IST 687: Week 8 Supplemental

Corey Jackson

2021-03-03 19:50:33

Today's Agenda

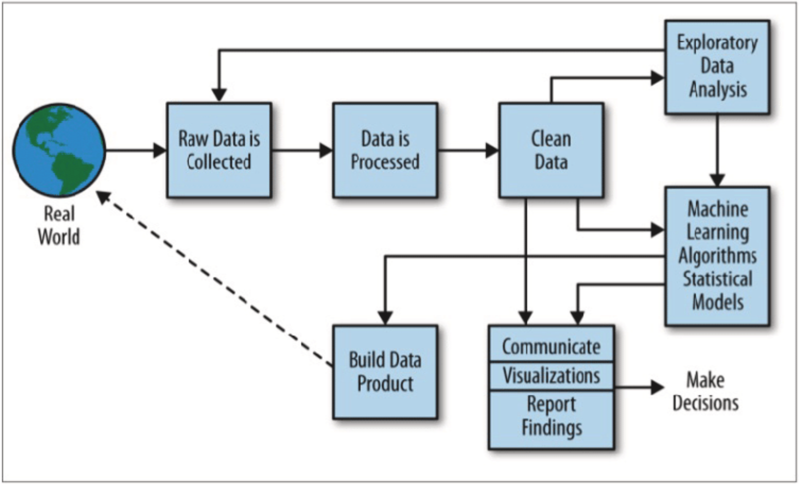
- Announcements
- Towards data modeling
- Next Week
- Exam Logistics/Code

Announcements

- Office hours Wednesday 6-7pm EDT or by appointment
- Homework/Lab 1 through 8 answers available on the syllabus
- HW 7 Update
- Final Project (Project Update III in Week 10)

Taking a Step Back

Towards data modeling: data science process



Towards data modeling: exploratory data analysis

EDA Methods

Descriptive statistics

Central tendency of a distribution

Mean	$\frac{\text{Sum of all values}}{\text{Total number of values}}$
Median	Middle value(when data are arranged in order
Mode	Most common value

Measure of Variation

Variance	how far a set of numbers are spread out from mean
Interquartile range	divides a data set into quartiles.
Standard deviation	dispersion of a set of data from mean

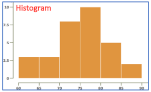
Skewness & Kurtosis

Skewness	**Measure of symmetry**
Kurtosis	**Kurtosis is a measure of "peakedness" relative to a Gaussian shape**

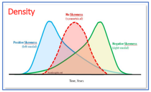
Visualizations

1-dimension

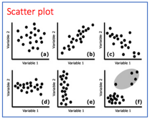
Few data points



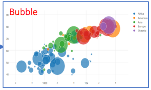
Many data points



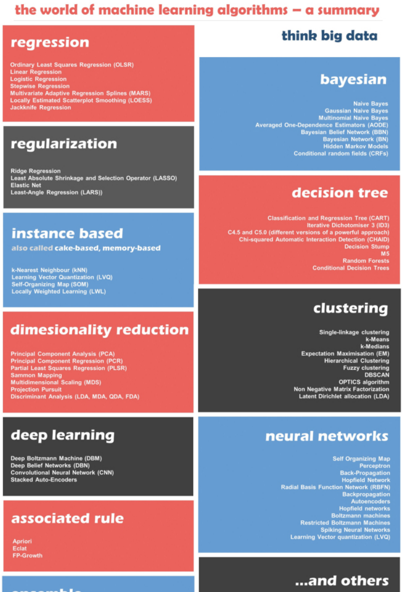
2-dimension



3-dimension



Towards data modeling: Many algorithms and modeling techniques



the world of machine learning algorithms – a summary

think big data

regression

Ordinary Least Squares Regression (OLSR)
Linear Regression
Logistic Regression
Stepwise Regression
Multivariate Adaptive Regression Splines (MARS)
Locally Estimated Scatterplot Smoothing (LOESS)
Jackknife Regression

regularization

Ridge Regression
Least Absolute Shrinkage and Selection Operator (LASSO)
Elastic Net
Least-Angle Regression (LARS))

instance based

also called cake-based, memory-based

k-Nearest Neighbour (kNN)
Learning Vector Quantization (LVQ)
Self-Organizing Map (SOM)
Locally Weighted Learning (LWL)

dimesionality reduction

Principal Component Analysis (PCA)
Principal Component Regression (PCR)
Partial Least Squares Regression (PLSR)
Sammon Mapping
Multidimensional Scaling (MDS)
Projection Pursuit
Discriminant Analysis (LDA, MDA, QDA, FDA)

deep learning

Deep Boltzmann Machine (DBM)
Deep Belief Networks (DBN)
Convolutional Neural Network (CNN)
Stacked Auto-Encoders

associated rule

Apriori
Eclat
FP-Growth

bayesian

Naive Bayes
Gaussian Naive Bayes
Multinomial Naive Bayes
Averaged One-Dependence Estimators (AODE)
Bayesian Belief Network (BBN)
Bayesian Network (BN)
Hidden Markov Models
Conditional random fields (CRFs)

decision tree

Classification and Regression Tree (CART)
Iterative Dichotomiser 3 (ID3)
C4.5 and C5.0 (different versions of a powerful approach)
Chi-squared Automatic Interaction Detection (CHAID)
Decision Stump
M5
Random Forests
Conditional Decision Trees

clustering

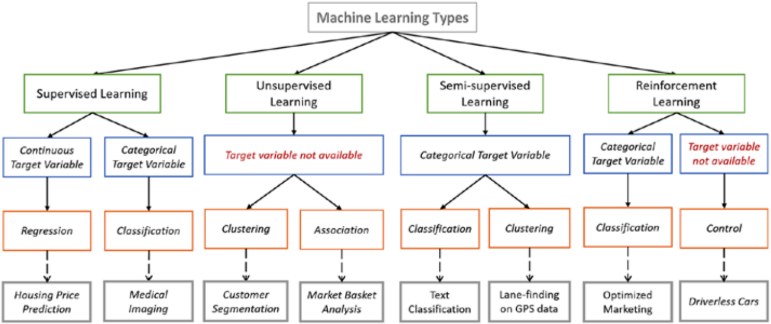
Single-linkage clustering
k-Means
k-Medians
Expectation Maximisation (EM)
Hierarchical Clustering
Fuzzy clustering
DBSCAN
OPTICS algorithm
Non Negative Matrix Factorization
Latent Dirichlet allocation (LDA)

neural networks

Self Organizing Map
Perceptron
Back-Propagation
Hopfield Network
Radial Basis Function Network (RBFN)
Backpropagation
Autoencoders
Hopfield networks
Boltzmann machines
Restricted Boltzmann Machines
Spiking Neural Networks
Learning Vector quantization (LVQ)

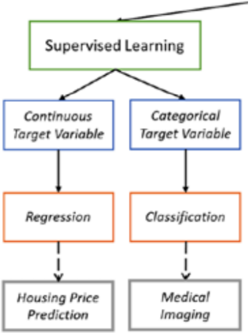
...and others

Towards data modeling: Machine learning algorithms and statistical models



Towards data modeling: Supervised Learning

Supervised learning algorithms consist of having a target/outcome variable (or dependent variable) which is to be predicted from a given set of predictors (independent variables).



Towards data modeling: Supervised Learning Regression

Speed and stopping distances of cars. The data give the speed of cars and the distances taken to stop. Note that the data were recorded in the 1920s.

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
## 4     7   22
## 5     8   16
## 6     9   10
```

Question: How does the speed of a car affect the distance needed to travel to come to a complete stop?

Supervised Learning Regression

```
##   speed dist
## 1     4    2
## 2     4   10
## 3     7    4
```

Question: How does the speed of a car affect the distance needed to travel to come to a complete stop?

Independent variable: speed
Dependent variable: dist (the distance depends on speed)

Supervised Learning: Regression

Modeling simple regression using R: `lm(dist ~  speed, data = cars)`

Model Output: dist = -17.579 + 3.932*speed
Application A car traveling 20 mph will take ? feet to stop

```
61.061 =  -17.579 + 3.932*(20)
```

Application A car traveling 20 mph will take 61.061 feet to stop

Supervised Learning: Classification

A dataset of student exam scores and whether they passed the course

Question: *What is the relationship between a student's exam score and passing?*

```
##   sid grade final
## 1 101     8     1
## 2 101    10     1
## 3 102     9     1
## 4 103     4     0
## 5 103     2     0
## 6 103     6     0
```

Supervised Learning: Classification

Question: *What is the relationship between a student's exam score and passing?*

```
##   sid grade final
## 1 101     8     1
## 2 101    10     1
## 3 102     9     1
```

Independent variable: grade
Dependent variable: final (the pass depends on exam score)

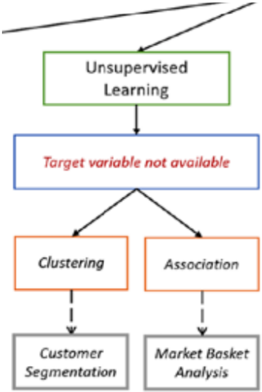
Supervised Learning: Classification

Modeling simple regression using R: `glm(final ~  grade, data = student_scores)`

Model Output: final = -0.4409 + 0.1466*grade
Application For a one-unit increase in exam score, we expect to see about 16% increase in the odds of passing the exam.

`(exp(0.1466)-1)*100` = 16%

Towards data modeling: Unsupervised Learning



- Discuss next week

Towards data modeling - In IST 687

Supervised Learning

- Linear Regression (Week 8)
- Support Vector Machines (Week 9)

Unsupervised Learning

- Association Rule Mining (Week 9)
- Text Mining (Week 9)

Next Week

- **Asynchronous Materials**
  - Week 9: Association Rule Mining and Support Vector Machines
  - Submit HW 8
- **Live Session**
  - Exam review
  - Complete lab 9: Using exploratory analysis and arules

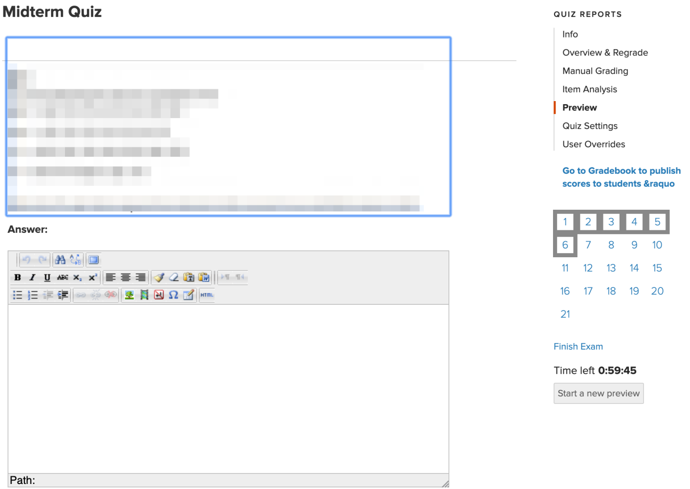
Exam Logistics

Exam Logistics

- **Format**
  - Closed book/notes/R
  - 1 hour time limit (no pausing)
- **Materials covered**: Weeks 1-8
- **Question types**
  - Given code what is the expected output: 2
  - Write code to perform: 10
  - Open-ended questions: 9

Exam Due: Saturday, August 22 9:30 PM EDT

Exam Logistics



Midterm Quiz

Answer:

Path:

QUIZ REPORTS

- Info
- Overview & Regrade
- Manual Grading
- Item Analysis
- **Preview**
- Quiz Settings
- User Overrides

Go to Gradebook to publish scores to students &raquo

1 2 3 4 5
6 7 8 9 10
11 12 13 14 15
16 17 18 19 20
21

Finish Exam

Time left **0:59:45**

Start a new preview

Exam Logistics

Final questions?

Exam Code:**p3nguins***