ADA Final Exam
Deadline
Friday, February 19, 2018, at 7:15 PM

Important Notes

Make sure you upload your iPython Notebook with this form at the end of the exam, with all the cells already evaluated.
Don't forget to add a textual description of your thought process, the assumptions you made, and your results!
Please write all your comments in English, and use meaningful variable names in your code.
As we have seen during the semester, data science is all about multiple iterations on the same dataset. Do not obsess over small details in the beginning, and try to complete as many tasks as possible during the first 2 hours. Then, go back to the obtained results, write meaningful comments, and debug your code if you have found any glaring mistake.
Remember, this is not a homework assignment -- no teamwork allowed!

Design the Pokedex 2.0

Pokémon are small creatures that fight in competitions. All Pokémon have different numerical characteristics (attack, defense, etc.) and belong to one or two “classes” (water, fire, etc.).

You just received a request from Professor Oak, asking you to update the software on his devices. He is the inventor of Pokedex, a useful portable device that keeps information about all the Pokémon available. Your task is to do a preliminary data exploration and to design a model to predict the outcome of one battle.

You dump the memory of one Pokedex, and you get these datasets to start your analysis:

Datasets description

pokemon.csv: It represents the features of the Pokémon

- #: Numeric - ID
- HP: Numeric - Health Points
- Attack: Numeric - Regular Attack
- Defense: Numeric - Regular Defense
- Sp. Atk: Numeric - Special Attack
- Sp. Def: Numeric - Special Defense
- Speed: Numeric - Moving Speed
- Legendary: Boolean - Rare Pokémon
- Type 1: Categorical - Pokémon Class
- Type 2: Categorical - Pokémon Class

Please note that a Pokémon can have 1 or 2 classes and both with the same importance.

combats.csv: Each row represents the outcome of one battle.

- First_pokemon: Numeric - ID (match with #)
- Second_pokemon: Numeric - ID (match with #)
- Winner: Numeric - ID of the winner

Question 1: Exploring the data

In this task, you have to explore the data to address the curiosities of the demanding Professor Oak.

- Explore the data and report some descriptive statistics, this will be useful later in your analysis (i.e., df.describe(), number of classes, attack and defense distributions).

- How does it look the Regular Attack vs. Regular Defense distribution? Plot the distribution **(1)** and provide the names of the top 3 most imbalanced Pokémon with highest **(2)** and lowest **(3)** Attack-over-Defense ratio. What are the names of the top 10 Pokémon with the highest number of victories **(4)**?

- Your colleague is convinced that the Grass Pokémon has a stronger Regular Attack than Rock Pokémon. Check if he is right and if not, convince him with reasonable arguments **(5)**. You can assume that the distributions are normal.

- Professor Oak designed a backdoor in every Pokedex of the world, and he revealed you that he is secretly dumping the memory of all the of devices online. He expects to have the data of 1 billion combats, and one machine is not enough to handle all this data. How

would you write the code of question 4 with a Spark implementation in Python **(6)**. The output must be a Python list with 10 names sorted in descending order.
You don't need to run the Spark code. Assume that you have 2 DataFrames (already loaded) called *pokemon* and *combats* with the same schema as the small CSVs.

Question 2: Prediction

Now you know what your data looks like, and you are ready to implement the new prediction app of the Pokedex 2.0. The app designed around your prediction model will take as input the features of 2 Pokémon, and it will generate a binary value to predict who will win.

Generate the feature vectors and the labels to train your model **(1)**. How do you handle the categorical data and Pokémon with multiple classes **(2)**?

Train a random forest to predict the winner of a match based on the available features. Split your data **(3)** into training and a testing set [90% / 10% random] and try different values of the number of estimators (n_estimators [10, 25, 50, 100]) and the maximum depth of the trees (max_depth [2, 4, 10]). What is the best combination of parameters **(4)**? Briefly describe your results.
Could this setup lead to wrong conclusions regarding the best hyperparameter setting? If so, describe why and how you would address the problem **(5).** (Hint: think about how you're making use of the data during training and testing). Implement your solution and show your results **(6)**.

Finally, Professor Oak asks you what the most predictive features are. Plot feature importance as a bar plot representing the ten most predictive variables **(7)**.

Question 3: Ranking

The new Pokedex is missing a smart way to rank the Pokémon. Now that you know the outcome of several battles you can sort the Pokémon from the strongest to the weaker.

Sort the Pokémon by winning ratio (wins-over-battles), show the top 10 and describe what you observe in the features **(1)**.

Pokémon tournaments can be represented as a dominance directed graph. In this graph, each Pokémon is represented by a single vertex and each battle outcome as a directed unweighted edge.

The graph represented as an adjacency matrix G has 1 in position e_ij to represent a superiority of node i over node j.
The elements of the matrix are non-zero only if Pokémon i won a strictly higher number of times against Pokémon j.

$$G_{i,j} = \begin{cases} 1, & \text{if pokémon i won the majority of the matches against j} \\ 0, & \text{otherwise} \end{cases}$$

A simple method to identify dominators out of this graph is to compute A = G + G^2, which compute first-stage dominance. The sum of the rows in A represent the dominance score of the Pokémons. What does this score represent **(2)**?

Compute A and extract the top-10 pokémon with the highest dominance score **(3)**. Compare them with their winning ratio. What do you observe? How do you explain the differences? **(4)**.

Now you have to integrate the final ranking in the new Pokedex. Which scoring method is better? Support your opinion with explanations and examples **(5)**.