Angela Christabel / enjel

1. Salah satu contoh case penggunaan machine learning adalah dalam menentukan spam email. Tidak mungkin manusia me-review semua email milik user-nya satu-satu lalu menentukan mana yang harus masuk inbox, mana yang spam. Maka disini kita bisa menggunakan machine learning untuk mengotomisasi proses itu.
2. Ada 2 tipe learning dalam machine learning, yaitu supervised learning dan unsupervised learning. Supervised learning adalah saat data yang sudah ada labelnya, baik input atau outpunya, jadi model semacam di-“supervise” dalam trainingnya, untuk memprediksi atau mengeluarkan output sesuai label yang ada. Sedangkan untuk unsupervised learning, datanya tidak memiliki label, jadi model tidak “diawasi” dalam mempelajari datanya.
3. Validation adalah proses mengevaluasi sebuah model setelah di-fitting dan sambil mengatur hyperparameternya. Data yang digunakan adalah sebagian data yang diambil dari data training awal dan tidak digunakan dalam proses fitting, supaya kita bisa mendapatkan perkiraan objektif terhadap kemampuan sebuah model. Testing adalah proses mengevaluasi model di akhir setelah di-fitting dan di-tuning. Data yang digunakan adalah data yang berbeda sama sekali dengan data yang digunakan waktu training. Pembagian dataset yang benar untuk validation dan testing set adalah, membagi seluruh data 80/20, 80% untuk training, 20% untuk testing. Lalu training data tersebut bisa dibagi lagi 80/20, 80% nya itu untuk training, dan 20% dari data training awal dipakai untuk validation.
4. Classification adalah proses memprediksi kelas/kategori yang dimiliki berbagai data. Regression adalah proses memprediksi nilai-nilai dari data yang memiliki sifat kontinu.
5. Training data dapat menyebabkan overfitting ketika datanya mungkin kurang banyak, sehingga model yang mempelajari training data tersebut bisa memahami korelasi/hubungan di dalam data tersebut, tapi tidak bisa menangkapnya dalam data yang baru. Training data juga dapat menyebabkan underfitting ketika datanya memiliki terlalu banyak fitur, atau kurang bervariasi, sehingga sebuah model tidak dapat menangkap esensi atau korelasi yang ada dalam data tersebut.
6. Cross-validation adalah cara melakukan training lalu validasi terhadap model kita dengan bermacam-macam subset dari data yang kita miliki. Teknik ini banyak digunakan, karena dengan melakukan cross-validation, kita bisa memastikan bahwa model kita benar-benar sudah bagus dan stabil. Selain itu, jika kita memiliki dataset yang lumayan kecil, kita bisa memanfaatkan teknik cross-validation untuk mengetes model kita.
7. Discrete random variable adalah variable yang memiliki kemungkinan nilai yang berjumlah finit, sedangkan continuous random variable adalah variable yang memiliki kemungkinan nilai yang berjumlah infinit, biasanya di antara suatu range.
8. Tipe distribution yang ada:
   1. Bernoulli Distribution

Distribusi ini terjadi ketika ada 2 outcome dan 1 kali percobaan. Contohnya, dalam kasus melempar koin. Dalam 1 kali melempar koin, terdapat 2 nilai probabilitas yang mungkin terjadi, yaitu muncul uang atau gambar. Hasil dari Bernoulli Distribution dapat digambarkan dengan *p* dan *(1 – p)*.

Misalkan, dalam kasus melempar koin, kita mendapatkan nilai peluang muncul angka adalah 0.5, sehingga peluang muncul gambar adalah 1- 0.5 = 0.5.

* 1. Uniform Distribution

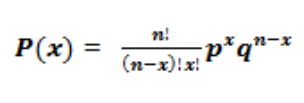
Distribusi ini terjadi ketika peluang terjadinya berbagai outcome itu sama besar, misalkan ketika kita melempar dadu. Peluang muncul angka 1 – 6 sama-sama 1/6. Rumus *probability density function* nya adalah:



Misalkan, jumlah bunga yang dijual di sebuah toko setiap hari didistribusi secara normal dengan maksimum 40 dan minimum 10. Peluang penjualan dalam sehari akan jatuh di antara 15 dan 30 adalah (30-15) \* (1/(40-10)) = 0.5.

* 1. Binomial Distribution

Distribusi ini terjadi ketika hanya 2 outcome dan lebih dari satu kali percobaan, dimana setiap percobaan bersifat independen, tidak mempengaruhi percobaan lain. Rumus *probability density function* nya adalah:



Misalkan saya melempar koin sebanyak 10 kali, saya ingin mengetahui peluang koin akan memunculkan gambar 3 kali. Misalkan p adalah peluang muncul gambar dalam 1 percobaan berarti p = 0.5 dan q = 0.5. Maka peluang saya memunculkan gambar 3 kali dari 10 percobaan adalah:

P(3) = 10!/(7! \* 3!) \* (p ^ 3) \* (q ^ 7)

1. Mean adalah rata-rata dari sebuah kumpulan bilangan. Variance adalah pengukuran seberapa jauh kumpulan bilangan tersebar.
2. Sebuah parametric model mengasumsi distribusi probabilitas yang memiliki sejumlah parameter yang fiks. Sedangkan non-parametric model tidak membuat asumsi tentang distribusi probabilitas data yang sedang dimodelling itu.
3. Loss function adalah sebuah fungsi untuk menghitung seberapa jauh sebuah nilai hasil prediksi dengan nilai aslinya. Loss function itu penting karena digunakan untuk melihat apakah model kita sudah pandai dalam memprediksi nilai aslinya atau belum, jika belum, kita bisa lihat sejauh apa hasil prediksi model kita. Dalam training data, kita meminimize loss function, karena semakin kecil loss yang dihasilkan oleh model kita, berarti model kita semakin dekat memprediksi nilai target.
4. Regularization adalah proses untuk memastikan model kita tetap berperforma stabil dan baik dengan data yang tidak pernah dilihat sebelumnya. Regularization berusaha mengurangi variance tanpa menambah bias. Optimization adalah proses mengurangi cost yang dihasilkan oleh model kita; semakin kecil semakin baik. Kita butuh regularization karena mungkin saja model kita berperforma baik dengan data yang sekarang, tapi jika variance nya besar maka model kita tidak akan berperforma baik dengan data yang asing. Akurasi dari model kita akan berubah-ubah dan tidak stabil karena tidak bisa berperforma baik dengan data yang asing. Cara kerjanya adalah mengurangi variance tanpa menambah bias.