Homework 1 (due 2/2)

DS-210 @ Boston University

Spring 2022

Before you start…

Collaboration policy: You may verbally collaborate on required homework problems. However, you must write your solutions independently without showing them to other students. If you choose to collaborate on a problem, you are allowed to discuss it with at most 2 other students currently enrolled in the class.

The header of each assignment you submit must include the field “Collaborators:” with the names of the students with whom you have had discussions concerning your solutions. If you didn’t collaborate with anyone, write “Collaborators: none.“ A failure to list collaborators may result in credit deduction.

You may use external resources such as software documentation, textbooks, lecture notes, and videos to supplement your general understanding of the course topics. You may use references such as books and online resources for well known facts. However, you must always cite the source.

You may **not** look up answers to a homework assignment in the published literature or on the web. You may **not** share written work with anyone else.

Submitting: Solutions should be submitted via Gradescope. More details will be provided on the course webpage and on Piazza. Please submit a solution to this homework as a single IPython notebook (`.ipynb`).

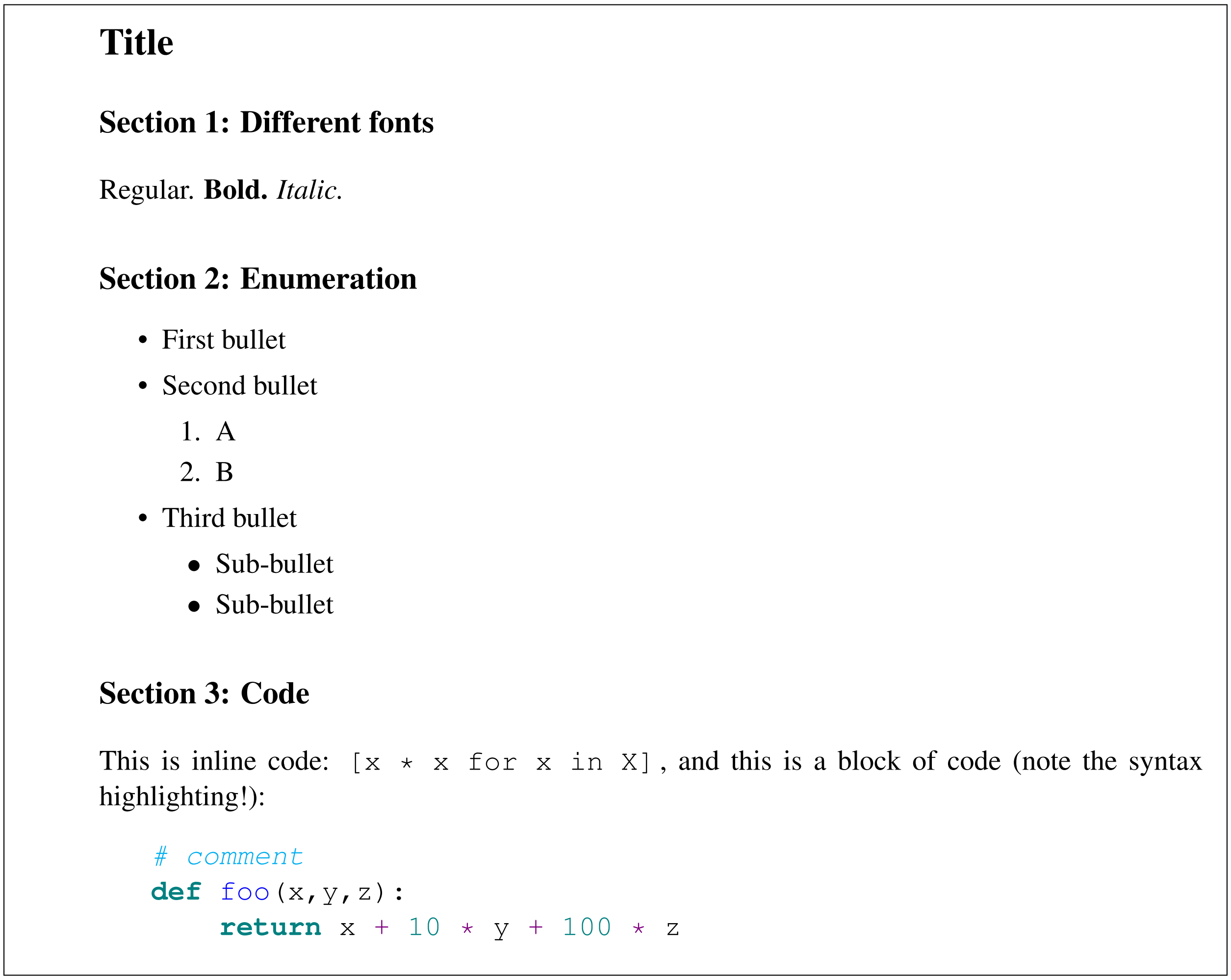
Grading: Whenever we ask for a solution, you may receive partial credit if your solution is not sufficiently efficient or close to optimal. For instance, if we ask you to solve a specific problem that has a polynomial–time algorithm that is easy to implement, but the solution you provide is exponentially slower, you are likely to receive partial credit.

Questions

Please submit a solution to this homework as a single IPython notebook (`.ipynb`).

1. Read the Markdown guide at https://medium.com/analytics-vidhya/the-ultimate-markdown-guide-for-jupyter-notebook-d5e5abf728fd

   Create a Markdown cell that roughly looks like the *content* of the following box (that is, do not include the box):



Title

Section 1: Different fonts

Regular. **Bold.** *Italic.*

Section 2: Enumeration

- First bullet
- Second bullet
  1. A
  2. B
- Third bullet
  - Sub-bullet
  - Sub-bullet

Section 3: Code

This is inline code: `[x * x for x in X]`, and this is a block of code (note the syntax highlighting!):

```
# comment
def foo(x,y,z):
    return x + 10 * y + 100 * z
```

2. Execute a simple data pipeline that involves:

   - Basic data validation (i.e., make sure no relevant attributes are missing) and—if needed—data cleansing.
   - Partitioning the data set into a training and test set.
   - Selection of the set of features that will be used in the learning process.
   - Training a decision tree.
   - Estimation of the quality of predictions by the final decision tree.

   Execute this pipeline for different target decision tree sizes and different sizes of the set of features used for learning and prediction. For the former, you can try various numbers of nodes that are multiples of 5. For the latter, you can select 3, 6, 9, etc. that you believe should be most important for what you are trying to predict. Compare the outcomes and plot a graph that displays the prediction accuracy.

   Suggested data set: https://archive.ics.uci.edu/ml/datasets/Student+Performance

Feel free to use a different data set if you find it more interesting for personal reasons, but if you do so, explain why you made this choice. Otherwise, if you use the suggested data set, predict attribute `G3` and do not use `G1` and `G2`. Additionally, this data set has grades for two subjects (Mathematics and Portuguese). Select just one of them.

Since in this data set the goal is to predict a numerical value, measure your accuracy as the expected square of the difference between your prediction and the actual value on the test set, or another similar quality measure

Summary: Write a short summary of what you learned. How did the accuracy depend on the size of a decision tree? How did the accuracy depend on the number of features you selected? Did you learn anything interesting about applying decision trees for predictive data analysis? Did you learn anything interesting about the data set?

Note: Please briefly explain all your design choices and what you do in the notebook whenever it is not obviously clear. Please use Markdown, which you learned in Question 1, whenever you create a Jupyter notebook in this and later homework.

3. (Optional, no credit) How much time did you spend on this homework? The answer will have no impact on the credit you receive, but it may help us adjust the difficulty of future homework assignments.