Introduction

An introduction to Machine Learning & Data Visualization

Tara McDarby
April 21st 2020
Machine Learning

Using computer programs such as python to process data

Advantage:
Can process large datasets quickly
Use

Volume/Size

- effect this has on resources
- collections growing at certain levels
- areas of dominance changing overall nature of collection
- new technical tools shaping subject headings and classification (computer generated art)
Python

(object oriented programming)

- abstraction (hiding unnecessary details from the user)
- encapsulation (combining data and methods that work on that data within one unit)
- inheritance (when an already existing class extends its features to a new class).
- polymorphism (when objects of different types can be accessed through the same interface)

https://stackify.com/oop-concept-polymorphism/  

Link: https://www.python.org/
Glossary

python
jupyter
notebook
script
textual data
Libraries

A library in python is a collection of functions and methods that you can ‘import’ into your script directly. This saves you having to write the code.

Numpy - scientific computing

Pandas - data manipulation and analysis

Scikit-learn - machine learning and data mining

NLTK - Language processing
Loading a corpus

In python on your jupyter notebook...

```python
from nltk.corpus import gutenberg
import matplotlib.pyplot as plt
import matplotlib

bible = gutenberg.open('bible-kjv.txt')
bible = bible.readlines()
Bible[:5]
```
Results

'\n',
'The Old Testament of the King James Bible\n',
'\n',
'The First Book of Moses: Called Genesis']
Stopwords

stopwords = nltk.corpus.stopwords.words('english')

words = [word.lower() for word in words if word.lower() not in stopwords()]

c = Counter(words)

c.most_common(10)
Results

[('the', 64014),
 ('and', 51313),
 ('of', 34634),
 ('to', 13567),
 ('that', 12784),
 ('in', 12503),
 ('he', 10261),
 ('shall', 9838),
 ('unto', 8987),
 ('for', 8810)]
Algorithms for text

- Bag of Words Model
- Bag of n-grams Model
- Document similarity
- Topic Models

Advanced Feature Engineering

- Word2Vec Model
- The GloVe Model
- The FastText Model
Text Classification

import pandas as pd
import numpy as np
import re
import nltk
import matplotlib.pyplot as plt
pd.options.display.max_colwidth = 200
%matplotlib inline
corpus = [['The sky is blue and beautiful.',
'Love this blue and beautiful sky!',
'The quick brown fox jumps over the lazy dog',
'I love eggs, ham, sausage and bacon',
'The brown fox is quick and the blue dog is lazy',
'The sky is very blue and the sky is beautiful today',
'The dog is lazy but the fox is quick']
]
Labelling

labels = ['weather', 'weather', 'animals', 'food', 'animals', 'weather', 'animals']

corpus = np.array(corpus)

corpus_df = pd.DataFrame({'Document': corpus, 'Category': labels})

corpus_df = corpus_df[['Document', 'Category']]

corpus_df
<table>
<thead>
<tr>
<th>Document</th>
<th>Category</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>The sky is blue and beautiful.</td>
</tr>
<tr>
<td>1</td>
<td>Love this blue and beautiful sky!</td>
</tr>
<tr>
<td>2</td>
<td>The quick brown fox jumps over the lazy dog</td>
</tr>
<tr>
<td>3</td>
<td>I love eggs, ham, sausage and bacon</td>
</tr>
<tr>
<td>4</td>
<td>The brown fox is quick and the blue dog is lazy</td>
</tr>
<tr>
<td>5</td>
<td>The sky is very blue and the sky is beautiful today</td>
</tr>
<tr>
<td>6</td>
<td>The dog is lazy but the fox is quick</td>
</tr>
</tbody>
</table>
Data Visualization
What kinds of Data Visualizations are there?

Data visualizations can be maps, plots, diagrams and graphs. Instead of reading densely written reports, we can use visualizations to see patterns or trends in data.
Selecting a visualization type

What do you want to find?

https://datavizcatalogue.com/

The data viz catalogue is a great interactive resource that can be used to discover which type of visualization suits which function best.
# A Periodic Table of Visualization Methods

## Data Visualization
- Visual representations of quantitative data in schematic form (either with or without axes).

## Strategy Visualization
- The systematic use of complementary visual representations in the analysis, development, formulation, communication, and implementation of strategies in organizations.

## Information Visualization
- The use of interactive visual representations of data to simplify cognition. This means that the data is transformed into an image, it is mapped to screen space. The image can be changed by users as they proceed working with it.

## Metaphor Visualization
- Visual Metaphors position information graphically to organize and structure information. They also convey an insight about the represented information through the key characteristics of the metaphor that is employed.

## Concept Visualization
- Methods to elaborate (mostly) qualitative concepts, ideas, plans, and analyses.

## Compound Visualization
- The complementary use of different graphic representation forms in one single schema or frame.

---

[https://www.visual-literacy.org/periodic_table/periodic_table.html](https://www.visual-literacy.org/periodic_table/periodic_table.html)
Functions

Comparisons  Proportions  Relationships

Part-to-a-whole  Processes & methods

Distribution  How things work  Range

Patterns  Locations  Concepts  Analysing Text

Movement or flow  Data over time
Creating visualizations

select  process  mine  visualize
We select the data then process it.

We identify what we want to do with it - group the content by theme or topic, analyse the content for features of language for example. Once we know what we are looking for, we can select a classifier to classify the data accordingly.

We did this at the beginning when we grouped our sentences together into topics (food, animals, weather)
Mining

Mining methods place the data into a context that enables it to be visualized.

Methods include sequences analysis, classifications, path analysis and clustering.
Clustering algorithms

- **Flat clustering** (creates a set of clusters without any explicit structure that would relate clusters to each other; It’s also called exclusive clustering)

- **Hierarchical clustering** (Creates a hierarchy of clusters)

- **Hard clustering** (Assigns each document/object as a member of exactly one cluster)

- **Soft clustering** (Distribute the document/object over all clusters)

https://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm
Algorithms

Agglomerative (Hierarchical clustering)

K-Means (Flat clustering, Hard clustering)

EM Algorithm (Flat clustering, Soft clustering)

https://www.codeproject.com/Articles/439890/Text-Documents-Clustering-using-K-Means-Algorithm
Clustering (unsupervised)

finding a *structure* in a collection of unlabeled data. The aim is to organize the data into groups based on common features or similarities.
Scatterplot

```python
import seaborn as sns
sns.set()

# Load the example planets dataset
planets = sns.load_dataset("planets")

cmap = sns.cubehelix_palette(rot=-.2, as_cmap=True)
ax = sns.scatterplot(x="distance", y="orbital_period",
                     hue="year", size="mass",
                     palette=cmap, sizes=(10, 200),
                     data=planets)
```
Scatterplot
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

sns.set(style="dark")
rs = np.random.RandomState(50)

# Set up the matplotlib figure
f, axes = plt.subplots(3, 3, figsize=(9, 9), sharex=True, sharey=True)

# Rotate the starting point around the cubehelix hue circle
for ax, s in zip(axes.flat, np.linspace(0, 3, 10)): 
# Create a cubehelix colormap to use with kdeplot
    cmap = sns.cubehelix_palette(start=s, light=1, as_cmap=True)

    # Generate and plot a random bivariate dataset
    x, y = rs.randn(2, 50)

    sns.kdeplot(x, y, cmap=cmap, shade=True, cut=5, ax=ax)
    ax.set(xlim=(-3, 3), ylim=(-3, 3))
    f.tight_layout()
## Dataset  BLL Theses

[https://bl.iro.bl.uk/work/86c21604-10d3-4367-a131-fb19a259ce07](https://bl.iro.bl.uk/work/86c21604-10d3-4367-a131-fb19a259ce07)

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Title</td>
<td>Author</td>
<td>Institution</td>
<td></td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Computation and measurement of turbulent flow through idealized turbins</td>
<td>Loizou, Panos A.</td>
<td>University of Manchester</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Prolactin and growth hormone secretion in normal, hyperprolactinaemic subjects</td>
<td>Prescott, R. W. G.</td>
<td>University of Newcastle upon Tyne</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Influence of strain fields on flame propagation</td>
<td>Mendes-Lopes, J. M. C.</td>
<td>University of Cambridge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5</td>
<td>Connectivity, flow and transport in network models of fractured media</td>
<td>Robinson, Peter Clive</td>
<td>University of Oxford</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>The theory and implementation of a high quality pulse width modulated voltage regulator</td>
<td>Lower, K. N.</td>
<td>University of Bristol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Separation bubbles at high Reynolds number: measurement and computation</td>
<td>Davenport, W. J.</td>
<td>University of Cambridge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>A unified approach to the identification of dynamic behaviour using the fuzzy systems</td>
<td>Brown, T. A.</td>
<td>University of Bristol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>PWM strategies for microprocessor control of variable speed drives</td>
<td>Midoun, A.</td>
<td>University of Bristol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>10</td>
<td>Theoretical investigations of stress concentrations in carbon fibre reinforced composites</td>
<td>Wu, C. M. L.</td>
<td>University of Bristol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>11</td>
<td>Speed-changing of induction motors by phase modulation</td>
<td>Ismail, K. S.</td>
<td>University of Bristol</td>
<td></td>
<td></td>
</tr>
<tr>
<td>12</td>
<td>The immune response of the bovine udder to Streptococcus agalactiae infection</td>
<td>MacKie, D. P.</td>
<td>Queen’s University Belfast</td>
<td></td>
<td></td>
</tr>
<tr>
<td>13</td>
<td>Metabolic effects of Bordetella pertussis</td>
<td>Sidey, Fiona M.</td>
<td>University of Strathclyde</td>
<td></td>
<td></td>
</tr>
<tr>
<td>14</td>
<td>Executing behavioural definitions in Higher Order Logic</td>
<td>Camilleri, Albert John</td>
<td>University of Cambridge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>15</td>
<td>A methodology for automated design of computer instruction sets</td>
<td>Bennett, J. P.</td>
<td>University of Cambridge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>16</td>
<td>Reasoning about the function and timing of integrated circuits with Prolog</td>
<td>Leeser, Miriam Ellen</td>
<td>University of Cambridge</td>
<td></td>
<td></td>
</tr>
<tr>
<td>17</td>
<td>Computer modelling of flow with a free surface</td>
<td>Xin, Liu</td>
<td>Imperial College London</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
https://textexture.com
Regression

The term regression is used when you try to find the relationship between variables.

In Machine Learning, and in statistical modeling, that relationship is used to predict the outcome of future events.

Linear Regression  https://www.w3schools.com/python/python_ml_polynomial_regression.asp

Linear regression uses the relationship between the data-points to draw a straight line through all them. This line can be used to predict future values.
Linear Regression

https://www.w3schools.com/python/python_ml_linear_regression.asp
Polynomial Regression

If your data points clearly will not fit a linear regression (a straight line through all data points), it might be ideal for polynomial regression.

Polynomial regression, like linear regression, uses the relationship between the variables x and y to find the best way to draw a line through the data points.

https://www.w3schools.com/python/python_ml_polynomial_regression.asp
Links & Tools

Machine Learning
- www.python.org

Visualization
- https://seaborn.pydata.org/

Excellent resources to explore
https://github.com/brianspiering/awesome-dl4nlp
https://datavizcatalogue.com/
www.tableau.com
https://densitydesign.org/
https://www.flickr.com/photos/densitydesign/sets/72157628222445801/with/6431913399/
https://www.flickr.com/photos/densitydesign/sets/72157624141332939/
https://densitydesign.org/research/minerva/

Stack Overflow
https://stackoverflow.com/questions/tagged/python

Tableau
https://www.tableau.com/learn/articles/data-visualization
Links

www.iskouk.org
https://twitter.com/ISKOUK
https://www.linkedin.com/groups/2079995/