diatas kita ketahui bahwa yang mempengaruhi posisi fungsi Sigmoid adalah persamaan b0+b1*X, artinya perubahan garis (merubah nilai koefisien b0 dan b1)akan mempengaruhi posisi dan nilai Likelihood dari Logistic Function.

Adapun langkah sederhana dalam mencari Maximum Likelihood adalah sebagai berikut:

- Tentukan suatu persamaan garis sembarang, ubah kedalam bentuk Sigmoid, dan hitung nilai Likelihood-nya
- Lakukan Rotasi (bisa disertai translasi juga) pada persamaan garis sebelumnya, kemudian ubah kembali kedalam bentuk Sigmoid, dan hitung nilai Likelihood garis baru.

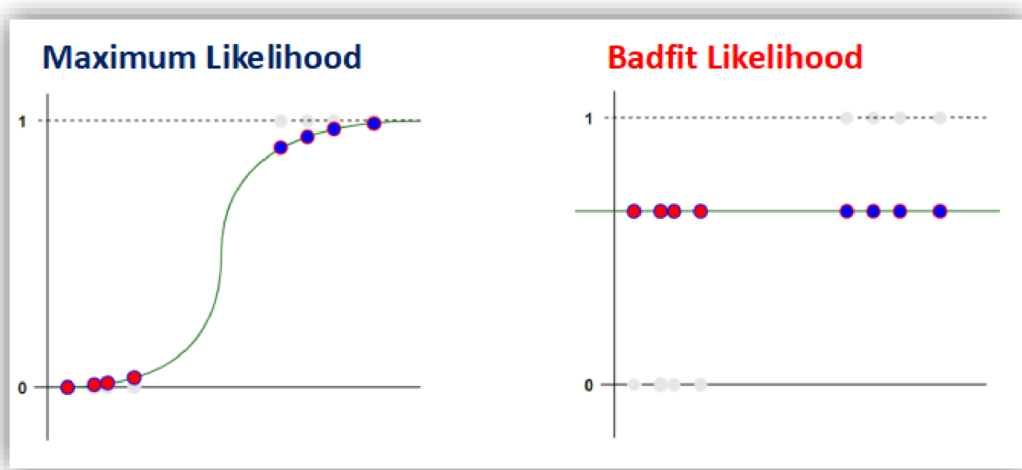
- Ulangi terus langkah kedua hingga mendapatkan nilai Likelihood tertinggi. Garis dengan Likelihood tertinggi berarti mengisyaratkan bahwa garis yang digunakan menentukan posisi Sigmoid yang merepresentasikan klasifikasi data dengan baik.

Nilai Likelihood dari setiap garis dapat dicari dengan formula berikut:

Likelihood = Log(P(datapositif 1))+Log(P(datapositif 2))+...+
Log(P(datapositif n-1))+Log((Pdatapositif n))+
Log(1-P(datanegatif 1))+Log(1-P(datanegatif 2))+...+
Log(1-P(datanegatif m-1))+Log(1-P(datanegatif m))

Perlu diingat juga bahwa pada formula diatas terdapat “1-P(datanegatif)” karena likelihood menghitung kemiripan peluang terhadap kelas positif (dalam konteks ini kelas 1 positif dan kelas 0 negatif), sehingga representasi positif dari “P(datanegatif)” adalah “1-P(datanegatif)”.

Istilah lain yang ada yaitu R-Squared (R^2), yaitu cara yang digunakan untuk mengetahui apakah Logistic Function dengan nilai Maximum Likelihood dapat merepresentasikan data dengan baik (baik jika R-squared = 1, Buruk jika R-squared = 0). Terdapat 2 parameter penting yang dibutuhkan dalam mencari nilai R-Squared, yaitu Maximum Likelihood dan Badfit Likelihood.



Hal yang perlu diingat adalah:

- Untuk parameter Badfit Likelihood, garis lurus didapat dengan persamaan:
- Y = banyak data kelas 1 atau banyak data keseluruhan
- sehingga kita dapat mencari nilai Badfit Likelihood dengan cara:
- Badfit Likelihood = $Log(Y) + Log(Y) + \ldots. + Log(1-Y) + Log(1-Y)$

- Dari Maximum Likelihood dan Badfit Likelihood dapat dibentuk formula R-Squared (R^2) sebagai berikut:

$$\text{R-Squared} = \frac{\text{Badfit Likelihood} - \text{Maximum Likelihood}}{\text{Badfit Likelihood}}$$

Contoh implementasi dapat dilihat pada link di bawah ini:

https://github.com/trueazp/Machine-Learning/blob/main/assignments/assignment06/Tugas06_ML_H071191035.ipynb