Applied Machine Learning Homework 1

Due: Wed 02/05/20 1pm
All assignments need to be submitted as jupyter notebook (.ipynb) via github classroom:
https://classroom.github.com/a/4BAF66CR
and as PDF via gradescope.

Make sure to not commit any data to github classroom.
All tasks can be done in a single jupyter notebook. Please ensure that your jupyter notebook does not have too much spurious output.

All exercises will work on the Australia Fire dataset:
https://www.kaggle.com/carlosparadis/fires-from-space-australia-and-new-zeland
We will work with fire_nrt_V1_96617.csv which is described here:
https://earthdata.nasa.gov/earth-observation-data/near-real-time/firms/viirs-i-band-active-fire-data

Task 1: Density Plots (50 points)

1.1 Plot the longitude vs latitude several ways within a single figure (each in its own axes):

1) Using the matplotlib defaults.
2) Adjusting alpha and marker size to compensate for overplotting.
3) Using a hexbin plot.
4) Subsampling the dataset.

For each but the first one, ensure that all the plotting area is used in a reasonable way and that as much information as possible is conveyed; this is somewhat subjective and there is no one right answer. [45 pts]
1.2 In what areas are most of the anomalies (measurements) located? [5pts]

Task 2: Visualizing class membership (50 points)

Visualize the distribution of Brightness temperature I-4 as a histogram (with appropriate settings). Let's assume we are certain of a fire if the value of temperature I-4 is saturated as visible from the histogram.
2.1 Do a small multiples plot of whether the brightness is saturated, i.e. do one plot of lat vs long for those points with brightness saturated and a separate for those who are not (within the same figure on separate axes). You can pick any of the methods from 1.1 that you find most suitable. Can you spot differences in the distributions? [20 pts]
2.2 Plot both groups in the same axes with different colors. Try changing the order of plotting the two classes (i.e. draw the saturated first then the non-saturated or the other way around). Make sure to include a legend. How does that impact the result? [20 pts]

2.3 Can you find a better way to compare the two distributions? [10pts]

Bonus

(no points and only semi-related to the rest of the class)

Find a dataset for classification or regression on Kaggle. Install dabl ('pip install dabl') and use dabl.plot to visualize the dataset. Briefly discuss the insights from this output. See how you can improve over these visualizations and what new insights you can get. As mentioned in class, dabl will not be allowed for any of the graded assignments. If you get an error, file an issue at https://github.com/amueller/dabl .

Helpful Resources

1. For anyone having trouble setting up your environment refer to this tutorial:
https://conda.io/projects/conda/en/latest/user-guide/tasks/manage-environments.html

2. For anyone struggling with Git / Github:
https://services.github.com/on-demand/downloads/github-git-cheat-sheet.pdf