**Data Analytics II:**

**Assignment #1**

Assigned readings

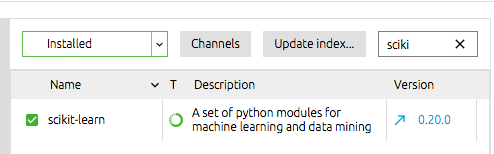
sk-learn tutorial at [http://scikit-learn.org/stable/tutorial/basic/tutorial.html#](http://scikit-learn.org/stable/tutorial/basic/tutorial.html)

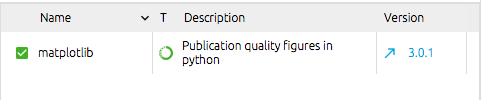
SKLearn Feature-extraction documentation: <http://scikit-learn.org/stable/modules/feature_extraction.html>

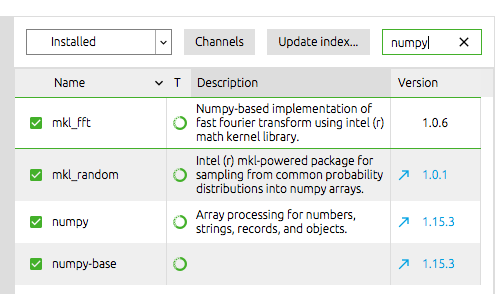
Pandas tutorial at <http://pandas.pydata.org/pandas-docs/version/0.18.1/tutorials.html>

NLTK book, chapter 3: <http://www.nltk.org/book/ch03.html>

Prior to completing this homework, download and install sklearn, matplotlib, & numpy to your machine. Provide screenshots showing successful imports of these packages.







1. (10 points) Provide a short definition of the following terms **IN YOUR OWN WORDS (no copy-paste)**. For each term, if relevant, indicate whether it is a supervised learning task or an unsupervised learning task. **Provide an example of each of these tasks:**
   1. Classification

Classification is when you already know the categories that you have and you are trying to see if something is in that category or not. Classification is supervised learning task. I.e., scam detection

* 1. Clustering

Clustering is when you don’t know the categories in the data and you are trying to let the data show the possible categories/similarities itself. Clustering is unsupervised learning task. I.e., the most popular topics in tweets in 2018

* 1. Density estimation

Density estimation is to construct a model or estimate based on the observed data of the unobservable probability density function. Density estimation is unsupervised task. I.e., density estimates for “glucose concentration” on the presence/absence/not conditional on diabetes

* 1. Regression

Regression is for finding the relationships between variables and using the relationship for prediction. Regression is supervised learning task. I.e., regression like the gas fee and the gas used

* 1. Samples

Samples indicate the data selected objectively from the data pool. I.e., the high schools selected randomly in the US

* 1. Features

Features indicate the values/properties/characteristics observed, sometimes might also be called variables. I.e., the height of the students in the high school randomly selected in the US

* 1. Multivariate data

Multivariate data indicate the data that is observed with multiple features/characteristics simultaneously, and the data will be analyzed with all/many of its features simultaneously observed at the same time. I.e., the height and weight of the students in the high school randomly selected in the US

* 1. Training set

The part of the data set apart from the whole data set that is used for the training of the model. Normally 80% but can vary. I.e., the temperature data in January from 2010-2015

* 1. Test set

The part of the data set apart from the whole data set that is used for the testing of the prediction accuracy of the model. It’s known data, and the main purpose is to score the accuracy of the model when being used for prediction. I.e., the temperature data in January from 2016-2017

* 1. Holdout set

Holdout set is also a type of testing data, but it is used for the final estimate of the model’s prediction accuracy/performance after the model has been trained and validated over and over again. i.e., the temperature data in Jan 2018

1. (10 points) Implement the tutorial found at <http://scikit-learn.org/stable/tutorial/basic/tutorial.html#introduction>

For each line of code, write a short sentence describing the meaning of that code (**in your own words**). Write a short paragraph describing the overall function of the code. Treat the Support Vector Classifier as a “black box” – don’t worry about how it works (that’s coming later).

I did this part in Jupyter notebook

Implement the tutorial here: <http://scikit-learn.org/stable/auto_examples/classification/plot_digits_classification.html#example-classification-plot-digits-classification-py>

For each line of code, write a short sentence describing the meaning of that code (**in your own words**). Write a short paragraph describing the overall function of the code. As above, treat the Support Vector Classifier as a “black box” – don’t worry about how it works (that’s coming later).

I did this part in Jupyter notebook

1. (10 points) Modify the digits tutorial for the Iris dataset (use 125 randomly selected samples as training data and the last 25 samples as test data).

I did this part in Jupyter notebook

1. (10 points) Load the “Boston Housing” dataset from SKLearn. Save it to a Pickle (.pkl) file and submit the file with your assignment.
   1. How many samples does this dataset have?

506 samples

* 1. How many features does this dataset have?

13 features

* 1. For each feature, indicate whether it is categorical or continuous. If categorical, how many levels does it have?

CRIM: continuous

ZN: continuous

INDUS: continuous

CHAS: categorical, 2 levels- 0 and/or 1

NOX: continuous

RM: continuous

AGE: continuous

DIS: continuous

RAD: categorical, 9 levels

TAX: continuous

PTRATIO: continuous

B: continuous

LSTAT: continuous

MEDV: continuous

* 1. Justify, in writing, your answers to part c. If a feature can be both categorical or continuous, present an argument for why your assignment is correct.

For CHAS (Charles River dummy variable ), as the description shows, it has a value of 0 if tract bounds river, and 0 otherwise.

For RAD (index of accessibility to radial highways), first of all, it only has countable (more importantly, few possible) values. Also, the possible values are all integers

* 1. For each feature, calculate its mean and median (if continuous) and its mode (if categorical)

Feature name Mean Median Mode

CRIM 3.6135 0.25651 N/A

ZN 11.3636 0.0 N/A

INDUS 11.1368 9.69 N/A

CHAS N/A N/A 0

NOX 0.5547 0.538 N/A

RM 6.2846 6.2085 N/A

AGE 68.5749 77.50 N/A

DIS 3.7950 3.20745 N/A

RAD N/A N/A 24

TAX 408.2372 330.00 N/A

PTRATIO 18.4555 19.05 N/A

B 356.6740 391.44 N/A

LSTAT 12.6531 11.36 N/A

1. (10 points) Load the airports only (airports.dat) dataset from <http://openflights.org/data.html> into SKLearn using the dictionary vectorizer function.
   1. First, use the read\_csv function in pandas to load the data into a dataframe. Next, use the to\_dict function to save the data as a dictionary (HINT: transpose the data and make sure to only keep the values)
   2. Use the sklearn DictVectorizer to load the data
   3. Save the dataset to a Pickle (.pkl) file and submit the file and your code with your assignment.
   4. How many samples does this dataset have?

7542 samples

* 1. How many features does this dataset have?

14

* 1. For each feature, indicate whether it is categorical or continuous. If categorical, how many levels does it have?

Feature1-airport id: Categorical, it has 7542 levels.

Feature2- name: Categorical, it has 7503 levels

Feature3-city: Categorical, it has 6822 levels

Feature4-country: Categorical, it has 237 levels

Feature5-IATA: Categorical, it has 5958 levels

Feature6-ICAO: Categorical, it has 7542 levels

Feature7-latitude: Continuous

Feature8-longitude: Continuous

Feature9-altitude: Continuous

Feature10-timezone: categorical, it has 40 levels

Feature11-DST: Categorical, it has 8 levels

Feature12- TZ database time zone: Categorical, it has 308 levels

Feature13- type: Categorical, it has 1 level

Feature14- source: Categorical, it has 1 level

* 1. Justify, in writing, your answers to part iii

Airport-id: you cannot compare the number, the number only to distinguish different airport

Name: the name of the airport. Categorical variable

City: the city of the airport. Categorical variable

Country: the country of the airport, categorical variable.

IATA: 3-letter alphabetic IATA code, categorical variable

ICAO: 4-letter alphabetic ICAO code, categorical variable

Latitude: number of the latitude of the airport, you can compare the number and say if the latitude number is bigger, the airport is at a higher latitude. Continuous variable in this case

Longitude: number of the longitude of the airport, you can compare the number and say if the longitude number is bigger, the airport is at a higher longitude. Continuous variable in this case

Altitude: number of the altitude of the airport, you can compare the number and say if the altitude number is bigger, the airport is at a higher altitude. Continuous variable in this case

Timezone: represent hours offset from UTC, has fractional numbers. But is string. Categorical variable.

DST: alphabetical values, categorical variable

Tz database time zone: alphabetical values, categorical variable

Type: alphabetical values, categorical variable

Source: alphabetical values, categorical variable

* 1. For each feature, calculate its mean and median (if continuous) and its mode (if categorical)

Airport\_id: mode=2

Name: mode= Capital City Airport

City: mode= London

Country: mode= United States

IATA: mode=\N

ICAO: mode= 03N

Latitude: mean=25.84246, median=34.1343498

Longitude: mean=-2.435413 median=5.692499

Altitude: mean=1004.147574 median=352.0

Timezone: mode=1

DST: mode=U

Tz database time zone: mode= \N

Type: mode=airport

Source: mode=ourAirports

1. (10 points) Load a non-JSON dataset of your choice from data.gov, <http://catalog.data.gov/dataset> into SKLearn. Save the dataset to a Pickle (.pkl) file and submit the file and your code with your assignment.
   1. Spend some time exploring this dataset, and brainstorm a question that you can use it to answer.

How the number of death affects age-adjusted death rate

* 1. First, use the read\_csv function in pandas to load the data into a dataframe. Next, use the to\_dict function to save the data as a dictionary (HINT: transpose the data and make sure to only keep the values)

I did this part in Jupyter notebook

* 1. Use the sklearn DictVectorizer to load the data

I did this part in Jupyter notebook

* 1. Save the dataset to a Pickle (.pkl) file and submit the file and your code with your assignment.

I did this part in Jupyter notebook

* 1. How many samples does this dataset have?

10296 samples

* 1. How many features does this dataset have?

6 features

* 1. For each feature, indicate whether it is categorical or continuous. If categorical, how many levels does it have?

Feature1- Year, categorical, 18 levels

Feature2- 113 Cause Name, categorical, 11 levels

Feature3-Cause Name, categorical, 11 levels

Feature4 – State, categorical, 52 levels

Feature5- Deaths, continuous, 5741 levels

Feature6- Age-adjusted Death Rate, continuous, 2419 levels

* 1. Justify, in writing, you answers to part iii

Feature1- Year. Comparison of the numerical greatness of year isn’t really meaningful numerically, therefore year should be considered categorical here.

Feature2- 113 Cause Name. Textual data, categorical

Feature3- Cause Name. Textual data, categorical

Feature4- State. Textual data, categorical

Feature5- deaths. Numerical data, also comparison makes numerical point, therefore should be considered continuous data.

Feature6- Age-adjusted Death Rate, comparison makes numerical sense, and the statistical operations like median/mean also makes sense. Therefore it should be considered continuous variable.

* 1. For each feature, calculate its mean and median (if continuous) and its mode (if categorical)

Year- mode :1999

113 Cause Name- mode: Accidents (unintentional injuries) (V01-X59,Y85-Y86)

Cause Name- mode: All causes

State: mode: Alabama

Deaths: mean: 15326.6767, median: 1704.5

Age-adjusted Death Rate: mean: 128.0264, median: 35.8

* 1. Propose a plan to use this dataset to answer your question in part a.

Build a linear regression model with age-adjusted death rate and rates

1. (10 points) Choose a (different) JSON dataset for #6.
   1. Spend some time exploring this dataset, and brainstorm a question that you can use it to answer.

If the headline sarcasm score could be predicted with model with headlines as input

* 1. First, use the read\_csv function in pandas to load the data into a dataframe. Next, use the to\_dict function to save the data as a dictionary (HINT: transpose the data and make sure to only keep the values)

I did this part in Jupyter notebook

* 1. Use the sklearn DictVectorizer to load the data

I did this part in Jupyter notebook

* 1. Save the dataset to a Pickle (.pkl) file and submit the file and your code with your assignment.

I did this part in Jupyter notebook

* 1. How many samples does this dataset have?

26709 samples

* 1. How many features does this dataset have?

3 features

* 1. For each feature, indicate whether it is categorical or continuous. If categorical, how many levels does it have?

Feature1- article link, categorical, 26708 levels

Feature2- headline, categorical, 26602 levels

Feature3-is\_sarcastic, categorical, 2 levels

* 1. Justify, in writing, you answers to part iii

Feature1- article link, categorical, because it’s the article link

Feature2- headline, categorical, because it’s textual data and the headline of the articles

Feature3-is\_sarcastic, categorical, because the number only indicates if the article is sarcastic or not

* 1. For each feature, calculate its mean and median (if continuous) and its mode (if categorical)

Article link: mode: [https://www.huffingtonpost.comhttp://nymag.com...](https://www.huffingtonpost.comhttp:/nymag.com...)

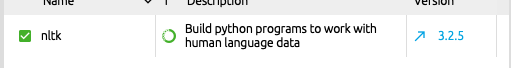
Headline, mode: sunday roundup

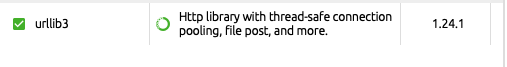
Is\_sarcastic: mode: 0

* 1. Propose a plan to use this dataset to answer your question in part a.

Build a logistic regression model and use the first 80% headline and is\_sarcastic results to train the model, and use the 20% to validate the model

For the following problems, download and install nltk, urllib2, bs4, feedparser, pypdf, and tweepy to your machine. Provide screenshots showing successful imports of these packages.











1. (20 points) Download text from each of the websites listed below. For each data source,
   1. list the top most frequently occurring bigrams (HINT: use the nltk collocations() function with 30 as an input).
   2. Split the text into documents
   3. Generate two term-document matrices from this dataset (one where each unigram is a token, and one where bigrams can also be tokens). Import the matrix into sklearn and save it as a PKL file. For each text file, indicate how many terms, documents, and unigram tokens are in the corpus
      1. George Washington’s Masonic Correspondence (UTF8) from Project Gutenberg. Treat each paragraph as one “document”

Term: 2

Documents: 1042

Unigram tokens:34373

* + 1. FDA Circulatory System’s Devices Panel Advisory Panel Meeting of February 23, 2017: <https://www.fda.gov/downloads/AdvisoryCommittees/CommitteesMeetingMaterials/MedicalDevices/MedicalDevicesAdvisoryCommittee/CirculatorySystemDevicesPanel/UCM547942.pdf>

Hint: use PyPDF2. One document is the end of a speaker’s statement??

Term: 2

Documents: 941

Unigram tokens:84693

* + 1. Nate Silver’s Sports RSS feed <http://fivethirtyeight.com/sports/feed/> One document is an article

Term: 2

Documents: 21

Unigram tokens:93125

* + 1. NASA’s Systems Engineering Handbook <http://ntrs.nasa.gov/archive/nasa/casi.ntrs.nasa.gov/20080008301.pdf> -- hint: use pypdf. One document is denoted by a carriage return

Term: 2

Documents: 2267

Unigram tokens:219442

* + 1. A sample of 10,000 tweets on a search string of your choice – hint: use tweepy or SFM. One document is a tweet.

Term: 2

Documents: 15622

Unigram tokens:382505

* + - 1. Extra credit: Redo your analysis on a sample of tweets that does not include retweets and tweets with URLs
    1. Download the text of the top 100 websites obtained by using the search string “data analytics” using a search engine API, such as Google. One document is a webpage.

Term: 2

Documents: 3813

Unigram tokens: 1981498

I did this part in Jupyter notebook

Note: this assignment will be much easier if you write a set of general functions that can take in any parsed input, instead of rewriting your code for each of i-vi.

1. (10 points) FINAL PROJECT Module 1:
   1. Identify a final project topic (not binding).

Analyzing the report of drug overdose death

* 1. Identify a data source for your final project topic

Data.gov, government database

* 1. Import data from this data source into a format readable by SKLearn.

I did this part in Jupyter notebook

* 1. Submit the name of the data source and a PKL file or executable Python script that will provide access to the data source.

I did this part in Jupyter notebook