Machine Learning Toolkit

fundera

What is Machine Learning?

Learning = Improving with experience at some task

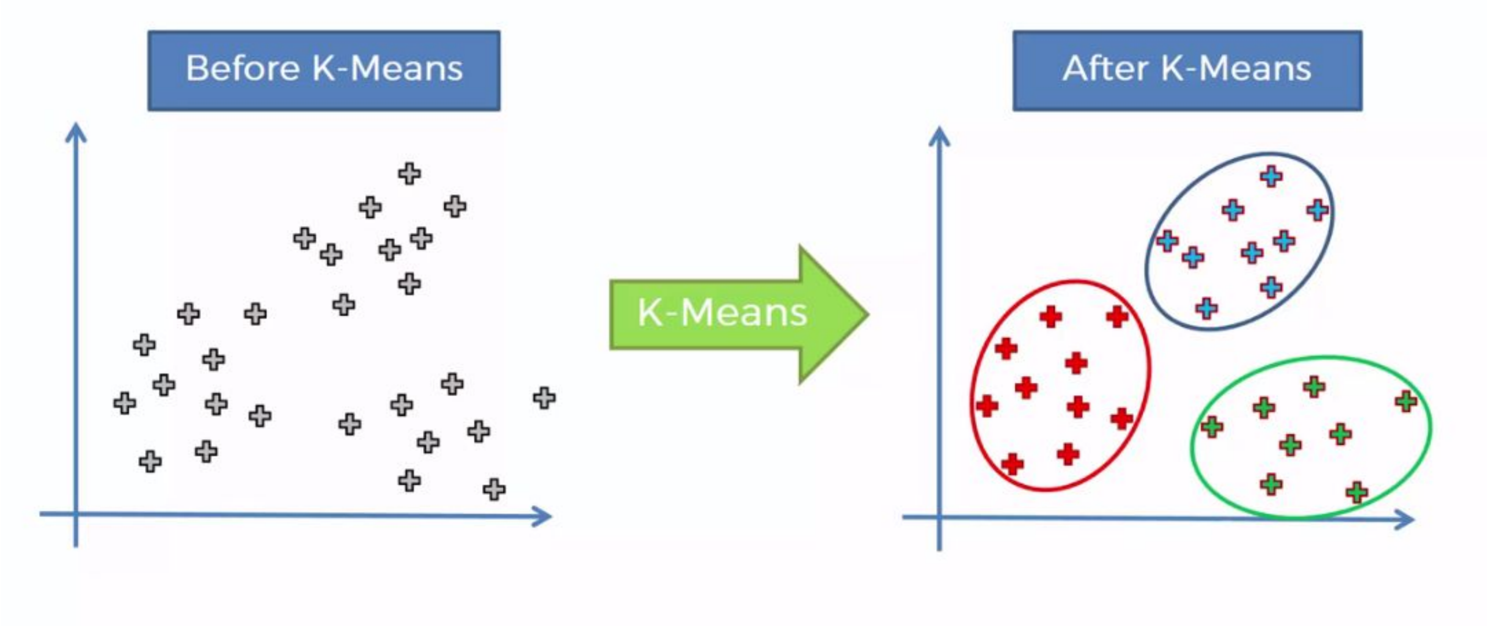
- Improve at task T
- With respect to performance measure P
- Based on experience E

Why is it cool now?

- Tons of data with advent of internet
- Increased computational power
- Progress in algorithms and related theory
- Support and interest from industries

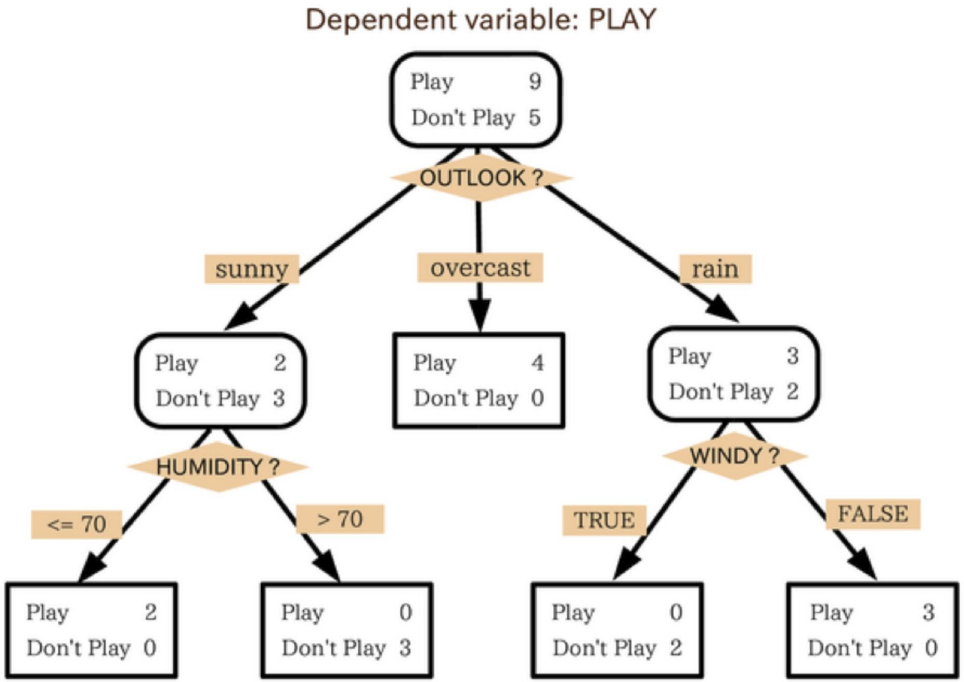
Types of Learning

<u>Unsupervised Learning:</u> K-means clustering example



Types of Learning

<u>Supervised Learning:</u> Decision Trees (Decision to play tennis)



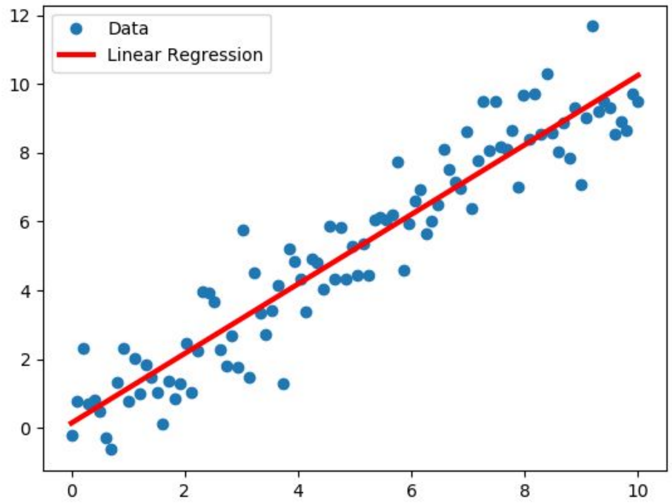
Types of Machine Learning Tasks

- Regression: predict a real value.
- Classification: predict a class a data point belongs to.

fundera

Linear Regression (Supervised)

- Predict height (y-axis) given age (x-axis)
- $(1/n)\sum(y - y')^2$ Mean Square Error.
  - y is actual value (**label**)
  - y' is value predicted by model
- $y' = w_0 + w_{age} * age$
  - w_0, w_{age} are the **weights**
  - age is the only **feature**
- We minimize MSE to get the weights (**model**)



fundera

Linear Regression (Examples & Instances)

Training Examples

Age	Weight
2	8
1	5
7	23
5	17

Testing Examples

Age	Weight
3	12
8	25

Prediction Instances

Age	Weight
4	?
6	?

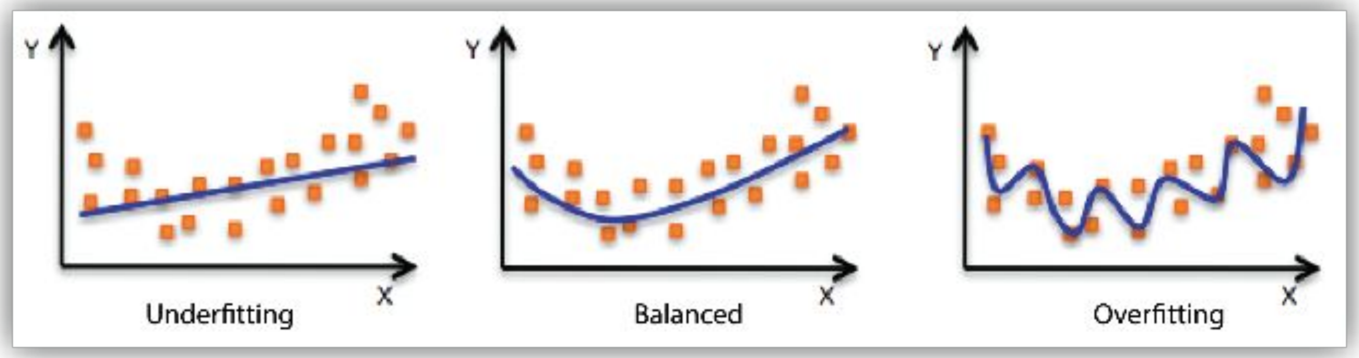
Mean Square Error

- Two approaches to minimizing mean square error:
  - Analytically: set derivatives with respect to the weights to zero and solve the resulting equations
  - Gradient Descent:
    - Gradient is direction of steepest increase of a function
    - By going in opposite direction we reach the minima
    - Global minima requires function to be convex; MSE is convex

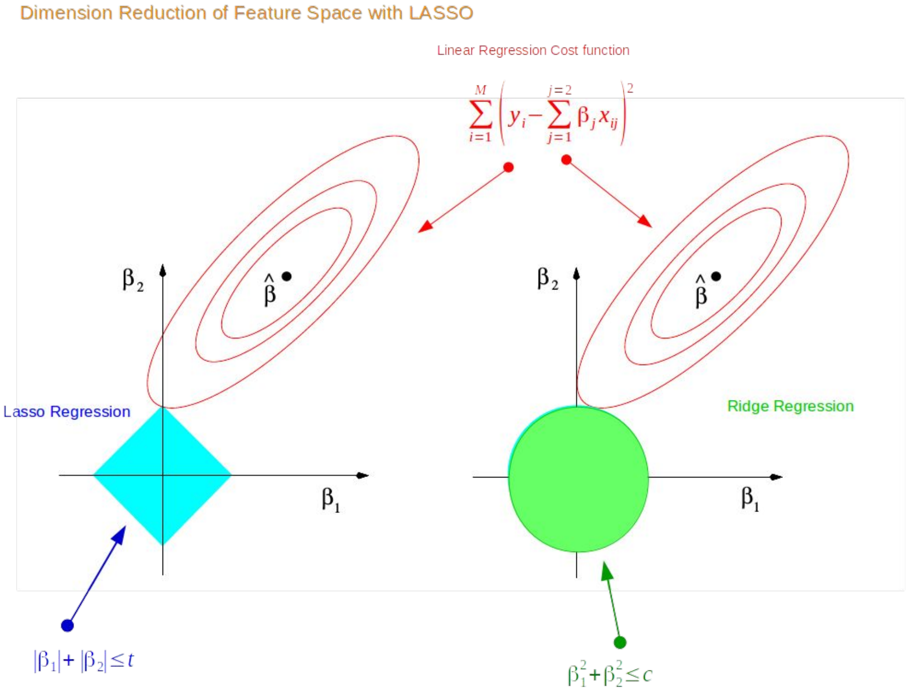
Loss(/Objective/Cost) Function And Maximum Likelihood

- MSE is a loss function we are trying to minimize
- Sometimes also referred to as the objective or cost function
- Likelihood:
  - Conditional probability: P(A|B) probability of A given B
  - As per Bayes' theorem:
    - P(model|data) $\propto$ **P(data|model)** * P (model)
  - Likelihood: what model maximizes the probability of the data seen?
- Maximum Likelihood Estimation: When minimizing MSE, we maximizing likelihood corresponding to likelihood.
- In general:
  - Likelihood $\propto$ $e^{-k(Loss)}$

Overfitting vs Underfitting

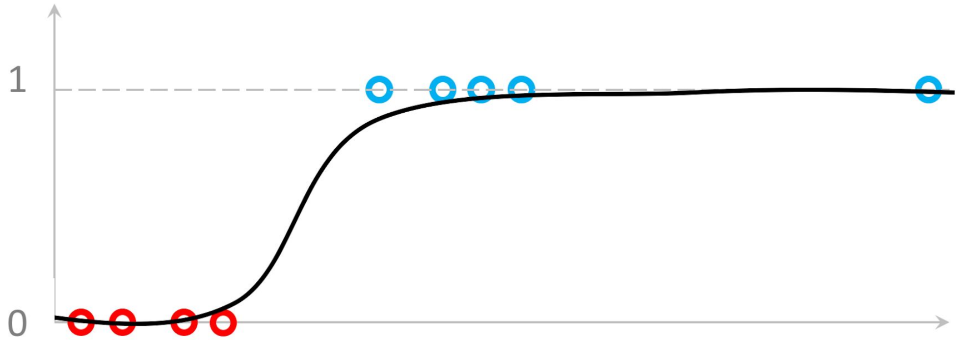


Regularization



- Each contour here is a curve of constant error value (eg: MSE value).
- The shaded area is the constraint on the weights.
- The error can't be too small otherwise the constraint is not met.
- L1 (Lasso) causes some weights to drop off.
- L2 (Ridge) causes the weights to become smaller in general.

Logistic Regression (Supervised, Classification)



- Logistic function gives values between 0 and 1. Can interpret as probability.
- Labels for training are 0 and 1, but the predictions are real values between 0 and 1.
- We use it so that we can do gradient descent on a continuous and differentiable function for a classification task.

Logistic Regression (Contd.)

- Pretend we are trying to predict whether a child is malnourished.
- Logistic function: $y = 1/(1 + e^{-J})$ where $J = w_0 + w_{age} * age + w_{weight} * weight$
- Training Data:

Age	Weight	Malnourished (Label)
1	5	0
1	2	1
2	10	0
2	4	1