PROCESSING OF BIG DATA

SPARK

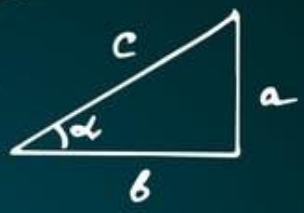
SESSION-2

$\sin\alpha = BC = \frac{a}{c}$;

$\cos\alpha = OB = \frac{b}{c}$;

$tg\,\alpha = OB = \frac{b}{c}$;

$ctg\,\alpha = OA = \frac{a}{b}$;



$\sin 2\alpha = 2\sin\alpha\cos\alpha$;

$\cos 2\alpha = \cos^2\alpha - \sin^2\alpha$;

$tg\,2\alpha = \frac{2\,tg\,\alpha}{1 - tg^2\alpha}$;

$\alpha^\circ = \frac{180}{\pi}\alpha$; $\alpha = \frac{\pi}{180}\alpha^\circ$;

$360^\circ = 2\pi$; $180^\circ = \pi$;

$\sin^2\alpha + \cos^2\alpha = 1$;

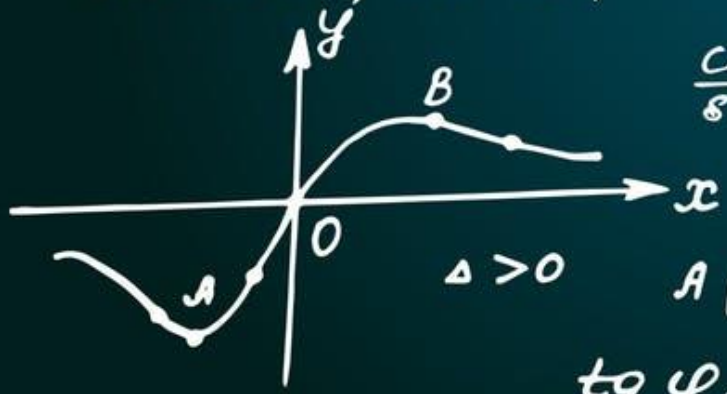
$\frac{\sin\alpha}{\cos\alpha} = tg\,\alpha$;

$\sin\alpha \cdot \csc\alpha = 1$;

$u = A\sin(\omega t + \varphi)$

$u = a\sin\omega t + b\cos\omega t$

$\frac{\cos\alpha}{\sin\alpha} = ctg\,\alpha$



$\Delta > 0$

$x = -\frac{b}{2a}$;

$\Delta = 4ac - b^2$

$A\left(-\frac{b}{2a}; \frac{4a}{\Delta}\right)$

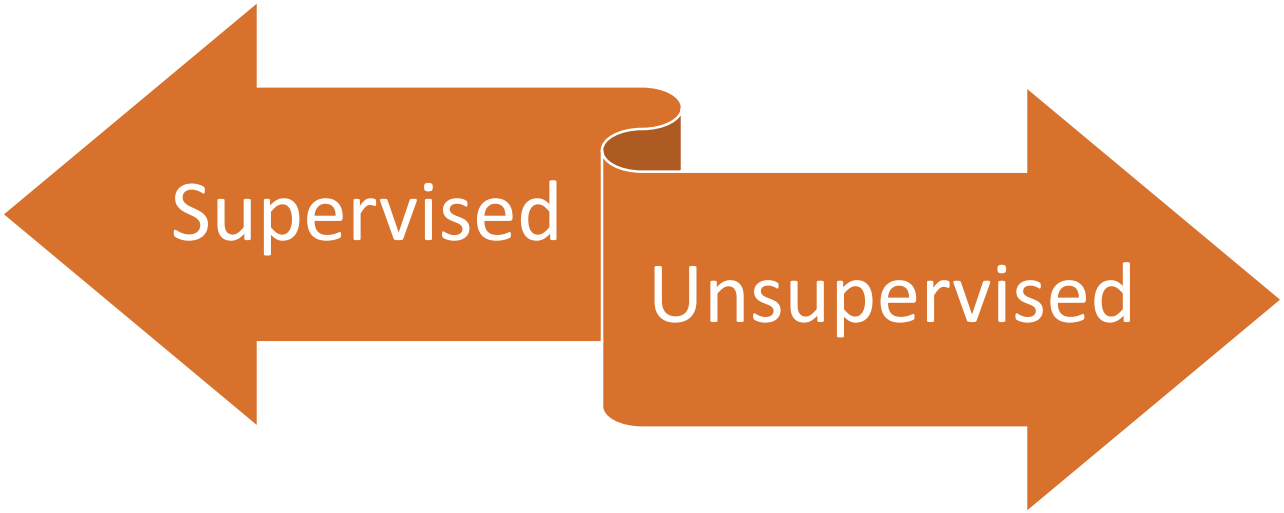
$a > 0$;

$tg\,\varphi = \pm a^2 \left(\frac{3}{\Delta}\right)^{\frac{3}{2}}$;

Agenda

- Common statistics functions
- Components of a typical ML program
- ML algorithms
  - Hotel review classification
  - Wine customer segmentation

ML algorithm types



Introduction

- Spark has two sets of ML libraries
  - Mllib: RDD based API under spark.mllib package
  - Ml: Dataframe based API spark.ml package

- Mllib is under maintenance mode only. There is no further enhancements except bug fixes

- ML is the latest API which will be carried forward in future

Correlation matrix

	Salary	Experience	Health
Salary	1.0	0.97	-0.56
Experience	0.97	1.0	0.0
Health	-0.56	0.0	1.0

Independence hypothesis

- Check dependency of two variables
- For example: Does voting preference depend on gender?
- Calculate chi-square for a sufficient sample data to validate pValue against hypothesis
- Variables must be exclusive and categorical

	Democrat	**Republican**
Male	20	30
Female	50	40

Independence hypothesis steps

1. State the hypothesis:
   - Gender and voting preferences are independent

2. Analyze sample data
   - Calculate chi-square and p-value

3. Compare p-value with significance
   - If p-value is less than significance (0.05 generally) then we can't accept hypothesis which means there is some relevance between Gender and Voting preference

Components of Spark ML

- Dataframe
- Transformer: transforms one dataframe to another
- Estimator: takes dataframe as input and generates model
- Parameter: parameters used while training the model
- Pipeline: discussed soon

Reference

- https://spark.apache.org/docs/latest/ml-guide.html
- http://spark.apache.org/docs/2.4.0/api/python/pyspark.ml.html
- https://stattrek.com/chi-square-test/independence.aspx
- https://www.khanacademy.org/math/multivariable-calculus/multivariable-derivatives