Glossary

Team Algoritma

1/21/2020

Programming for Data Science

- ***R Markdown***

  Tipe file pada R yang dapat mengkombinasikan narasi dan code.

- ***Console***

  Tempat untuk menjalankan perintah/code.

- ***Chunk***

  Tempat untuk menempatkan dan menjalankan code pada `R Markdown`.

- ***Package***

  Kumpulan fungsi, sering disebut sebagai `library`.

- ***Implicit coercion***

  Konversi ke tipe data yang paling umum.

- ***Vector***

  Suatu objek di R yang dapat menyimpan lebih dari 1 elemen. `Vector` hanya dapat menyimpan tipe data yang sama. Jika terdapat lebih dari satu tipe data pada satu elemen, maka akan dilakukan `implicit coercion`.

- ***Character***

  Tipe data ini digunakan untuk menyimpan nilai dalam bentuk karakter, sering disebut sebagai `string`.

- ***Numeric***

  Tipe data ini digunakan untuk menyimpan nilai dalam bentuk angka (bilangan desimal).

- ***Integer***

  Tipe data ini digunakan untuk menyimpan nilai dalam bentuk angka (bilangan bulat).

- ***Complex***

  Tipe data ini digunakan untuk menyimpan gabungan nilai `numeric` dan imajiner. Contoh: `1+3i`.

- ***Logical***

  Tipe data ini digunakan untuk menyimpan nilai logika (`TRUE` dan `FALSE`).

- ***Matrix***

  Objek di R yang dapat menyimpan elemen dalam dua dimensi. `Matrix` hanya dapat menyimpan nilai dengan tipe data yang sama, kemudian disusun secara baris dan kolom.

- ***List***

  Objek di R yang dapat menyimpan elemen yang memiliki tipe data dan jumlah elemen yang berbeda.

- ***Data frame***

  Objek di R yang dapat menyimpan elemen yang memiliki tipe data yang berbeda namum memiliki jumlah elemen yang sama.

- ***Factor***

  Tipe data ini digunakan untuk menyimpan nilai yang berulang (nilai dengan tipe data kategorik). Contoh: `Gender` terdiri dari `Laki-laki` dan `Perempuan`.

Practical Statistics

- ***Population***

  Keseluruhan data.

- ***Sample***

  Bagian kecil/subset dari `population`.

- ***Descriptive statistics***

  Suatu nilai yag merangkum data, tujuannya untuk menggambarkan keadaan data secara umum.

- ***Inferential Statistics***

  Metode yang digunakan untuk menarik kesimpulan terhadap `population` dengan memanfaatkan informasi dari `sample`.

- ***Measures of centaral tendency***

  Ukuran pemusatan data, menjelaskan titik sentral/pusat data.

- ***Mean***

  Nilai rata-rata (total seluruh data dibagi dengan banyaknya data) dari data yang bertipe numerik/angka. Nilai rata-rata suatu `population` dinotasikan oleh μ, sedangkan nilai rata-rata suatu `sample` dinotasikan oleh $\bar{x}$.

- ***Quartil***

  Nilai yang membagi data terurut menjadi 4 bagian sama besar (—Q1—Q2—Q3—).

- ***Median***

  Nilai yang membagi data terurut menjadi 2 bagian sama besar, sering disebut sebagai `Q2`.

- ***Outlier***

  Data yang nilainya sangat ekstrim, sering disebut sebagai data yang anomali.

- ***Trimmed mean***

  Nilai rata-rata yang diperoleh dari data terurut, yang sudah tidak terdapat `outlier` (`outlier` sudah dibuang).

- ***Mode***

  Nilai yang paling sering muncul/keluar dari data.

- ***Measures of spread***

  Ukuran penyebaran data, menjelaskan bagaimana persebaran suatu data.

- ***Variance***

  Nilai yang menggamabarkan seberapa bervariasi/beragamnya suatu data bertipe numerik/angka. Semakin besar nilai `variance` maka semakin beragam suatu data (heterogen), sedangkan semakin kecil nilai `variance` maka semakin sama/mirip setiap observasi pada data (homogen). Data yang observasinya bernilai sama, maka `variance` sama dengan 0.

- ***Standard deviation***

  Rata-rata selisih/jarak setiap observasi dengan nilai `mean`, diperoleh dari akar kuadrat `variance`.

- ***Range***

  Selisih/jarak antara observasi yang nilainya paling kecil (minimum) dengan yang nilainya paling besar (maksimum).

- ***IQR***

  Selisih/jarak antara `Q1` dan `Q3`.

- ***Standard error***

  Nilai yang menggambarkan kedekatan antara `sample` dan `population`. Semakin kecil nilai `standard error` maka semakin dekat/representatif pula `sample` menggambarkan `population`, dan sebaliknya.

- ***Covariance***

  Nilai yang menggambarkan hubungan (positif/negatif/tidak ada hubungan) antara dua variabel numerik. Namun `covariance` tidak dapat menggambarkan seberapa erat/kuat hubungan tersebut karena nilai `covariance` tidak memilki batasan yang mutlak (- inf, + inf).

- ***Correlation***

  Nilai yang menggambarkan keeratan hubungan (positif/negatif/tidak ada hubungan) antara dua variabel numerik. Nilai `correlation` mendekati 1 artinya kedua variabel berhubungan erat dan hubungannya positif, nilai `correlation` mendekati -1 artinya kedua variabel berhubungan erat dan hubungannya negatif, nilai `correlation` mendekati 0 artinya kedua variabel tidak saling berhubungan.

- ***Data distribution***

  Fungsi/bentuk yang menggambarkan persebaran data, sering disebut sebagai kumpulan nilai yang mungkin terjadi.

- ***Normal distribution***

  Fungsi/bentuk yang menggambarkan persebaran data bertipe numerik/angka, bentuknya hampir menyerupai sebuah lonceng (simetris). Data yang memiliki distribusi normal cenderung mengelompok di sekitar `mean` (pusat lonceng).

- ***Central limit theorem***

  Teorema ini menyatakan bahwa persebaran `sample` dengan distribusi tertentu yang diambil berulang kali dalam jumlah yang besar cenderung mengikuti `normal distribution`.

- ***Discreate variable***

  Variabel yang berisi data bertipe numerik/angka bulat, contoh: jumlah siswa dalam satu kelas.

- ***Continous variable***

  Variabel yang berisi data bertipe numerik/angka desimal, contoh: tinggi badan.

- ***Probability mass function***

  Peluang yang digunakan untuk mengambarkan kejadian pada `discreate variable`. Contoh: Dari total 50 siswa hanya 40 siswa yang hadir di kelas, maka peluang kehadiran siswa adalah 40/50 (0.8).

- ***Probability density function***

  Peluang yang digunakan untuk mengambarkan kejadian pada `continous variable`. Contoh: Peluang bertemu dengan orang yang memiliki tinggi badan minimal 165 cm di Algoritma.

- ***Standarization***

  Proses untuk menyeragamkan skala data yang berbeda.

- ***Standard scores***

  Nilai yang dihasilkan dari proses `Standarization`.

- ***z-score***

  `Standard score` yang dihasilkan dari proses `standarization` dengan memanfaatkan `normal distribution`.

- ***Confidence interval***

  Rentang nilai yang kemungkinan mengandung nilai parameter `population`, diperoleh dari informasi statistik `sample`.

- ***Margin of error***

  Nilai yang menggambarkan besar kesalahan dari pengambilan `sample`.

- ***Significance level (alpha)***

  Batas toleransi kesalahan/error yang diperbolehkan pada suatu pengujian.

- ***Hypothesis***

  Dugaan sementara terhadap masalah pada `population` yang harus diuji kebenarannya.

- ***Hypothesis test***

  Suatu metode yang digunakan untuk menarik kesimpulan/mengambil keputusan dari dua pernyataan/`hypothesis` yang saling bertolak belakang.

- ***Null hypothesis***

  Menyatakan bahwa nilai parameter populasi sama dengan nilai yang ditentukan atau tidak memiliki pengaruh yang signifikan.

- ***Alternative hypothesis***

  Menyatakan bahwa nilai parameter populasi tidak sama dengan nilai yang ditentukan atau memiliki pengaruh yang signifikan.

- ***p-value***

  Nilai/peluang kesalahan yang diperoleh dari hasil perhitungan statistik.

Data Visualization

- ***Plot***

  Sebuah visualisasi yang menggambarkan/merepresentasikan suatu data, sering disebut sebagai `chart`.

- ***Data wrangling***

  Serangkaian proses yang dilakukan untuk mebersihkan/merapikan data mentah/awal, sering disebut sebagai `data preparation`/`data cleansing`/`data pre-processing`.

- ***Aggregation***

  Rangkuman informasi suatu data yang sudah dikelompokkan berdasarkan suatu variabel/kolom tertentu.

- ***Exploratory***

  Langkah awal dalam melakukan analisis data yang bertujuan untuk megetahui karakteristik suatu data, biasanya dilakukan dengan membuat visualisasi sederhana.

- ***Explanatory***

  Hal yang dilakukan untuk menyampaikan informasi yang terkandung dalam suatu data, biasanya dilakukan dengan membuat visualisasi yang lebih kompleks dan informatif.

- ***Box plot***

  `Plot` yang digunakan untuk menggambrakan persebaran data yang bertipe numerik/angka.

- ***Scatter plot***

  `Plot` yang digunakan untuk menggambarkan persebaran dan hubungan (`correlation`) antara dua variabel bertipe numerik/angka.

- ***Regresion line***

  Sebuah garis yang menggambarkan pola linier dan hubungan (`correlation`) antar variabel.

- ***Histogram***

  `Plot` yang digunakan untuk menggambarkan persebaran data yang bertipe numerik/angka.

- ***Bins***

  Rentang nilai setiap batang pada `histogram`.

- ***Density Plot***

  `Plot` yang digunakan untuk menggambarkan persebaran data yang bertipe numerik/angka.

- ***Line chart***

  `Plot` yang digunakan untuk menggambarkan pola trend suatu data bertipe numerik/angka.

- ***Pie chart***

  `Plot` yang digunakan untuk menggambarkan jumlah/frekusensi data yang bertipe kategorik. `Pie chart` tidak disarankan untuk digunakan dalam `exploratory` maupun `explanatory` karena bentuknya yang menggunakan luas area untuk menggambarkan jumlah/frekuensi data.

- ***Bar plot***

  `Plot` yang digunakan untuk menggambarkan jumlah/frekusensi data yang bertipe kategorik.

- ***ggplot2***

  Kumpulan fungsi yang digunakan untuk membuat visualisasi.

- ***Mapping***

  Memetakan/menentukan `axis` ataupun elemen lain (warna, ukuran, dll) pada `plot` berdasarkan variabel-variabel yang tedapat pada data.

- ***Axis***

  Sumbu pada `plot`, terbagi menjadi sumbu horizontal(x) dan sumbu vertikal(y).

- ***Aesthetic***

  "Sesuatu yang dapat ditampilkan" (`axis`, warna, ukuran, bentuk, dll).

- ***Layer***

  Pada `ggplot2` dalam membuat `plot` menggunakan konsep lukisan, dimana plot dibuat dengan cara `layer-by-layer`.

- ***Geometry***

  Jenis `plot` yang ingin digunakan/ditambahkan.

- ***Reshaping***

  Mengubah bentuk data yang awalnya melebar menjadi memanjang, atau sebaliknya.

- ***Faceting***

  Membuat jenis `plot` yang sama menjadi beberapa bagian berdasarkan variabel tertentu.

- ***Leaflet***

  Salah satu `package` yang cukup populer untuk membuat visualisasi berupa peta interaktif.

Interactive Plotting and Web Dashboard

- ***User Interface (UI)***

  User interface merupakan tampilan atau bagian visual yang dapat dilihat oleh user/pengguna.

- ***Server***

  Server berisi perintah untuk proses `aggregation` dan membuat visualisasi yang akan ditampilkan pada dashboard.

- ***Dashboard***

  Dashboard merupakan media untuk mengumpulkan visualisasi terkait yang dapat meningkatkan pemahaman informasi yang lebih lengkap.

- ***Deployment***

  Deployment merupakan suatu proses yang bertujuan untuk mendistribusikan suatu `dashboard` yang dapat beroperasi dan diakses secara publik.

- ***Interactive plotting***

  Interactive plotting merupakan upaya meningkatkan informasi yang terdapat pada sebuah visualisasi (`plot`).

Regression Model

- ***Observation***

  Data yang dikumpulkan sebagai informasi, secara umum mengacu pada 1 baris data yang terdiri dari beberapa variabel.

- ***Target variable***

  Variabel yang ingin diprediksi/dimodelkan, sering disebut sebagai `respon/dependen variable`.

- ***Predictor***

  Variabel yang digunakan untuk memprediksi `target variable`, sering disebut sebagi `independen variable`.

- ***Dummy variable***

  `Dummy Variable` adalah hasil transformasi variabel kategorik dengan nilai 0 atau 1. Variabel ini digunakan untuk membuat data kategorik yang bersifat kualitatif menjadi kuantitatif.

- ***Residual/Error***

  Selisih antara nilai yang diprediksi dan nilai sebenarnya.

- ***Ordinary least square***

  Memperkirakan parameter model regresi dengan meminimumkan nilai Sum Squared Error (SSE).

- ***Intercept***

  Titik perpotongan antara sumbu y dengan garis regresi.

- ***Slope***

  Kemiringan garis regresi.

- ***R Squared***

  Persentase variansi target variabel yang dapat dijelaskan oleh model (`predictor`). Ukuran yang bisa digunakan untuk mengukur `kebaikan model`.

- ***Adjusted R Squared***

  Persentase variansi target variabel yang dijelaskan oleh model, perbedaannya dengan `R Squared` adalah `Adjusted R squared` memmperhitungkan banyak observasi dan signifikansi prediktor yang digunakan.

- ***Akaike Information Criterion (AIC)***

  Nilai yang menjelaskan besar informasi yang hilang pada model.

- ***Feature selection***

  Feature selection merupakan tahapan dalam memilih variabel yang akan digunakan.

- ***Stepwise***

  Stepwise adalah algoritma yang secara bertahap menambahkan atau mengurangkan variabel dengan mengacu pada nilai `AIC` terkecil.

- ***Mean squared error***

  Rata-rata dari `error` kuadrat. Ukuran yang bisa digunakan untuk mengukur `kebaikan model`.

- ***Homoscedasticity***

  Residual yang dihasilkan bernilai konstan dan tidak membentuk pola apapun.

- ***Multicolinearity***

  Keadaan dimana terdapat hubungan/`correlation` yang tinggi antar `predictor` pada model regresi.

Classification 1

- ***Classification***

  `Classification` adalah metode yang digunakan untuk memprediksi `target variable` bertipe kategorik (`factor`).

- ***Probability***

  Kemungkinan terjadinya suatu kejadian.

- ***Odds***

  Ukuran yang dapat menjelaskan probability. Dimana odds bisa di dapatkan dari `p/(1-p)` dimana `p` adalah peluang suatu kejadian terjadi.

- ***Odds ratio***

  Odds Ratio adalah perbandingan antara dua `odds`.

- ***Sigmoid function***

  `Sigmoid function` merupakan fungsi yang digunakan untuk mentransformasi nilai prediksi ke nilai peluang yaitu antara 0 hingga 1.

- ***Class imbalance***

  Keadaan dimana jumlah observasi antar kelas tidak seimbang.

- ***Data train***

  Data yang digunakan untuk membuat model.

- ***Data test***

  Data yang digunakan untuk menguji kebaikan model.

- ***Overfitting***

  Keadaan dimana model yang dibuat hanya dapat memprediksi dengan baik `data train`. Namun, ketika melakukan prediksi pada `data test`, model tersebut tidak dapat memprediksi dengan baik.

- ***Independence of observations***

  Antar observasi independen satu sama lain.

- ***Null deviance***

  `Null deviance` menunjukkan seberapa baik `target variable` diprediksi oleh model berdasarkan nilai `intercept`.

- ***Residual deviance***

  `Residual deviance` menunjukkan seberapa baik `target variable` diprediksi oleh model berdasarkan `intercept` dan semua `predictor` yang digunakan.

- ***Maximum likelihood estimator***

  `Maximum likelihood estimator` merupakan pendekatan statistik untuk memperkirakan paramater pada model `logistic regression`.

Classification 2

- ***Independent Events***

  Kejadian yang masing-masing tidak saling berkaitan satu sama lain.

- ***Dependent Events***

  Kejadian yang masing-masing saling berkaitan satu sama lain (kemungkinan terjadinya suatu kejadian akan memepengaruhi kemungkinan terjadinya kejadian lain.

- ***Laplace estimator***

  Menambahkan nilai pada setiap prediktor, biasanya adalah 1 pada masing-masing jumlah kejadian.

- ***Text mining***

  Analisis yang dilakukan dengan memanfaatkan teks sebagai prediktor.

- ***Corpus***

  Objek yang digunakan pada `text mining`, berisi kumpulan teks yang disebut sebagai dokumen.

- ***Punctuation***

  Tanda baca pada teks.

- ***Stopwords***

  Kumpulan kata yang tidak memiliki makna jika diikutsertakan dalam pemodelan klasifikasi.

- ***Stemming***

  Proses untuk mengambil kata dasar dari sebuah kata berimbuhan.

- ***DocumentTermMatrix***

  Proses untuk membagi setiap konten teks menjadi kata-kata yang mewakili `predictor`.

- ***ROC (Receiver Operating Characteristic)***

  Kurva yang menggambarkan performa model klasifikasi untuk seluruh `threshold`.

- ***AUC***

  Luas area di bawah kurva `ROC`, menggambarkan keberhasilan model klasifikasi dalam memprediksi/membedakan kedua kelas dari `target variable`.

- ***Entropy***

  Derajat kehomogenan.

- ***Information gain***

  Penurunan `entropy` setelah terjadi pembagian/spliting data.

- ***Pruning***

  Membatasi pembentukan cabang pada pohon (menyederhanakan pohon yang dibentuk).

- ***Terminal node***

  Bagian data yang sudah tidak dapat terbagi lagi.

- ***K-fold cross validation***

  Membagi data sebanyak k bagian, setiap bagian akan digunakan menjadi train dan test secara bergantian.

- ***Ensamble method***

  Kumpulan beberapa algoritma prediktif untuk memperoleh performa yang lebih baik.

- ***OOB (Out of Bag)***

  Besar error dari hasil prediksi pada data aktual yang belum dilihat oleh model.

- ***Variable importance***

  `Predictor` yang dianggap penting dalam model.

Unsupervised Learning

- ***Principal Component***

  Dimensi/variabel baru yang berisi rangkuman informasi dari keseluruhan variabel awal (data awal).

- ***Principal Components Analysis***

  Proses untuk membuat `principal comonent`.

- ***Eigen values***

  Nilai yang merepresentasikan jumlah/besar informasi (variansi) yang dimiliki oleh setiap PC.

- ***Eigen vector***

  Kumpulan nilai yang memproyeksikan data awal ke setiap `principal component`

- ***Biplot***

  Plot yang menggambarkan posisi data berdasarkan hasil `principal component analysis` dan besarnya pengaruh setiap variabel ke `principal component` 1 dan `principal component` 2.

- ***Reconstruct***

  Proses transformasi hasil `principal component analysis` ke data awal.

- ***Clustering***

  Proses mengelompokkan data berdasarkan jarak terdekat (kemiripan).

- ***Centroid***

  Pusat cluster.

- ***Between sum of square***

  Jarak tiap pusat cluster (`centroid`) ke pusat data secara keseluruhan.

- ***Within sum of square***

  Jarak tiap observasi ke `centroid` (pusat cluster) tiap cluster.

- ***Total sum of square***

  Jumlah nilai `Between sum of square` dan nilai `Within sum of square`

- ***Elbow method***

  Salah satu metode yang digunakan untuk menentukan jumlah cluster yang optimum.

Time Series

- ***Autocorrelation***

  `Correlation` antar data `observation` pada periode waktu yang berbeda.

- ***Smoothing***

  Smoothing merupakan transformasi data time series yang dapat membantu melihat pola pada data.

- ***Differencing***

  Differencing merupakan tranformasi data yang digunakan untuk membuat data time series stasioner.

- ***Autoregressive model (AR)***

  `Autoregressive model` hampir sama dengan model `linear regression`, namun `predictor` yang digunakan adalah nilai `target variable` itu sendiri pada masa lampau.

- ***Integrated (I)***

  Menjelaskan mengenai berapa kali model melakukan `differencing`.

- ***Moving average (MA)***

  `Moving average` digunakan untuk melakukan `smoothing` terhadap `error.`

- ***Trend***

  Data `trend` merupakan keadaan ketika observasi cenderung naik atau turun pada suatu periode waktu.

- ***Seasonal***

  Efek `Seasonal` terjadi jika data time series memiliki pola berulang pada siklus tertentu.

- ***Stationary patern***

  Keadaan dimana data time series berada di sekitar rata-rata.

Neural Network

- ***Optimization***

  Metode yang digunakan untuk meminimumkan error/kesalahan pada model neural network.

- ***Node***

  Unit terkecil pada arsitektur neural network yang berfungsi untuk melakukan transfer informasi, sering disebut sebagai `neuron`.

- ***Input layer***

  Lapisan pertama pada arsitektur neural network. Jumlah `node` pada `input layer` bergantung pada jumlah `predictor`.

- ***Output layer***

  Lapisan terakhir pada arsitektur neural network. Jumlah `node` pada `output layer` bergantung pada jenis `target variable`.

- ***Hidden layer***

  Lapisan yang terletak di antara `input layer` dan `output layer`. Jumlah `hidden layer` dan jumlah `node` di setiap `hidden layer` ditentukan oleh peneliti.

- ***Weight***

  Besar bobot yang menggambarkan besar informasi yang diteruskan dari setiap `node`. `Weight` ditetapkan secara random (acak).

- ***Linear regression***

  Suatu metode yang digunakan untuk memprediksi `target variable` bertipe numerik/angka.

- ***Bias***

  Pada `linear regression` sama seperti nilai `intersept (b0)`. `Bias` ditetapkan secara random (acak).

- ***Activation funcion***

  Fungsi yang digunakan untuk mengubah interval nilai (informasi) yang masuk ke setiap `node` pada `hidden layer` dan `output layer`.

- ***Cost function***

  Fungsi error.

- ***Feedforward***

  Proses pada neural network yang dimulai dari `input layer` hingga menghasilkan nilai prediksi.

- ***Backpropagation***

  Proses pada neural network ketika melakukan optimisasi dan melakukan update `weight`.

- ***Epoch***

  Satu kali proses `feedforward` dan `backpropagation`.

- ***Gradient***

  Hasil turunan dari `cost function`.

- ***Dummy Variable***

  Hasil transformasi varibel bertipe kategorik.

- ***Learning rate***

  Besar nilai yang menentukan seberapa besar `gradient` yang digunakan untuk melakukan update `weight`.

- ***Batch size***

  Jumlah observasi yang diikutsertakan untuk satu iterasi. Besar nilai yang menentukan seberapa besar `gradient` yang digunakan untuk melakukan update `weight`.

- ***Batch size***

  Jumlah observasi yang diikutsertakan untuk satu iterasi.