Olfaction Prediction from Molecule Structure

Xiufeng Zhao, Haocheng Zhu, Junliang Yu

**Introduction:**

It is supposed that different kinds of molecule has different kinds of smell. We therefore want to study the link between them. To be more specific, the chemical structure of molecules affects our olfaction. In this project, we study the relationship between the chemical structure and olfaction and compare different machine learning models to predict the smell of a molecule based on its chemical structure.

**Problem Definition:**

In this project, we are provided with two datasets: 1) molecular descriptor dataset which contains 4870 molecular descriptors for each of the 476 molecules; 2) perceptual dataset which contains 21 olfaction descriptors from each of 49 volunteers for each of the 476 molecules, and each molecule are tested in 2 kinds of dilutions. This is, in fact, a regression problem, and there are two subproblems:1) predict 21 descriptor values for each volunteer given molecule and dilution; 2) predict the mean and standard deviation of 21 descriptors across 49 volunteers given molecule and dilution.

As for the dataset, 397 of 476 molecules are split as training samples and the other 69 molecules are used as testing samples to evaluate the performance of learned models.

**General Pipeline:**

To solve this problem jointly, we train one regression model for each volunteer and the result will be the prediction answer to the first subproblem. And then, we compute the mean and standard deviation of predicted values across 49 volunteers , as the predicted answer to the second subproblem.

To deal with the unnormalized values in molecular descriptor dataset, we first perform normalization and then Principal Components Analysis to reduce data dimension from 4869 to 49. As for dilution values, we set dilution value as 1 for samples in high dilution and as 0.5 for samples in low dilution. They are then concatenated with reduced molecule descriptors as input for machine learning models. The output of machine learning models is a prediction on 21 olfaction descriptors for each of 49 volunteers.

**Regression Methods:**

As for regression method, we apply four machine learning algorithms includes: Lasso, Ridge Regression, Bayesian Ridge Regression, and Random Forest.

**Lasso Regression**

Lasso, or least absolute shrinkage and selection operator, is a method for a regression problem. Basically, it is defined as a linear regression model with L1 regularized loss. Say if we have N samples in the p-dimensional space: . We also have to be the outcome of . Then, the Lagrangian form of loss function becomes

and we want to minimize L for is a p dimensional vector.

**Ridge Regression**

Here the regularization term becomes quadratic. The loss function is

**Bayesian Ridge Regression**

Assume that the output is Gaussian distributed around :

where is a random variable to be estimated, The prior for the coefficient is

The priors over and are chosen to be gamma distributions.

And the distribution of w can be written using Baye’s theorem:

**Random Forest**

Random Forest is actually a bagging strategy. It ensembles several decision tree models to make generalization. Use ID3 algorithm to build a tree. First calculate the standard deviation of current data . Second, for every attribute , split the data and calculate the standard deviation between the two attributes, where

Then we choose the attribute that maximize the standard deviation reduction to form a new branch::

Recursively find new branches until coefficient of variation is smaller than a threshold:

**Analysis:**

In a paper[1] published by the competition organizers, they use Z-scores to evaluate the correlation between predicted values and the actual data. But the description on how to compute the Z-scores is unclear, so we turn to mean square error(MSE) as an evaluation metric because it is the most commonly used metric for regression problem. The results of the different methods are shown in the following. The best results are underlined.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Method | Individual MSE | mean MSE | std MSE | Overall MSE |
| Lasso | 431.176 | 112.275 | 347.692 | 891.143 |
| Ridge | 443.465 | 125.066 | 347.692 | 916.223 |
| Random Forest | 429.449 | 107.272 | 306.058 | 842.779 |
| Bayesian Ridge | 400.691 | 81.471 | 347.692 | 829.854 |

Individual MSE denotes the MSE between each predicted value and the actual data, mean MSE denotes the MSE between the mean of each predicted value across 49 volunteers and std MSE denotes the MSE between the standard deviation of each predicted value across 49 volunteers. Overall MSE is the sum of the previous three MSE.

In the aforementioned paper, Lasso Regression and Random Forest Regression are among the best models, but details about how to implement these models are not included. It can be seen from the table that Lasso Regression and Random Forest Regression perform moderately well(overall MSE less than 900), and among other experiments we did these results are not bad compared to all other regression methods, which proves the correctness of results in the paper.

But Bayesian Ridge Regression outperforms both of them. It is not mentioned whether researcher of [1] tried this method, but the reasons for this outperform can be as following: It consider not an optimal value but a range of values with continuous probability, so it take all possible situation into consideration and this is why it is better than Lasso and Ridge Regression. Random Forest here ensambles several decision trees. The branches of the trees are picked greedily, which may be the reason it is not performing so well.

**Conclusion:**

The result we got is reasonable because the mean square error is around 400, then the average difference between real scores and predict scores is about 20 percent. For example, it’s hard to tell the score difference between 30% and 50% on an odor. So our result is close to the real score. From our experiment, the best machine learning algorithm on this problem is Bayesian Ridge Regressor.

**Reference:**

[1] Keller A, Gerkin R C, Guan Y, et al. Predicting human olfactory perception from chemical features of odor molecules[J]. Science, 2017, 355(6327): 820-826.