Alicia Milloz
Fanny Strasser

Mini Project BIO-322 : Project 1

The goal of this project is to predict how pleasantly a new molecule smells in a regression task. To do so, the smell of some molecules is analysed regarding their physical and chemical features. Indeed, the data set to build this project consists of a csv file with valence pleasantness, that is to be predicted, intensity (low or high), sweet or sour (true or false), complexity, and specific molecular features values.

Firstly, we started the exploration of the data by removing all the possible NAs values in our data set to reduce its initial size (708 smell experiences and 4872 predictors). Then, the boxplot of pleasantness versus intensity, taking high or low value, reveals that odor molecules with a high and with a low intensity have nearly the same pleasantness mean. However, odors with low intensity values have a bigger interquartile range, meaning that the pleasantness takes more spread values. We produced a matrix of scatterplots for the first 60 predictors to see if some of them are linearly dependent. We can indeed notice a linear dependency between some of the predictors. To perform prediction methods efficiently, the data set needs to be cleaned. Indeed, we removed the predictors with a null variance, that is, the predictors that are constant. Therefore, they do not have any influence on the pleasantness. Then, we also suppressed the correlated predictors (corresponding to a correlation above 0.9) to reduce the size of the data set without losing precision on the future predictions. Finally, data sets can have different magnitudes that can induce lower performances. So we choose to normalize the entire dataset to perform linear regression. Normalization can simplify visualization of data and values interpretation. We also randomly divided the data set into a training and a test set of the same size.We performed all the following methods on the train set and evaluated the MSE on the test set.

To compute multilinear regression, we need to reduce the number of predictors by choosing the ones with the most influence on valence pleasantness. To do so, we tried to use forward regression, but the value of the adjusted R-squared was too low, meaning that this method is not useful in this case.We then performed multilinear regression using all the predictors. Some predictors were not defined (NA values) because of singularity meaning that the variables of the training set are not linearly independent. To overcome this issue, we can reduce the number of linearly dependent predictors. To do so, Lasso-regularization can be used. This subset selection method employs the least squares method to fit a linear model that contains a subset of the predictors. In our project, in order to perform linear regression with L1 regularization we loaded the library glmnet and represented the input data as a matrix (x.train) and the response as a vector. The glmnet function performs linear regression with L1 regularization if alpha = 1 and L2 regularization if alpha = 0. The lambda argument sets the size of the allowed area.Therefore, we choose a sequence of 800 points between 10 and 10e-2 to perform linear regression for different values of the size of the allowed area. We then selected the best lambda with cross-validation. We obtained a model with 28 predictors. Unfortunately, the test error is still very high (0.9) due to the fact that we may have suppressed too much information by removing a high number of predictors.Similarly, the L2 regularization did not fit the model very well, with a test error even higher of 1.7. As we have a dimensional issue due to the very high number of predictors, we performed PCA : this method can reduces the dimensionality of a dataset consisting of a large number of interrelated variables, while retaining as much as possible of the variation present in the dataset. We found that the first 35 Principal components are sufficient to explain the 80% of the variance meaning that we can select them to perform linear regression. The following multilinear regression performed with those 35 components returned a test error of 0.93. It is the multilinear regression performed with Lasso L1-regularization that gave the best result. However, this method is not an optimal solution to predict pleasantness according to the corresponding MSE value.

Alicia Milloz
Fanny Strasser

As we previously noticed, it seems that our model indicates a highly non-linear and complex relationship between the molecular features and the pleasantness. Consequently, we then performed nonlinear methods in order to try to find a better model to fit our data. Contrary to linear methods, we used for this part non scaled data. The test set allows us to find a magnitude for the MSE errors of the different methods we choose to assess in order to find the best one. Indeed, we firstly build a simple regression tree using the training set. The MSE found on the test set is already considerably lower than the one computed with linear methods considering that the data is non scaled. However, we can still find a better fit by performing tree prunning, as the last regression tree might be too complex and is likely to overfit the data. The prunning tree method allows to select a subtree that leads to a lowest test error rate. In order to find the optimal number of tree leaves, we performed six-fold cross- validation to estimate the cross-validated MSE of the trees. The plot of the tree size versus the train, test and CV errors indicates that the best tree size is 20. Even if the test error is higher than the one found with a simple regression tree previously, the CV error is noticeably lower at size 20, explaining this choice. We plotted the corresponding tree and calculated the corresponding errors. This model predicts Mor24m as the most important predictor and the training error is 3.35e-31, the test error is 2.26 and the CV error is 8.93. The big difference between the training error and the two other ones suggests that this method can still be too flexible. Generally, tree based methods are simple and useful for interpretations but they are not always accurate enough for prediction. Indeed, in order to seek the best model to predict valence pleasantness, we used another tree approach named boosting. As an alternative to fitting a single large decision tree to the data train, boosting method works in a learning way : instead of building trees independent of others, boosting method builds trees using information from previously grown trees. Moreover, to improve the boosting method and avoid overfitting, we decided to implement Gradient Boosting. We first performed boosting with the RSS as a loss function on the training set with 1000 trees for a range of the shrinkage parameter to find the optimal lambda to reduce the mean squared error. We found a lambda of 0.03 and used it to perform boosting to find our model. This method gave a lower test error of 2.8.

To conclude, the prediction of human olfactory perception from chemical features of odor molecules can be done by building different models that seems to be more or less accurate. Indeed, a linear model such as multilinear regression is not an efficient method to predict valence pleasantness, revealing that chemical features of odor molecules are not linked to pleasantness in a linear way. However, the performance of the non-linear method of trees, such as prunned tree boosting trees to predict valence pleasantness is unquestionably better. Indeed, the lower MSE we found is 2.26 by using a prunned tree. However, by trying to predict valence pleasantness using a provided dataset (test_data.csv), we noticed that the boosting tree method returns the lowest MSE and is consequently the strongest prediction method. This can be explained by the fact that prunning tree overfits a lot the data. Therefore, knowing that the pleasantness takes values from 0 to 100, we can surely say that tree methods and especially boosting trees are an efficient way to predict the valence pleasantness of an odor from chemical features of odor molecules.