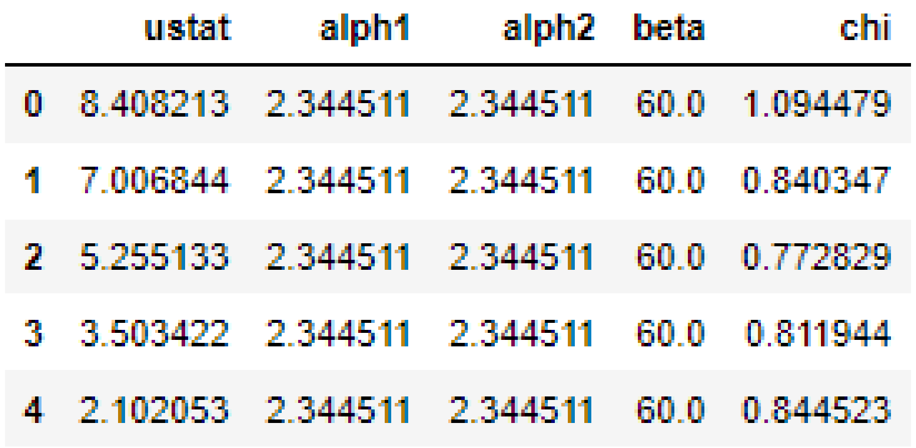
Week 01 and 02 Report

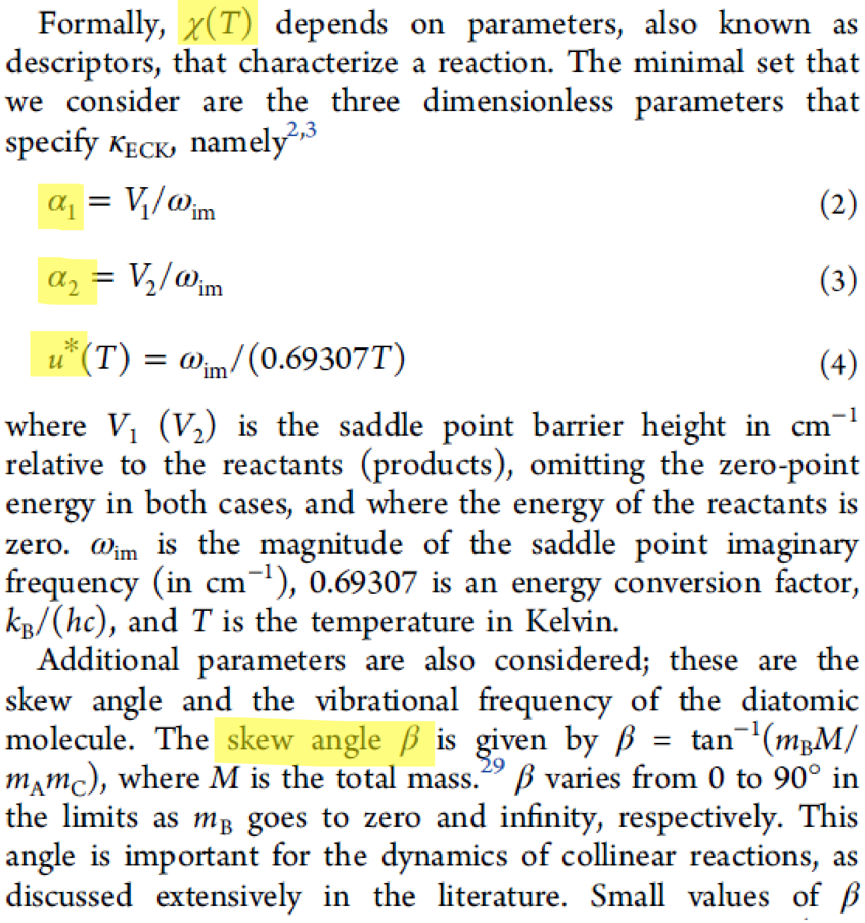
Jeffrey Li

Current Dataset

- **Features:** ustat, alph1, alph2, beta
- **Target:** chi
- 404 data instances



	ustat	alph1	alph2	beta	chi
0	8.408213	2.344511	2.344511	60.0	1.094479
1	7.006844	2.344511	2.344511	60.0	0.840347
2	5.255133	2.344511	2.344511	60.0	0.772829
3	3.503422	2.344511	2.344511	60.0	0.811944
4	2.102053	2.344511	2.344511	60.0	0.844523



Formally, $\chi(T)$ depends on parameters, also known as descriptors, that characterize a reaction. The minimal set that we consider are the three dimensionless parameters that specify κ_{ECK}, namely[2,3]

$$\alpha_1 = V_1/\omega_{\mathrm{im}} \tag{2}$$

$$\alpha_2 = V_2/\omega_{\mathrm{im}} \tag{3}$$

$$u^*(T) = \omega_{\mathrm{im}}/(0.69307T) \tag{4}$$

where V_1 (V_2) is the saddle point barrier height in cm^{-1} relative to the reactants (products), omitting the zero-point energy in both cases, and where the energy of the reactants is zero. ω_{im} is the magnitude of the saddle point imaginary frequency (in cm^{-1}), 0.69307 is an energy conversion factor, $k_B/(hc)$, and T is the temperature in Kelvin.

Additional parameters are also considered; these are the skew angle and the vibrational frequency of the diatomic molecule. The skew angle β is given by $\beta = \tan^{-1}(m_B M/m_A m_C)$, where M is the total mass.[29] β varies from 0 to 90° in the limits as m_B goes to zero and infinity, respectively. This angle is important for the dynamics of collinear reactions, as discussed extensively in the literature. Small values of β

Purpose

- How can we divide existing data into reasonable clusters?
    - Previously, **chi = 3** was used as a boundary to separate two clusters (classes): small_chi and large_chi.
    - These clusters will be used later in the training of Gaussian Process Regression models.
- Classification Problem: How can we categorize data into classes?
    - Given reaction features (ustat, alph1, alph2, beta), we want to be able to identify small_chi and large_chi data instances.
    - This would be a good indicator of whether classes are separated well.

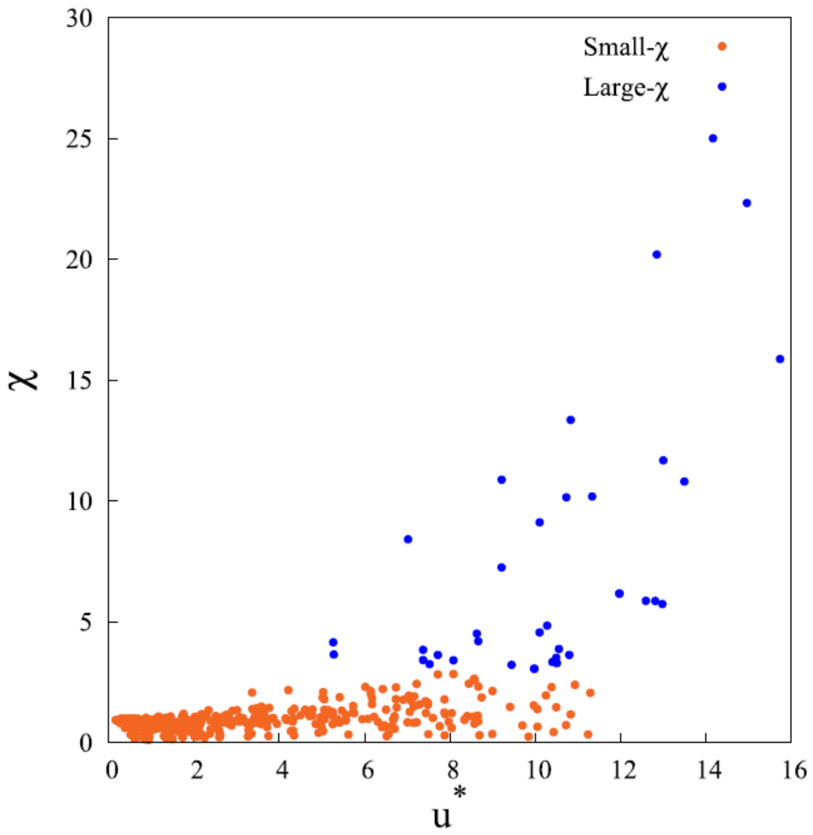
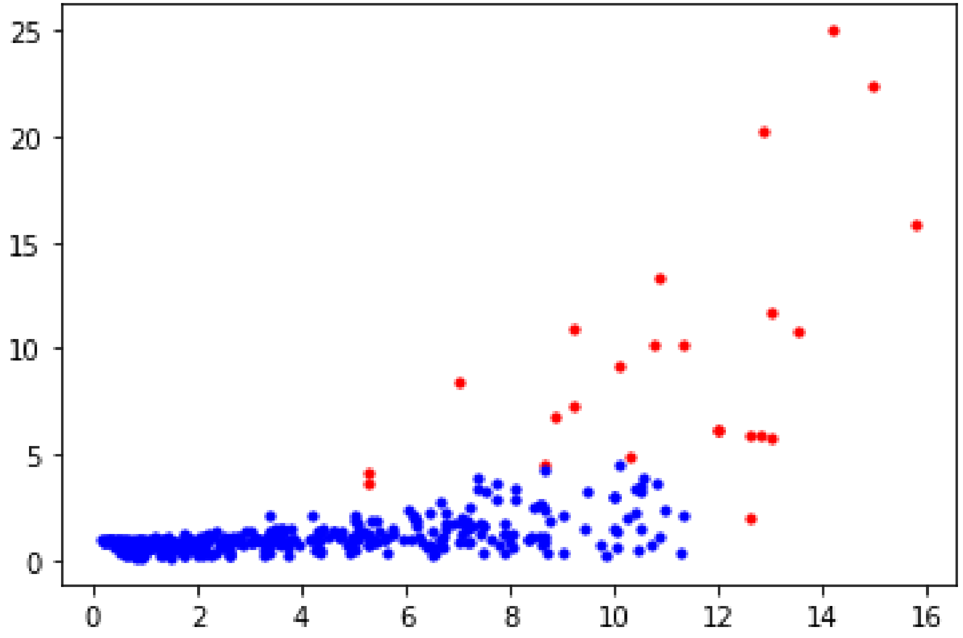
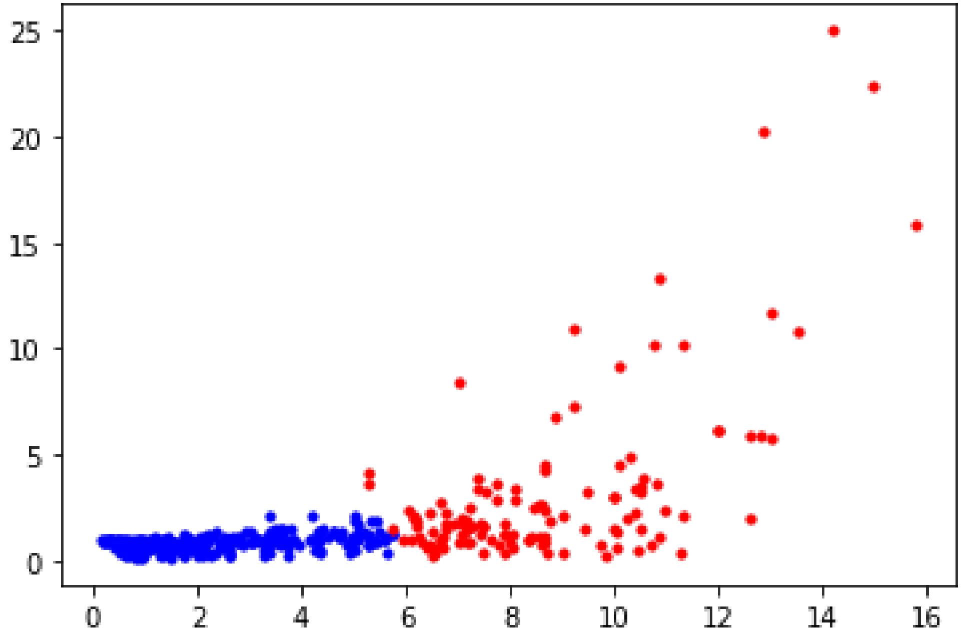


Figure 2. Exact χ versus u^* and for the entire χ data set.

Clustering Results

- Unsupervised Learning: So we cannot use SVM.
- DBSCAN Clustering: groups together points that are closely packed together
  - eps=1, min_samples=6
- K-means Clustering: partition n observations into k clusters
  - n_clusters=2
  - Centers: [ustat, chi]
    - [2.33877914, 0.83184532]
    - [8.66781375, 3.23312814]



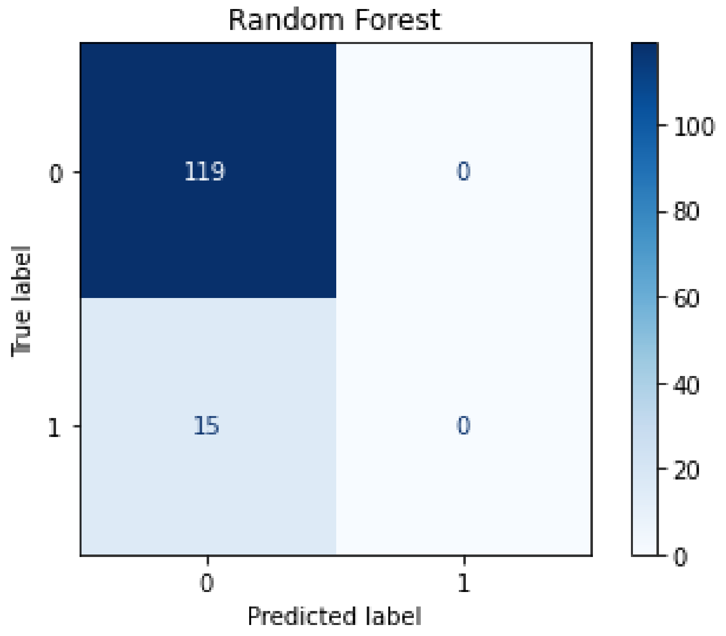


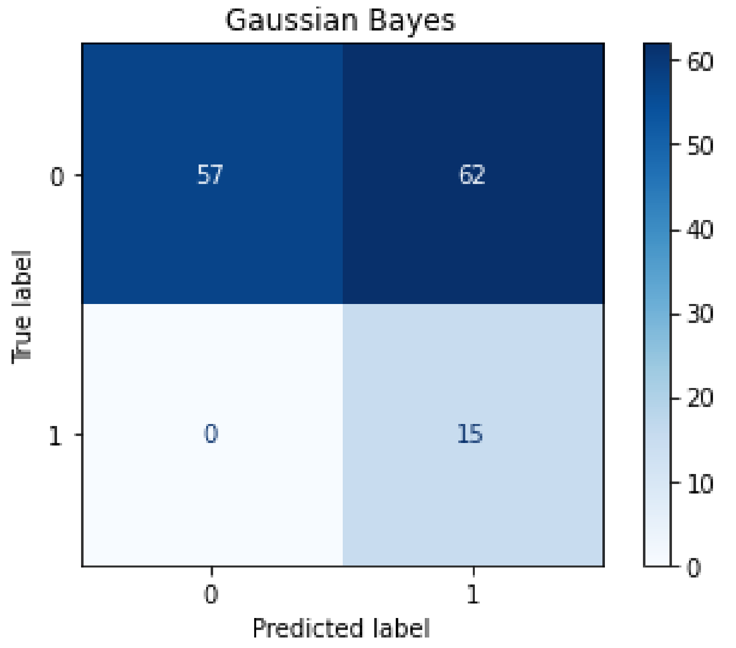
Classification

- Binary classification: Using small_chi (chi < 3) and large_chi (chi >= 3).
- Label encoding:
  - Chi: 0 = small chi (chi < 3), 1 = large chi (chi >= 3)
  - Ustat: 0 = (ustat < 3), 1 = (ustat >= 3)
  - Alph1: 0 = (alph1 < 1), 1 = (1 <= alph1 < 2), 2 = (alph1 >= 2)
  - Alph2: 0 = (alph2 < 2), 1 = (alph2 >= 2)
  - Beta: 0 = (beta < 50), 1 = (beta >= 50)

Classification Results

- For LR, RF, and AB, the confusion matrices display only results for TN and FN. This means that the models only predicts small chi for all data instances. This may be a result of having a large disproportion of small chi to large chi data.
- For GB: The test score was very low. The model does attempt to identify large chi (shown through TP and FP), but often incorrectly identifies small chi (shown through large FP value).





Questions

- Why are we performing clustering using ustat and chi?
    - Normally, we want to cluster using two feature variables, but chi is a target variable.
- How can we overcome overfitting due to scarcity of data, especially for large_chi?
- How is GP regression used on clustered data?