Practical 1:Predicting the Efficiency of Organic Photovoltaics

Haoqing Wang, Willie Jin, Youbin Kim

February 10, 2017

quantity in your submission.

1 Technical Approach

Overall, we tried several approaches to attempting to solve the problem. This can be divided into feature engineering, random forest tuning, and gradient boosting tuning.

Before trying various regression techniques, we realized the importance of feature engineering. In order to accurately predict the HOMO-LUMO gap for various molecules, we need to learn using relavent features. We believe that molecules that are structurally similar are likely to have similar HOMO-LUMO gaps. In order to judge similarity between molecules, we use RDkit to generate structural fingerprints of each molecule. The type of fingerprint we used are Morgan (circular) fingerprints, which hashes sections of the molecule with increasing radii up to a preset radius. Given a wide diversity of molecules, a diameter of 4 or 6 gives the best results in measuring similarity (O'Boyle & Sayle, 2016). We thus fingerprinted each molecule with a diameter of 4 into a 256 bit vector. Finally, we appended these 256 new "features" to the existing data file given.

Just listing things we should talk about: first writing for loops to try to tune RF parameters–¿ gridsearch–¿ random search(because took too much time) attempting to use oob for rf but got error (UserWarning: Some inputs do not have OOB scores. This probably means too few trees were used to compute any reliable oob estimates. warn("Some inputs do not have OOB scores. " done loading)

Thus did 80/20 split on test data. Attempted to do 5-fold cross-validation for parameter tuning using grid search, took too much time.

How did you tackle the problem? Credit will be given for:

2 Results

This section should report on the following questions:

- Did you create and submit a set of predictions?
- Did your methods give reasonable performance?

You must have *at least one plot or table* that details the performances of different methods tried. Credit will be given for quantitatively reporting (with clearly labeled and captioned figures and/or tables) on the performance of the methods you tried compared to your baselines.

Mention Features	
Feature	Value Set
Mention Head	$\mathcal{V}$
Mention First Word	$\mathcal{V}$
Mention Last Word	$\mathcal{V}$
Word Preceding Mention	$\mathcal{V}$
Word Following Mention	$\mathcal{V}$
# Words in Mention	$\{1, 2, \ldots\}$
Mention Type	$\mathcal{T}$

Table 1: Feature lists are a good way of illustrating problem specific tuning.

Model	Acc.
Baseline 1	0.45
Baseline 2	2.59
Model 1	10.59
Model 2	13.42
Model 3	7.49

Table 2: Result tables can compactly illustrate absolute performance, but a plot may be more effective at illustrating a trend.

3 Discussion

End your report by discussing the thought process behind your analysis. This section does not need to be as technical as the others but should summarize why you took the approach that your did. Credit will be given for:

- Explaining the your reasoning for why you seqentially chose to try the approaches you did (i.e. what was it about your initial approach that made you try the next change?).
- Explaining the results. Did the adaptations you tried improve the results? Why or why not? Did you do additional tests to determine if your reasoning was correct?