Dataset

I found this dataset uploaded to Kaggle from the University of California Irvine Machine Learning Repository. It contains data relating to red variations of the Portuguese wine "Vinho Verde" collected by Dr. Paulo Cortez at the University of Minho and used in the paper *Modeling wine preferences by data mining from physicochemical properties*[1]. From the notes on Kaggle, I found an expanded version with data for both red and white variations. This data seems quite simple to analyze which is useful in a short term project such as this, however I was hopeful to add a certain degree of additional complexity; thus, I was very happy to find supplementary data on white wines which would allow me analyze how the relationships would change between red and white variations.

Methodology

Preprocessing

The data is provided in a very easy to use format. It is a .csv file (although it was initially formatted with semi-colons and I had to replace these with commas) with 11 different physicochemical properties each in a different column followed by a column for the quality of the wine with the aforementioned attributes. Every line represents a different wine. Depending on how initial tests proceed, I may transform the quality values to categories of “good” and “bad” (perhaps with an “ok” category as well). There is also a concerning issue that this data is imbalanced as the vast majority is centred on quality ratings of 5 and 6, there are very few low or high ratings, and no absolute extreme ratings (0 or 10). I will have to be careful of how I treat the data, hopefully I can use some form of cross-validation to ensure the data is usable.

Model

From what I’ve learned about machine learning models, as well as what I’ve gathered from the Kaggle discussion on this dataset, this data should be simple enough to approach with some form of regression algorithm. There is also some suggestion that a random forests approach would be effective, however this would later impose a limit on the predicted outcomes as the random trees model is not capable of extrapolation (since this dataset is limited and only includes quality values between 3 and 8 for red wines and 3 and nine for white ones, the predicted outputs would be limited to this same range). This issue could be solved if I were to convert my outputs to a categorical system where anything above 7 is considered “good” (then have the lower values be “bad”, or split between “ok” and “bad”), but I feel that a direct numerical prediction would be far more interesting. I feel that ridge regression is a good choice as an initial approach for this problem, due to the relatively simple relationships in the data and the fact that some of the properties will be correlated.

Conceptualization

My current intention is to use this data to create a web app using Flask. This would provide users the ability to select values for the components with sliders and then the quality of the resulting wine would be estimated.

---

[1] P. Cortez, A. Cerdeira, F. Almeida, T. Matos and J. Reis. Modeling wine preferences by data mining from physicochemical properties. In Decision Support Systems, Elsevier, 47(4):547-553, 2009.