Homework 3

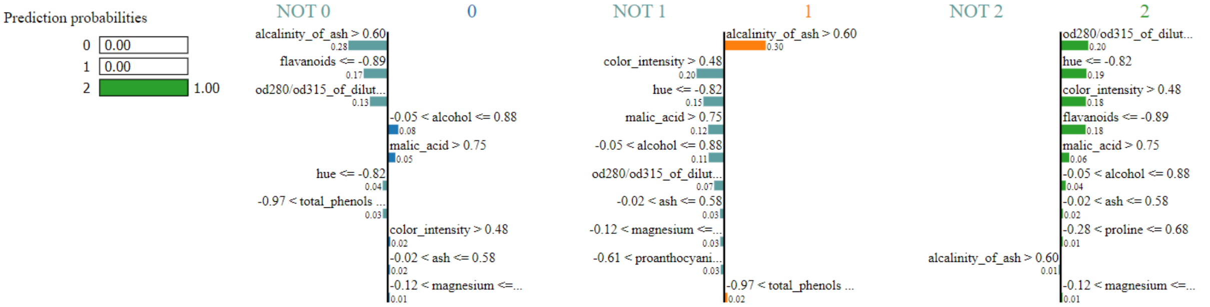
For this analysis I'll use the UCI ML wine recognition dataset to predict wine cultivator using data from chemical analysis of wine. My main model will be a multi-layer perceptron with one hidden layer with 100 neurons. It'll be compared later to a gradient boosting classifier with 200 regression trees.

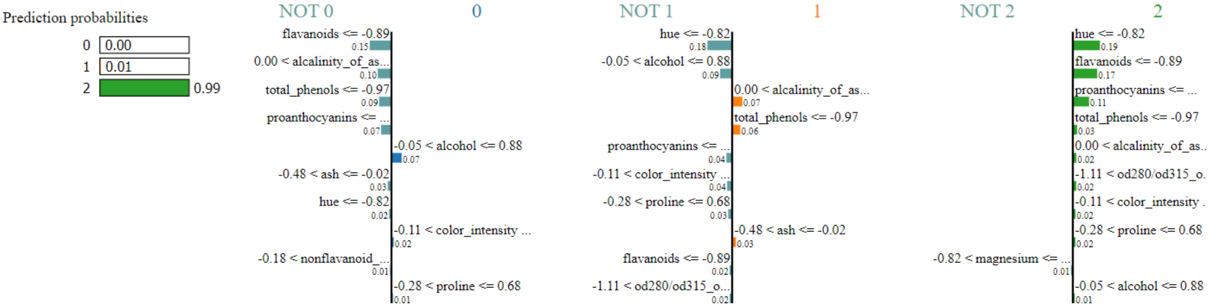
Basic LIME interpretation

Most of the LIME decompositions in this document use discrete features. Which means that we will often see things like: feature "$hue \leq -0.87$" has a coefficient of 0.15. That means that on average (considering the training data distribution), having *hue* in this bucket (from minus infinity to -0.87 inclusive) raises the prediction for a given class by 0.15.

Task 4

I'll start by comparing LIME decompositions for different observations. Since LIME explains predictions locally we may get different decompositions in different areas or feature space.





In both cases the same prediction was made with high probability. Top features that caused predicting second class in the first case were:

- *od280/od315_of_diluted_wines* (a method for determining the protein concentration)
- *hue*
- *color_intensity*
- *flavanoids* (polyphenolic secondary metabolites found in plants)

and in the second case:

- *hue*
- *flavanoids*
- *proanthocyanins* (chemical compounds that give the fruit or flowers their red, blue, or purple colors)

As we can see there are some differences. For example, the most predictive feature for the first observation has almost no predictive power for the second observation. What is more, decompositions for other classes differ as well. *alcalinity_of_ash* has over four times larger coefficient that predicts that wine comes from cultivator encoded as “1”.

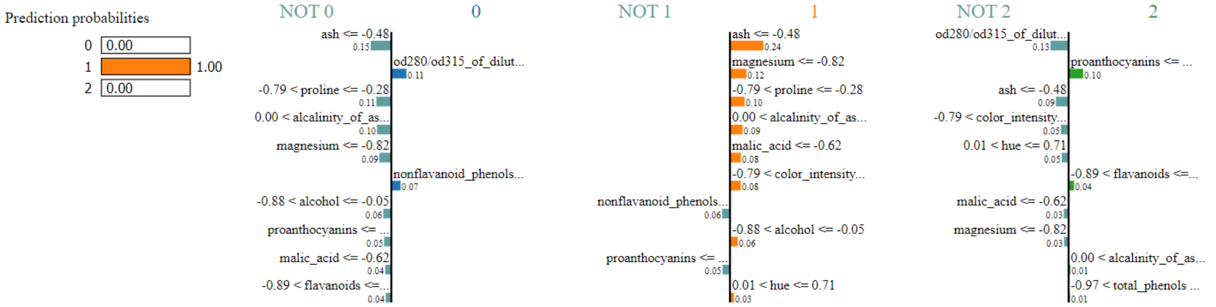
Now lets look more globally on those coefficients. After taking all coefficients for predictions of class 2 on all test observations we get following statistics:

```
              Feature Name           |    Mean    |    Std
-----------------------------------------------------------
alcohol                              |      0.007 |      0.016
malic_acid                           |      0.005 |      0.024
ash                                  |      0.003 |      0.033
alcalinity_of_ash                    |      0.002 |      0.007
magnesium                            |       -0.0 |      0.009
total_phenols                        |      0.007 |      0.016
flavanoids                           |      0.042 |      0.074
nonflavanoid_phenols                 |      0.001 |      0.005
proanthocyanins                      |      0.016 |      0.052
color_intensity                      |      0.021 |      0.065
hue                                  |      0.037 |      0.072
od280/od315_of_diluted_wines         |      0.035 |      0.071
proline                              |      0.001 |      0.004
```

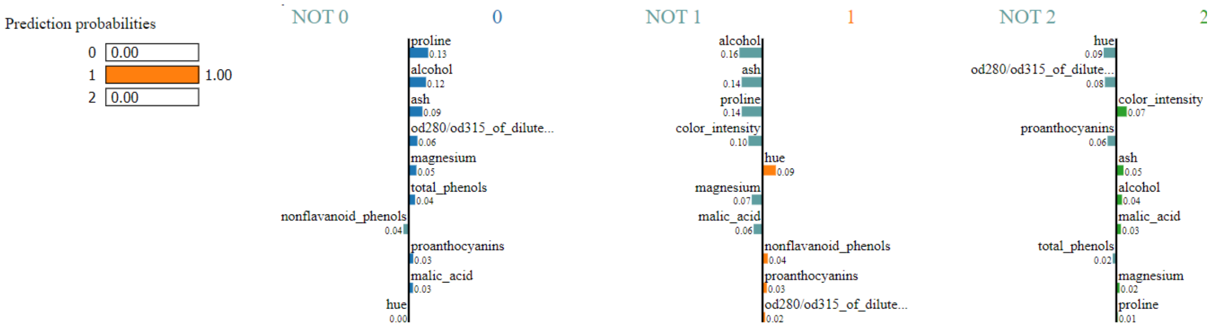
In fact standard deviation of these coefficients seems to be pretty high.

Discretization

Python implementation of LIME by default discretizes continuous features into quartiles. That’s why in decomposition we can see features like: “$ash \leq -0.48$”. Here is an example:



We can switch off discretization. If we do that above image would change into this:



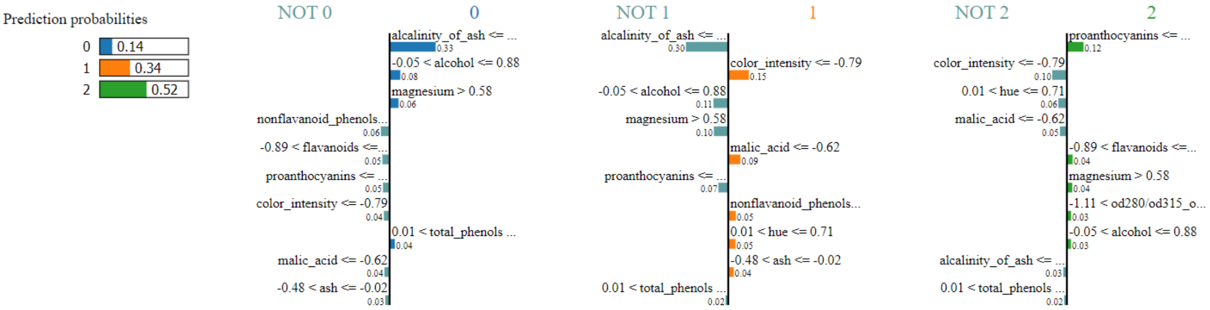
As we can see they are completely different. Moreover from bottom decomposition we should get a prediction of 0, not 1. Based on this representative example we can say that decomposition without discretization makes no sense. Interpreting results for numerical features is much harder because of two reasons (according to creators of this Python library):

- It's hard to think about double negatives (i.e. negative weight for a negative feature = positive contribution).
- The values may be in different ranges. We can always standardize the data, but then the meaning of the coefficients changes.

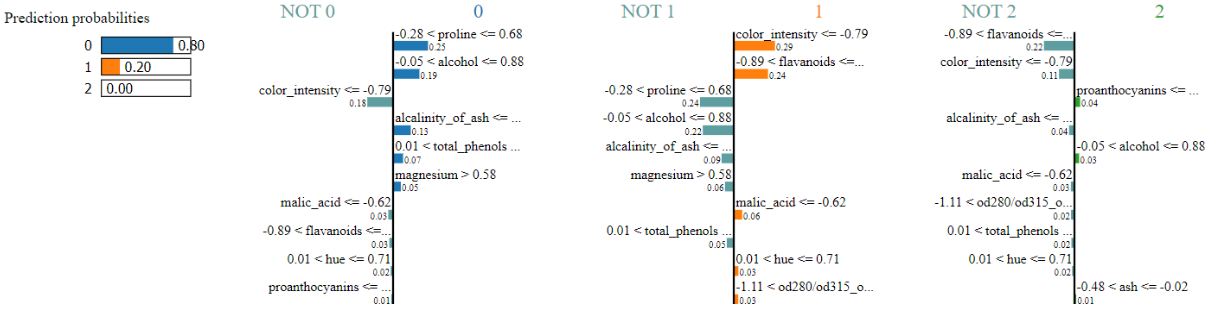
But that's not the only problem with LIME decomposition. Apparently often there are many equivalent linear models that fit the data equally well. In this case LIME picks an arbitrary one, so the weights are not necessarily going to be the same. That means that LIME is very sensitive to high correlation between variables.

Task 5

After training the second model I realized that both of them have exactly the same accuracy on the test set (98%). That means that both models failed only on one test case (out of 59). During examination I found out that it was the same observation in both cases. I'll use this observation to compare models because they were the most different there. Decomposition for the first model:



Decomposition for the second model:



As we can see both models are less certain about their prediction (usually both models made predictions with probability 99% - 100%). Decompositions are completely different, as well as predicted class, but we should remember that that's an extreme case. For sake of completeness I include the same table with global coefficient statistics for second model:

```
         Feature Name          |    Mean    |    Std
--------------------------------------------------------
alcohol                        |      0.007 |       0.02
malic_acid                     |      0.002 |      0.007
ash                            |       -0.0 |      0.008
alcalinity_of_ash              |      0.002 |      0.006
magnesium                      |      0.004 |      0.021
total_phenols                  |      0.002 |      0.008
flavanoids                     |      0.155 |      0.309
nonflavanoid_phenols           |     -0.001 |      0.004
proanthocyanins                |      0.001 |      0.007
color_intensity                |      0.013 |      0.029
hue                            |       0.03 |      0.065
od280/od315_of_diluted_wines   |      0.021 |      0.058
proline                        |      0.005 |      0.011
```