Meenu Ravi
May 7 2021
Project Proposal

What dataset do you want to use?
https://www.kaggle.com/ahmaurya/iclr2017reviews
ICLR (International Conference on Learning Representations) is a machine learning conference where they accept various papers for submission. There is data regarding the papers that were accepted and rejected for this conference for the year 2017 and information about the papers such as the authors and the title. This dataset is available on kaggle.

- **How many samples are in the data?**
  There are 490 samples in the data set

- **How many Mb or Gb?**
  About 1 mb

- **What are the features?**
  The features present in this dataset are:
    - Abstract
    - Authorids
    - Authors
    - Conflicts
    - Keywords
    - Paper_id
    - Paperhash
    - Title
    - Excerpt
    - decision

  However, I would like to add more features when data cleaning. Below are the features I would like to add
    - Number of authors (count authors)
    - Length of title (count number of words in title)
    - Length of abstract (count words in abstract)
    - Ivy affiliation (check if conflict contains any of the ivy league schools)
    - keywords_ML (check if the keywords contain machine learning keywords)
    - Number of keywords (count number of keywords)
    - Title_ML (check if the title contains any machine learning keywords)
    - Acceptance type (whether the acceptance was through oral, or poster or workshop- here I can just split the acceptance columns and pull this word out)
    - Number of conflicts
    - Paper word count
    - Company_affliation (whether the affiliations are to facebook or google companies)

- **Is it labelled, partially labelled, unlabelled? If there is a label, what is it?**

It is labelled as accepted or rejected

- **What's the data format?**
  It is a csv file

What problem do you want to solve with this dataset?
I would like to build a model to classify whether a conference paper can be accepted or rejected to the machine learning conference.

What methods that we have studied so far are applicable to your problem?
First, I can perform PCA to reduce dimensionality.
Then, some models I can consider are:

- Multivariate logistic regression: since the dependent variable is binary (Accept/Reject)
- SVM: I can try using different kernels and hypertuning parameters
- Random forest classification
- Neural networks: binary dependent variable with one output layer

I can use these topics that we learning this class to see which is the best and most accurate model for classifying and predicting whether a paper will be accepted or rejected into the conference

Discuss your proposals in your group and rank them. Provide some justification for your ranking.

Proposal	Ranking	Justification
Predicting Stock Market Price	2	We thought this might be nice to observe and there was lot of data available for this
Predicting Data Science occupation	3	We thought this would be interesting but there weren't as much feature to look at
Predicting Conference paper acceptance	1	This seemed interesting for the group. It was something we thought was new and might be interesting to look at