Codes for STEM

Noushin Nabavi

2020-09-10

Contents

Chapter 1

Coding for STEM

Tools and capabilities of data science is changing everyday!

This is how I understand it today:

Data can: * Describe the current state of an organization or process
* Detec anomalous events
* Diagnose the causes of events and behaviors
* Predict future events

Data Science workflows can be developed for:
* Data collection and management
* Exploration and visualization
* Experimentation and prediction

Applications of data science can include:
* Traditional machine learning: e.g. finding probabilities of events, labeled data, and algorithms
* Deep learning: neurons work together for image and natural language recognition but requires more training data
* Internet of things (IOT): e.g. smart watch algorithms to detect and analyze motion sensors

Data science teams can consist of: * Data engineers: SQL, Java, Scala, Python
* Data analysts: Dashboards, hypothesis tests and visualization using spreadsheets, SQL, BI (Tableau, power BI, looker)
* Machine learning scientists: predictions and extrapolations, classification, etc. and use R or python * Data employees can be isolated, embedded, or hybrid

Data use can come with risks of identification of personal information. Policies for personally identifiable information may need to consider:

* sensitivity and caution
* pseudonymization and anonymization

Preferences can be stated or revealed through the data so questions need to be specific, avoid loaded language, calibrate, require actionable results.

Data storage and retrieval may include: * parallel storage solutions (e.g. cluster or server)
* cloud storage (google, amazon, azure)
* types of data: 1) unstructured (email, text, video, audio, web, and social media = document database); 2) structured = relational databases
* Data querying: NoSQL and SQL

Communication of data can include:
* Dashboards
* Markdowns
* BI tools
* rshiny or d3.js

Team management around data can use: * Trello, slack, rocket chat, or JIRA to communicate due data and priority

A/B Testing: * Control and Variation in samples
* 4 steps in A/B testing: pick metric to track, calculate sample size, run the experiment, and check significance

Machine learning (ML) can be used for time series forecasting (investigate seasonality on any time scale), natural language processing (word count, word embeddings to create features that group similar words), neural networks, deep learning, and AI.
Learning can be classified into: *Supervised*: labels and features/ Model evaluation on test and train data with applications in: * recommendation systems
* subscription predictions
* email subject optimization
Unsupervised: unlabeled data with only features
* clustering

Deep learning and AI requirements: * prediction is more feasible than explanations
* lots of very large amount of training data

Chapter 2

Introduction

Chapter 3

Literature

Chapter 4

Methods

We describe our methods in this chapter.

Chapter 5

Applications

Chapter 6

Final Words