(From Intro to Machine Learning, Women Who Code)

## Types of Machine Learning

Based on the availability and data types of example data and target responses machine learning can be broadly classified into: <BR/>

\*\*▸ Supervised learning:\*\*<BR/>

algorithm learns from example data and

associated target responses, under supervision<BR/>

- Classification - for categorical/nominal data<BR/>

- Regression - for continuous/numeric data<BR/><BR/>

\*\*▸ Unsupervised learning\*\*<BR/>

algorithm learns from example data

without any associated response, it determined data patterns

on its own.<BR/><BR/>

\*\*▸ Semi-supervised learning:\*\*<BR/>

algorithm learns from partially labeled data sets.<BR/><BR/>

\*\*▸ Reinforcement learning\*\*<BR/>

algorithm learns from example data

along with positive/negative feedback to

each prediction the algorithm proposes.

algorithm must make decisions and the

decisions bear “consequences”.

## Other concepts:<BR/>

\*\*Normal Distribution:\*\* <BR/>

It is a type of population distribution that is commonly found.<BR/>

\*\*Hypothesis Test\*\*

H0: null hypothesis - regular<BR/>

H1: violation of the assumption<BR/>

- p-value is a quantity that we can use to

interpret or quantify the result of the test

- α is the significance level that is used to either accept or reject the hypothesis <BR/>

- α commonly is 5% or 0.05.<BR/>

- p-value > α: Fail to reject the H0

(not significant result).<BR/>

- p-value <= α: Reject th e null H0

(significant result).<BR/>

\*H0 = Data is normally distributed\* <BR/>

\*H1 = Data is not normally distributed\* <BR/>

\*Confidence level: subtracting significance level from 1 (confidence = 1 - α)\* <BR/>

## Errors in Statistical Tests

The interpretation of a statistical hypothesis test is probabilistic - the evidence of the test may suggest an outcome and be mistaken. <BR/>

Types errors: <BR/>

▸ Type I Error: The incorrect rejection of a

true H0 (ie. false positive). <BR/>

▸ Type II Error: The incorrect failure of

rejection of a false H0 (ie. false negative)<BR/>

<p align="center">

<img src = "../images/typesOfTests.png" > <br/>

<p>

## z-test

Used to determine whether two population means are different when<br/>

- their variances are known<br/>

- the sample size is large.<br/>

It is assumed to follow normal distribution<br/>

A z-statistic, or z-score, is a number representing how many standard

deviations above or below the mean population a score derived from a z-test

is. <br/>

## t-test

Used to evaluate if the means of two sets of data are statistically significantly different from each other. <br/>

Types of t-tests <br/>

- One-sample - compare the mean of a population with a theoretical value. <br/>

- Unpaired Two-sample - compare the mean of two independent samples. <br/>

- Paired - compare the means between two related groups of samples. <br/>

Degrees of freedom is 1 less than sample size (n-1).<br/>

<p align="center">

<img src = "../images/MLprocesses.png" > <br/>

<p>

\*\*Open-source Datasets:\*\* <br/>

[UCI Machine Learning Repository](https://archive.ics.uci.edu/ml/datasets.php)<br/>

[Kaggle](https://www.kaggle.com/datasets)<br/>

[US Government](https://www.data.gov/)<br/>

[US Census](https://www.census.gov/data.html)<br/>

## Classification

- dependent events

- independent events

### Probability

Used to defined probability between dependent events

- Consider 2 events, A and B where A occurs

before B

- Conditional probability is used to calculate

the probability of B occuring after A has

occurred

- The occurrence of A changes the probability

of the occurrence of B

<p align="center">

<img src = "../images/condProbab.png" > <br/>

<p>

<p align="center">

<img src = "../images/condProbab2.png" > <br/>

<p>

<p align="center">

<img src = "../images/Bayes1.png" > <br/>

<p>

<p align="center">

<img src = "../images/Bayes2.png" > <br/>

<p>

<p align="center">

<img src = "../images/Bayes3.png" > <br/>

<p>

## Naive Bayes Assumption

Each feature makes an independent and equal contribution to the outcome

## Terminology

- Feature(s), x <br/>

Column(s) in a dataset that the target response is dependent on

- Class Variable, y <br/>

Column in a dataset that the user would like to predict

### about probability

- Evidence<br/>

The event that has already occurred (Event A)

- Prior Probability<br/>

Probability of the event that has already occured, P(A)

- Posterior Probability<br/>

Probability of the event that is to occur, P(B)

## INSERIR FORMULAS DE BAYES PARA ML

## Naive Bayes Drawbacks

- Features need to be categorical in nature

- Assumption of no dependence between features (naive)

- Features need to be categorical in

nature

- No occurrence of a class value drops

its posterior probability to