

Tipologie di modelli:

Random tree forest

Evoluzione del decision tree: albero con due tipologie di nodi,
Nei primi vi è una condizione if/else sulla proprietà dell'input
L'obiettivo è di isolare le classi del problema rappresentato nei nomi foglia dell'albero.

Tende a non essere influenzato troppo dal feature selection
Aiuta però a calcolare l'entropia delle variabili predittive

La condizione if/else viene determinata in base all'entropia: ciò che lo rende un ML
Ad ogni possibile condizione viene determinato il livello di information gain e viene scelto quello migliore

Alternativa alla regressione lineare, visualizzare il dataset per determinare quale metodologia applicare; se una classe è raggruppata in una porzione del piano è adatto il decision tree; se una classe ha gli elementi disposti su una retta è roba da regressione lineare

SVM - Macchine a vettori di supporto

Modello di classificazione supervisionato
Ogni elemento da classificare viene rappresentato tramite un punto in uno spazio N dimensionale. Le coordinate del punto saranno le features.
L'obiettivo di un SVM è tracciare un iperpiano sicché i semispazi ottenuti contengano classi omogenee.

Margine: distanza tra l'iperpiano di tradeoff e il punto più vicino della classe

Supporting vector: iperpiano parallelo al tradeoff che passa per il punto estremo della classe

Se i punti dello spazio non possono essere separati da un iperpiano:

1. Aggiungere una feature
2. The kernel trick

Data Analysis

In questa fase vengono analizzate le variabili del dataset, valutate delle ipotesi e calcolato p per dimostrare la validità di tale ipotesi

Null hypothesis: ipotesi secondo la quale non esiste un rapporto statistico tra diverse feature; non esiste una differenza tra due gruppi

The null hypothesis states there is no relationship between the measured phenomenon (the dependent variable) and the independent variable.





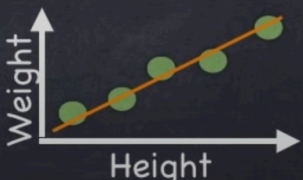
Ex il colore della luce non influisce sulla crescita della pianta

Alpha value:

Se il p value è inferiore all'alpha value la tesi è valida

Once you have the null and alternative hypothesis in hand, you choose a significance level: a probability threshold that determines when you reject the null hypothesis.

After carrying out a test, if the probability of getting a result as extreme as the one you observe due to chance is lower than the significance level, you reject the null hypothesis in favor of the alternative. This probability of seeing a result as extreme or more extreme than the one observed is known as the p-value

	What we observe in our sample data	Is it real?
One categorical		1 sample proportion test
Two categorical		Chi squared
One numeric		t-test
One numeric and one categorical		t-test or ANOVA
Two numeric		<u>correlation test</u>

A seconda della combinazione di variabili utilizzate ne dipende il test effettuato

QQ-plot

High level speaking, QQ-plot (Quantile-Quantile plot) is a scatter plot, often be used to check if a variable follows the normal distribution (or any other distributions). If all points on the QQ-plot form (or almost form) a straight line, it is a high chance that the examining variable is normally distributed.

ppplot : Probability-Probability plot Compares the sample and theoretical probabilities (percentiles).

qqplot : Quantile-Quantile plot Compares the sample and theoretical quantiles

probplot : Probability plot Same as a Q-Q plot, however probabilities are shown in the scale of the theoretical distribution (x-axis) and the y-axis contains unscaled quantiles of the sample data.

the QQ-plot has a higher deviation at 2 tails (i.e. QQ-plot has fewer points at the 2 tails), while for PP-plot, the deviation is higher in the middle. And as researchers often give more attention to the tails, **the QQ-plot is more popular in practice.**

T-test

The T-test is a statistical test used to determine whether a numeric data sample differs significantly from the population or whether two samples differ from one another.

Chi-squared test

common statistical test for categorical variables

Correlation Test

Determinare unicità delle variabili tramite metodo della correlazione e/o correlation matrix

Oltre a mostrare linearità, aiuta a far vedere l'eterogeneità delle feature: se i punti delle classi formano insiemi sovrapposti o quasi, tali feature non sono adatte sicché è impossibile separare i punti nello spazio

To do list per raffinare il modello

1. Analisi sull'unicità delle variabili predittive tramite metodo correlazione
2. Confrontare SVM, Random Tree
dal random tree è possibile determinare l'information gain delle variabili predittive
3. Analisi sull'information gain
4. Analizzare performance modello