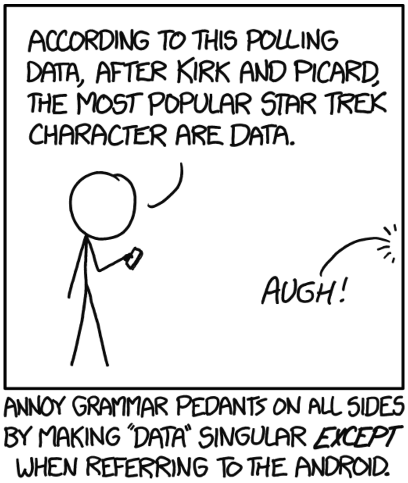
CS158 - Assignment 1
Data

Due: Friday, September 1 at 5pm


ACCORDING TO THIS POLLING DATA, AFTER KIRK AND PICARD, THE MOST POPULAR STAR TREK CHARACTER ARE DATA.
AUGH!
ANNOY GRAMMAR PEDANTS ON ALL SIDES BY MAKING "DATA" SINGULAR *EXCEPT* WHEN REFERRING TO THE ANDROID.

`https://xkcd.com/1429/`

0. Slack

   If you are not a member of the `cs158-fa2023` slack channel, message me and I'll add you.

1. **You know the drill**

   Read through the administrative handout on the course web page.

   What is the late policy for the course?

2. **Data, data, data**

   There are now lots of really interesting data sets publicly available to play with. They range in size, quality and the type of features and have resulted in many new machine learning techniques being developed.

   Find a public, free, supervised (i.e. it must have features *and* labels), machine learning dataset. You may NOT list a data set from 1) The UCI Machine Learning Repository or 2) from Kaggle.com. Once you've found the data set, provide the following information:

   (a) The name of the data set.
   The name of the data set is "The-Big-Five-European-soccer-leagues-data".

   (b) Where the data can be obtained.
   The data can be obtained at the OpenML if you visit this link `https://www.openml.org/search?type=data&status=active&id=43535`

(c) A brief (i.e. 1-2 sentences) description of the data set including what the features are and what is being predicted.
The data set contains soccer match information for teams in Europe's top 5 leagues - English Premier League, French Ligue 1, Spanish La Liga, Italian Serie A, and German Bundesliga - between the 1995/96 to 2019/20 seasons. Some attributes of the data set include the team names, the home and away scores, the match dates and days etc (a total of 15 attributes). The data set can be used to predict specific team trends over the next seasons to see which team is most likely to be relegated or which team is likely to clinch the league title.

(d) The number of examples in the data set.
The number of examples in the data set is 44,269.

(e) The number of features for each example. If this isn't concrete (i.e. it's text), then a short description of the features.
The total number of features for each example is 15.

Extra credit will be given for particularly interesting data sets, e.g. the most unique, the data set with the largest number of examples and the data set with the largest number of features.

3. **Data analysis**

One of the first things to do before trying any formal machine learning technique is to dive into the data. This can include looking for funny values in the data, looking for outliers, looking at the range of feature values, what features seem important, etc. At:

`http://www.cs.pomona.edu/classes/cs158/assignments/assign1/titanic-train.csv`

I have provided a modified version of passenger survival data for the Titanic[1].

This data set has six binary features:

- `First_class` (whether the passenger was in first class or not)
- `Sex` (0 = Male, 1 = Female)[2]
- `Age` (0 = <25, 1 = 25+)
- `SibSp` (had siblings/spouses aboard?)
- `ParCh` (had parents/children aboard?)
- `Embarked` (Left from Southhampton?)

Based on these features, the Titanic task is to learn to predict the last column, whether or not the passenger survived (1 = survived).

(a) For each of the features calculate (and write down) the *training error* if you used **only** that feature to classify the data. To do this you will need to do the following for each feature:

[1] The original data can be found at: `http://www.kaggle.com/c/titanic-gettingStarted`

[2] I recognize that there are diverse gender identities. In this case, we are limited by the data and will focus on sex as a binary feature.

- Split the data based on that feature. Call bin_0 all examples that have 0 for that features and bin_1 all examples that have 1 for that feature.
- Calculate the majority count for the label in each bin, i.e. for bin_0,

$$majority(bin_0) = max(count(bin_0 = survive), count(bin_0 = notsurvive))$$

  This value is how many examples you would get right in bin_0 if you split on that feature. Make sure you understand why!
- Calculate the training error for that feature. The accuracy on the training set (i.e. percentage correct) can be calculate as:

$$accuracy = \frac{majority(bin_0) + majority(bin_1)}{totalNumberOfTrainingExamples}$$

  and then

$$error = 1 - accuracy$$

You can either write a program to do this in any language you'd like (you don't need to submit the code) or you could also do this in a spreadsheet program like excel. Your answer to this problem should be 6 error training error rates, one for each feature.
Class training error = 0.3249299719887955
Sex training error = 0.21988795518207283
Age training error = 0.4061624649859944
SibSp training error = 0.4061624649859944
ParCh training error = 0.3851540616246498
Embarked training error = 0.3837535014005602

(b) Which feature would be the best to use? Put another way, if we were building a 1-level decision tree using Algorithm 1 from the book, which feature would it pick?
I would use the feature that has the lowest training error. In this case, it is the sex training error with a low training error of 0.21988795518207283.

(c) Do you agree that this is the best choice to make? Just 1-2 sentences explaining yes/no is sufficient.
I agree that this is the best choice to make if we were just considering a decision stump because the low training error of 0.21988795518207283 would ensure that we are right of the survival rate for most of the time.

4. Ethics and ML

As machine learning becomes more commonplace, it is increasingly important to think about the ethical implications, including when and how we apply the models, data collection practices, and biases that can be introduced into the models from the data. Find one article that discusses machine learning ethics:

(a) What is the article (i.e., title and url)?
The title of the article is Great Promise but Potential for Peril by Christina Pazzanese: The Harvard Gazette. You can find it at `https://news.harvard.edu/gazette/story/2020/10/ethical-concerns-mount-as-ai-takes-bigger-decision-making-role/`.

(b) Give a couple of sentence sentence summary of the article?
The article is actually about the adoption of AI which is basically an ML application. It especially focuses on the application of AI in healthcare where for example AI can be used to initiate faster diagnoses since essentially doctors will have a plethora of information about certain diseases at their disposal. Critics says that instead of eliminating human biases in certain areas like job reviews, AI could potentially perpetuate those biases while other proponents of AI say that AI is here to complement the role of humans instead of full-on replacing them. The article concludes by stating that there is a need for regulation of AI and the USA has been slow to adopt a federal guideline to control the spread of AI, unlike their European counterparts via the EU. Finally, the article gives a call to action to higher-education institutions to institute tech ethics courses in their curriculum.

(c) Are there other ethical concerns that you can think of related to the article topic that weren't mentioned?
Another concern arises when applying for loans from banks. If a bank's training data is biased and historically inclined to deny loans to minority groups like Black/African American individuals, then there is a risk that the bank may continue to use machine learning models to deny loans to other minority groups, thus perpetuating the bias.

When you're done

Put your answers to the questions above in either a `.txt` or `.pdf` file *with your name at the top of the file* and named *lastname.1* (with the corresponding file extension).

Submit your assignment via gradescope.