DATA 23700 Autumn 2023

Assignment 3: Persuasive or deceptive visualization?

Due November 13, 2023

"Interactivity is a multiplier" insofar as it enables us to navigate many more views of data than we would be able to in a static display. This is especially helpful when using visualizations to explore many dimensions of a dataset simultaniously. Models provide a complementary way of characterizing multidimensional data by constructing latent parameter spaces which can sometimes project datasets onto lower dimensional representations called *embeddings*. However, this *dimensionality reduction* that we get from certain types of models can be difficult to interpret. Here too interactive visualization can be helpful!

In this assignment, we will build a simple dashboard of interactive visualizations in order to explore the output of a dimensionality reduction technique called UMAP. We have applied this model to a dataset on California home prices and saved its output alongside the data in a file on the course website called `ca-housing-umap.csv`. Students will create an interactive visualization using either D3 or Altair, following the instructions below.

Students will *work alone*.

Students should submit their assignment on Gradescope as either: (1) a .zip file containing source code for a webpage constructed with D3; (2) or a .ipynb file containing a multi-view interactive visualization constructed with Altair.

Technical specification

Load the provided dataset `ca-housing-umap.csv` from the course website. This table contains all of the variables in the California housing dataset plus two columns labeled x and y, which represent the embedding space learned by the UMAP algorithm.

Students' task is to **build a dashboard** for the purpose of exploring the meaning of the embedding coordinates x, y produced by this dimensionality reduction technique. The dashboard must use crossfiltering to *connect a scatterplot of the x, y embedding space to at least five other views* of data showing other features in the dataset provided. Each view should be a visualization showing the distribution of a distinct feature or set of features (i.e., the same feature should not appear in multiple of these charts). The design of these at least five additional views is left to students' discretion, but we encourage students not to add so many encodings that it becomes difficult to add interactivity. Students should choose a layout for these charts that makes it easy to see them all simultaniously.

These **charts should be linked** such that brushing or otherwise making a selection on any chart in the dashboard filters all the data shown on all charts in view. The user should be able make selections in any and all of these charts in order to further refine their filter set. While the filtered data should be shown in the foreground in a pop-out color, the unfiltered dataset should be displayed in the background in gray.

The idea is to be able to understand the embedding coordinates in terms of the original feature values associated with them.

Students may use either D3 or Altair. For the D3 option, students should submit a zipped `src/` folder containing the following files: `ca-housing-umap.csv`, `chart.js`, `index.html`, and `README.md`. Following the demonstration in class, the `index.html` file should contain the structure and copy of the webpage, plus any styling. The `chart.js` file should contain all the code used to render the visualizations. Students who wish to break out D3 charts into separate components (e.g., for histograms, scatterplots, etc.) may submit these in different `.js` files, however, writing modular code is not a requirement for this assignment. The `README.md` file should contain students' written reflection (see below).

For the Altair option, students should submit a `.ipynb` file containing a notebook as in previous assignments. The whole dashboard should be rendered in one code block, so the multiple charts can interact with each other and share a common set of selections. The written reflection (see below) should be in a separate text block below the dashboard. Students taking this option will find this tuitorial helpful. It covers interactivity in Altair, including the functionality required for this assignment.

Students taking the D3 option will earn an additional S on this assignment as long as they score above U for quality of visualization and quality of write up. This extra S is an acknowledgment that, generally, programming in D3 is more difficult than programming in Altair. It is meant to provide a cushion for students who choose to push themselves outside their comfort zone on this assignment.

Finally, **students must write a two paragraph reflection**. In the first paragraph, students should comment on their choice to use D3 vs Altair to build the dashboard. What do we gain from using a lower-level graphics tool like D3 over Altair? What are the benefits of using Altair to add interactivity to data science workflows? In what kinds of situations would you prefer to use each of these levels of abstraction?

In the second paragraph, students should describe their design process. What decisions did you have to make that weren't outlined in this technical specification? Why did you make those design choices? What if anything did you struggle with about this assignment? What do you see as the benefits and drawbacks of interactivity writ broadly after attempting to build interactive visualizations?