EE621 Adaptation and Learning

Suggestions for Projects

Instructor: A. H. Sayed Teaching Assistant: S. Vlaski

February 26, 2018

The purpose of the project in EE621 is to allow students to apply concepts and algorithms learned in class on real world-data sets and experience the challenges that arise when departing from the well-defined models and simplifying assumptions that are typically applied during the theoretical development and analysis of learning algorithms.

1 Data Set Suggestions

We list here a few suggestions for data sets with ideas for interesting questions to ask. The scope of the project is deliberately left wide open, allowing students to explore data sets and questions while using the tools learned in EE621.

1.1 Soccer

`https://www.kaggle.com/hugomathien/soccer`
A large collection of data on soccer matches, results, and meta data such as player skills and betting odds. The overarching task here is to build a model that takes into account the available information and allows for the prediction of game results.

- Is it possible to predict who will win a soccer match? How well?
- Assign probabilities to each outcome and compare with betting odds as well as the actual outcome.
- It is known that home-advantage is a strong indicator of outcome. What are others?
- Are there team-specific indicators of outcome, such as key players?
- Are there league-specific indicators of outcome?

1.2 Spam messages

`https://www.kaggle.com/uciml/sms-spam-collection-dataset`
A collection of spam messages. The broad objective is to train a classifier to identify whether messages are spam or not.

- How can a message be converted into a feature vector for classification?
- Design a classifier for the decision spam/no spam.
- Assign a probability of a message being spam or not.
- Is it possible to tune the sensitivity of the classifier? What are the trade-offs?
- Allow for user input in the form of a manual correction and design a mechanism for an online update.

1.3 Traffic

`https://www.kaggle.com/daveianhickey/2000-16-traffic-flow-england-scotland-wales`
This data set contains traffic and accident information from the United Kingdom.

- How can such a big data set be visualized?
- Which parts of the UK are prone to accidents? How has this changed over time?
- Why are these trends occurring?

1.4 Financial Distress

`https://www.kaggle.com/shebrahimi/financial-distress`
Financial and non-financial data that can be used to predict whether a company will end up in financial trouble. Banks frequently perform this type of analysis to determine whether to extend a loan to applicants or not.

- Train a classifier to predict whether a company will end up in financial distress or not.
- Assign a probability to the event of financial distress. How would this information inform interest rates and how would these be set as a function of the probability of distress?
- What are key indicators of potential for financial distress?

1.5 Stock Market Data

`https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs`
A large collection of stock market information for publicly traded companies.

- Is it possible to predict the movement in the stock market?
- Are there correlations between the performance of certain companies? Is it possible to infer competitive/symbiotoic relationships from these correlations?

1.6 News

`https://www.kaggle.com/aaron7sun/stocknews`
A collection of news which can be correlated to stock data, to investigate whether this additional information can be leveraged to improve prediction performance.

1.7 Cryptocurrency

`https://www.kaggle.com/philmohun/cryptocurrency-financial-data`
Prices of various crypto currencies. Similar questions as in the stock data arise.

- Is there any structure to this modern phenomenon?
- Which currencies are volatile? Which are correlated? Which are in competition?
- How do cryptocurrency rates correlate with traditional investment avenues?
- How do they correlate to current news?

1.8 Netflix Data Set

`https://www.kaggle.com/netflix-inc/netflix-prize-data`
This data set contains movie ratings by user.

- Is it possible to predict whether a user will like a certain movie they have not seen before?
- How can we decide if two users have similar taste? How can we decide if two movies are similar?

1.9 Fake News

`https://www.kaggle.com/mrisdal/fake-news`
A collection of news articles, which have been identified as fake.

- Is it possible for a machine to decide whether an article is fake or not?
- How would one construct a feature vector for classification?
- What are indicators of a fake news article?

2 Students' Suggestions

Students are welcome to propose other, publicly available, data sets to work on, so long as the data sets allow for the application of algorithms and concepts learned in EE621. If this is what you want to do, please send a short email to `stefan.vlaski@epfl.ch` describing:

- What kind of data set is this and where was it obtained from?
- What information is available in the data?
- What types of questions do you wish to investigate?
- What types of algorithms and concepts are you planning on using?