CS 696 Intro to Big Data: Tools and Methods
Fall Semester, 2016
Doc 17 Review
Sep 26, 2016

Reading

Getting started with Julia Programming Language Chapters 2-5

Julia for Data Science Chapters 2-11

Julia

Basic Types
Regular Expressions
Destructuring
Compound expressions
Control - if while for ranges
Type declarations
functions
exceptions
user defined types
immutable types, subtypes
Dictionaries, Arrays
    Vectors, column & row
Matrix operations
Column storage
vectorizing functions
@parallel
@devec

Kahan summation
comprehensions
function parameter options
lambda, Anonymous Functions
Multiple dispatch
map, reduce, filter
|>
Unit testing
Performance Issues
    Top level code, Type stability
    Change types of variables
DataFrames
Parallel Processing
    @spawn, addprocs
    @parallel
Speedup, Amdahl's law
Pleasingly parallel
Shared arrays
ArrayFire

Statistics

mean
median
mode
variance
standard variation, Bessel's correction
quantiles
ploting
    box plots, beeswarm, violin
Distributions
    Normal
Hypothesis testing
confidence interval, standard error
Central limit theorem

Regression

supervised, unsupervised learning
No Free Lunch Theorems
Linear regression

Pearsons correlation
r^2

Logistic (Logit) Regression or Logit Model

Multiple linear regression

Generalized linear regression (model)

Is the dependent variable related to the independent variable

Generating the model

Error in the model

Effect of independent variables

Clustering

Distances

normalization

max-min, mean-standard dev,

Sigmoidal, softmax

Text

Jaccard Distance, cosine

k-means, k-medoids

DBSCAN

Curse of Dimensionality

PCA - Principle Component Analysis

Picking initial means

Picking number of clusters

Measuring how good the clusters are

Silhouettes, Dunn index, Davies-Bouldin

Normalization of data

What is distance