

Nama : Anindita Khusnul Oktavia
Program : Introduction to Python for Data Science

SUMMARY

REGRESSION MODEL

Regression

Regression analysis adalah salah satu bidang terpenting dalam statistik dan machine learning. Ada banyak metode regresi yang tersedia. Linear regression adalah salah satunya. model yang paling sederhana yang kemungkinan besar akan paling sering digunakan.

Regression mencari hubungan antar variabel. Regression adalah bentuk lain dari supervised learning. Perbedaan antara classification dan regression adalah bahwa regression **menghasilkan angka** daripada kelas. Oleh karena itu, regression berguna saat memprediksi masalah berbasis angka seperti harga pasar saham, suhu untuk hari tertentu, atau probabilitas suatu peristiwa.

Linear Regression

Linear regression mungkin salah satu teknik regresi yang paling penting dan banyak digunakan. Saat menerapkan linear regression dari beberapa variabel dependen y pada himpunan variabel independen $\mathbf{x} = (x_1, \dots, x_r)$, di mana r adalah jumlah prediktor, kita mengasumsikan linear relationship antara y dan \mathbf{x} : $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_r x_r + \epsilon$. Persamaan ini adalah **persamaan regresi**. $\beta_0, \beta_1, \dots, \beta_r$ adalah **koefisien regresi**, dan ϵ adalah **random error**.

Regression Performance

Variasi respon aktual $y_i, i = 1, \dots, n$, terjadi sebagian karena ketergantungan pada prediktor \mathbf{x}_i . Namun, ada juga varian tambahan yang melekat pada output.

Coefficient of determination, dilambangkan sebagai R^2 , akan memberi berapa banyak variasi dalam y yang dapat dijelaskan oleh ketergantungan pada \mathbf{x} menggunakan model regresi tertentu. Lebih besar R^2 menunjukkan kesesuaian yang lebih baik dan berarti bahwa model dapat menjelaskan variasi keluaran dengan masukan yang berbeda dengan lebih baik.

Simple Linear Regression

Simple atau single-variate linear regression adalah kasus regresi linier yang paling sederhana dengan variabel independen tunggal, $\mathbf{x} = x$.

Multiple Linear Regression

Multiple atau multivariate linear regression adalah kasus regresi linier dengan dua atau lebih independent variables.

Polynomial Regression

diaasumsikan polynomial dependence antara output dan inputs , dan menghasilkan, polynomial estimated regression function. Contoh paling sederhana dari polynomial regression memiliki satu independent variable, dan estimated regression function adalah polynomial of degree 2: $f(x) = b_0 + b_1x + b_2x^2$. Sekarang, ingatlah bahwa kita ingin menghitung b_0 , b_1 , dan b_2 , yang meminimalkan SSR.

Underfitting and Overfitting

Underfitting terjadi jika model tidak dapat menangkap dependensi di antara data secara akurat, biasanya sebagai konsekuensi dari kesederhanaannya sendiri. Seringkali menghasilkan R^2 rendah dengan data yang diketahui dan kemampuan generalisasi yang buruk saat diterapkan dengan data baru.

Overfitting terjadi saat model mempelajari dependensi di antara data dan fluktuasi acak. Dengan kata lain, model mempelajari data yang ada dengan terlalu baik. Model kompleks, yang memiliki banyak fitur atau istilah, sering kali cenderung overfitting. Ketika diterapkan pada data yang diketahui, model seperti itu biasanya menghasilkan R^2 yang tinggi. Namun, mereka sering tidak menggeneralisasi dengan baik dan memiliki lebih rendah secara signifikan R^2 saat digunakan dengan data baru.

Untuk implementasinya ada pada file .ipynb