Introduction to Statistical Machine Learning CSC/DSCC 265/465

Kaggle Challenge II

Cantay Caliskan

Kaggle Challenge II

Kaggle Challenge II

- A prediction challenge
- You will be asked to predict the ***winner ratio*** for a large set of political contributors from the US
- **Input (X)** – all of the variables in the training dataset (and more…)
- **Output (Y)** – ***winner ratio***

Information about the dataset

Aggregated Campaign Contributor Data:

- training_data.csv
- test_data.csv

Bipartite Networks Between Contributors and Candidates:

- all_candidates_state_bipartite_weighted_network.csv
- federal_contributor_top100_contributors_network.csv
- state_contributor_top100_contributors_network.csv
- winning_candidates_state_bipartite_weighted_network.csv

Sample Solution File:

- sample_solutions.csv

Training and test data contain information on the campaign behavior of contributors.

Networks show the connections between contributors (themselves) and contributors to candidates.

For more detailed information, please refer to instructions.

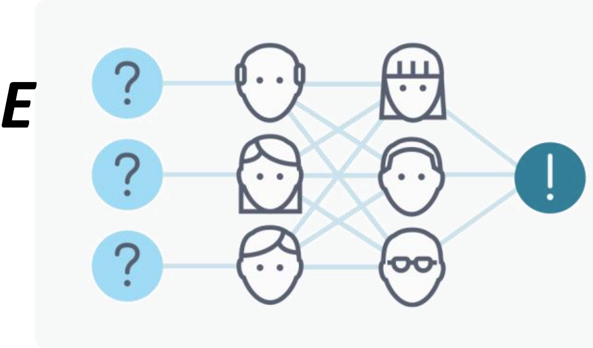
UNIVERSITY *of* ROCHESTER

Tasks

Tasks

- Slightly less amount of work needed

- You will work on one (1) task:
  - ***Kaggle Competition (100 points)***
    - You will create a model that provides the lowest ***MSE*** value for predicting correct **‘winner ratio’** by using the contributor information and *lobbying networks* formed by contributors



Prediction: Steps

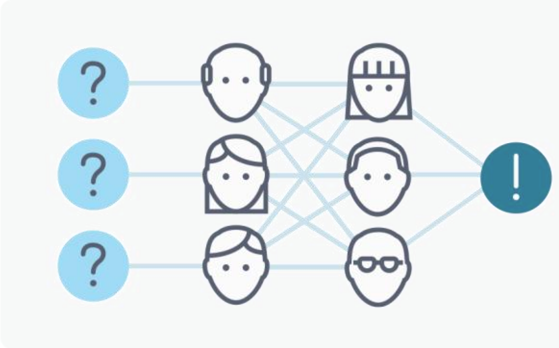
1) Develop a prediction model using the **`training`** dataset

2) Using the model, classify the observations in the **`test`** dataset

3) Use the **`sample submission`** file (a smaller version of the **`test`** dataset) to submit your solutions [solutions submitted according to the Index variable]

4) If not happy with the results, repeat the Steps 1), 2), and 3)

kaggle



Online Competition

- Online competition you can enter on ***Kaggle***:
  - https://www.kaggle.com/t/8e30548690334c5095c8f0cf2970d891
  - Goal: Develop a prediction model that predicts the observations with the lowest ***MSE*** possible
  - **No model restrictions!**
  - You can:
    - Use any prediction algorithm that you think will give the highest accuracy
    - Perform any type of feature engineering
    - Perform weighting, dimensionality reduction etc.
    - Use any external dataset to enrich your training and test datasets
    - Note: You can use any external dataset.

Important:

- Use **training_data.csv** to *train your model*
- Use **sample_solutions.csv** to submit your answers
- You can send up to 10 submissions every day (competition is currently open!)
- Provide the *MSE* score in your code

Online Competition: Further Do’s and Don’ts

- Code:
  - **You cannot post your solution / code online.**
  - You can use `Python` (only)
  - Your code should be ***executable***, i.e.:
    - We should be able to run your code by running the cells ***consecutively***
    - We should also be able to run your code on a ***laptop*** (for instance, a new MacBook Pro) in a reasonable amount of time (in max. a few hours)
    - We should be able to ***understand*** what your code is doing. So, please make sure that:
      - you write **a lot of comments** describing your code
      - you only include the code that works
      - you only include your best solution
      - you name your variables mutually intelligibly (i.e. **case_data**, not **td123** etc.)
- Model:
  - Your model must give a number as the prediction result

Lab Report

- **No lab report needed!**

Deliverables

Deliverables

- Your code in ***.ipynb*** format
  - Add a lot of comments to your code!
- Your ranking in ***Kaggle*** system
- Submit the ***code*** through *BlackBoard*

Grading

Kaggle Competition: Grading

- You will be graded based on the following criteria:
  - **Code**
    - Cleanliness/understandability (i), executability (ii), format (iii)
  - **Ranking**
    - Ranking in the **`Kaggle`** competition

More about Grading

- Other important information about Kaggle competition:
  - The lowest grade you can get from the **ranking** component will be **60/100**.
  - The highest ranked project will get **100/100** for the **ranking** component.
  - However:
    - If your accuracy is close to the benchmark reported in the guidelines, your grade may be lower (and it may be zero, as well).

Deadlines

- Please submit your code, solution submission, and report by:
  - **Deadline: Sunday, April 27, 11:59 PM**
  - **You *must* send everything by the deadline.**
  - **Unfortunately, no late submission is possible for this challenge.**

And one last reminder…

- Let's say you have achieved a really good (or maybe a really bad) MSE and you are done with model training:
  - **Please do not post the solutions online!**
  - **Or, simply said, please do not post any related code online ☺**