# Introduction to Machine Learning

According to the father of Artificial Intelligence, John McCarthy, it is "The science and engineering of making intelligent machines, especially intelligent computer programs".

Now the question is how will the computer learn automatically?

Nah, jawabannya adalah data. Kita memasukkan data yang memiliki atribut atau fitur berbeda yang harus dipahami oleh algoritma dan memberi kita batasan keputusan berdasarkan data yang kita berikan. Setelah algoritma mempelajari dan menafsirkan data, yang berarti telah melatih dirinya sendiri, kita kemudian dapat menempatkan algoritma kita dalam fase pengujian dan tanpa memprogramnya secara eksplisit, masukkan titik data pengujian dan diharapkan memberi kita beberapa hasil.

# Regression

Regression analysis adalah salah satu bidang terpenting dalam statistik dan machine learning. Ada banyak metode regresi yang tersedia.Linear regression adalah salah satunya.

# What is Regression

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.Ini adalah salah satu metode regresi yang paling sederhana.Salah satu keuntungan utamanya adalah kemudahan interpretasi hasil.

## Problem Formlation

Saat menerapkan linear regression dari beberaoa variabel dependen 𝑦 pada himpunan variabel independen 𝐱 = (𝑥₁,…, 𝑥ᵣ), di mana 𝑟 adalah jumlah prediktor, kita mengasumsikan linear relationship antara 𝑦 dan 𝐱: 𝑦 = 𝛽₀ + 𝛽₁𝑥₁ + 𝛽₁𝑥₁ + 𝛽ᵣ𝑥ᵣ + 𝜀.Persamaan ini adalah persamaan regresi.𝛽₀, 𝛽₁,…, 𝛽ᵣ adalah koefisien regresi, dan 𝜀 adalah random error.

## Regression Performace

Variasi respon aktual 𝑦ᵢ, 𝑖 = 1,…, 𝑛, terjadi sebagian karena ketergantungan pada prediktor 𝐱ᵢ.Namun, ada juga varian tambahan yang melekat pada output.

Coefficient of determination, dilambangkan sebagai 𝑅², memberi tahu kita berapa banyak variasi dalam 𝑦 yang dapat dijelaskan oleh ketergantungan pada 𝐱 menggunakan model regresi tertentu.Lebih besar 𝑅² menunjukkan kesesuaian yang lebih baik dan berarti bahwa model dapat menjelaskan variasi keluaran dengan masukan yang berbeda dengan lebih baik.

Nilai 𝑅² = 1 sesuai dengan SSR = 0, yaitu perfect fit karena nilai prediksi dan respons aktual saling cocok satu sama lain.

## Simple Linear Regression

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

## Multiple Linear Regression

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

Jika hanya ada dua independent variables, estimated regression function-nya adalah 𝑓 (𝑥₁, 𝑥₂) = 𝑏₀ + 𝑏₁𝑥₁ + 𝑏₂𝑥₂.Ini mewakili bidang regresi dalam ruang tiga dimensi.Tujuan dari regresi adalah untuk menentukan nilai bobot 𝑏₀, 𝑏₁, dan 𝑏₂ sedemikian rupa sehingga bidang ini sedekat mungkin dengan respons aktual dan menghasilkan SSR minimal.

Kasus lebih dari dua independent variables adalah serupa, tetapi lebih umum.Estimated regression function adalah 𝑓 (𝑥₁,…, 𝑥ᵣ) = 𝑏₀ + 𝑏₁𝑥₁ + ⋯ + 𝑏ᵣ𝑥ᵣ, dan ada 𝑟 + 1 bobot yang harus ditentukan jika jumlah masukan adalah 𝑟.

## Polynomial Regression

Kita dapat menganggap polynomial regression sebagai kasus umum linear regression.Kita mengasumsikan polynomial dependence antara output dan inputs , dan menghasilkan, polynomial estimated regression function.

Dengan kata lain, selain istilah linier seperti 𝑏₁𝑥₁, fungsi regresi kita 𝑓 dapat menyertakan istilah non-linier seperti 𝑏₂𝑥₁², 𝑏₃𝑥₁³, atau bahkan 𝑏₄𝑥₁𝑥₂, 𝑏₅𝑥₁²𝑥₂, dan seterusnya.

Contoh paling sederhana dari polynomial regression memiliki satu independent variable, dan estimated regression function adalah polynomial of degree 2: 𝑓 (𝑥) = 𝑏₀ + 𝑏₁𝑥 + 𝑏₂𝑥².Sekarang, ingatlah bahwa kita ingin menghitung 𝑏₀, 𝑏₁, dan 𝑏₂, yang meminimalkan SSR.

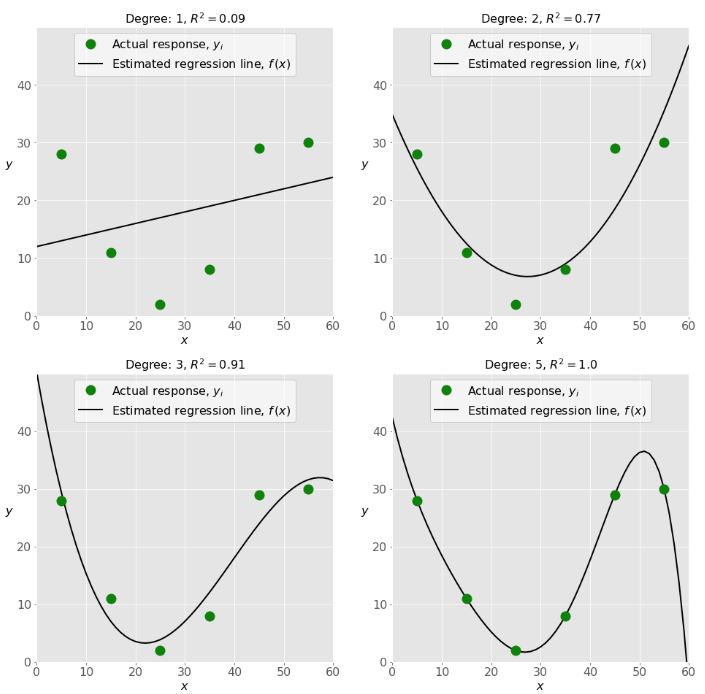
## Underfitting & Overfitting

Satu pertanyaan sangat penting yang mungkin muncul saat kita menerapkan polynomial regression terkait dengan pilihan optimal degree dari polynomial regression function.

Tidak ada aturan langsung untuk melakukan ini.Tergantung kasusnya.Namun, kita harus menyadari dua masalah yang mungkin mengikuti choice of the degree: underfitting dan overfitting.

Underfitting terjadi jika model tidak dapat menangkap dependensi di antara data secara akurat, biasanya sebagai konsekuensi dari kesederhanaannya sendiri.Seringkali menghasilkan 𝑅² 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 𝑅² yang tinggi.Namun, mereka sering tidak menggeneralisasi dengan baik dan memiliki lebih rendah secara signifikan 𝑅² saat digunakan dengan data baru.



Plot kiri atas menunjukkan garis linear regression yang memiliki 𝑅² rendah.Mungkin juga penting bahwa garis lurus tidak dapat mengambil fakta bahwa actual response meningkat saat 𝑥 menjauh dari 25 menuju nol.Ini adalah contoh underfitting.

Plot kanan atas menggambarkanpolynomial regression dengan degree sama dengan 2. Dalam contoh ini, mungkin derajat optimal untuk memodelkan data tersebut.Model ini memiliki nilai 𝑅² yang memuaskan dalam banyak kasus dan menunjukkan tren dengan baik.

Plot kiri bawah menampilkan polynomial regression dengan degree sama dengan 3. Nilai 𝑅² lebih tinggi dari pada kasus sebelumnya.Model ini berperilaku lebih baik dengan data yang diketahui daripada yang sebelumnya.Namun, ini menunjukkan beberapa tanda overfitting, terutama untuk nilai input yang mendekati 60 saat garis mulai menurun, meskipun data sebenarnya tidak menunjukkannya.

Akhirnya, pada plot kanan bawah, kita dapat melihat kesesuaian yang sempurna: enam titik dan polynomial line dengan derajat 5 (atau lebih tinggi) menghasilkan 𝑅² = 1. Setiap respons aktual sama dengan prediksi yang sesuai.

Dalam beberapa situasi, ini mungkin yang kita cari.Namun, dalam banyak kasus, ini adalah model yang terlalu pas.Ini cenderung memiliki perilaku yang buruk dengan data yang tidak terlihat, terutama dengan input yang lebih besar dari 50.

Sebagai contoh, ini mengasumsikan, tanpa bukti apapun, bahwa terdapat penurunan yang signifikan dalam respon untuk 𝑥> 50 dan 𝑦 mencapai nol untuk 𝑥 mendekati 60. Perilaku tersebut merupakan konsekuensi dari usaha yang berlebihan untuk mempelajari dan menyesuaikan dengan data yang ada.