JOINT INSTITUTE
交大密西根学院

VE414 Bayesian Analysis

Prediction Model based on Bayesian Hierarchical Clustering

Group 2
Group Members: Ziyang Chen, Yi Wen, Tianning Zhu
Instructor: Prof. Jing Liu

Problem Statement

In the Forbidden Forrest of Hogwarts, there is a kind of magic tree named Jiuling which is invisible but its fruit Tayes can be observed. Hermione can obtain the information about Tayes with her spell:

- Positions at a 1-minute interval
- Number of Tayes close by (less 1 meter away)
- Number of Tayes no far way (less 3 meters away) (Inaccurate and unknown its accuracy)

This project is to apply the knowledge of Bayesian analysis to build a model to determine the number of Jiuling trees in the Forbidden Forest with the limited data provided.

Concept Generation

Through the data analysis, we find that the number of Tayes only change a little with different time and the spell of different people. Therefore, we assume those factors don't affect the number of Tayes. Then we use original data to plot the locations with parameter "close". We assume the fruit observed in 1*1 grid should be the same set. As a result, we create a new point to represent the set. Finally, we generated our new data.

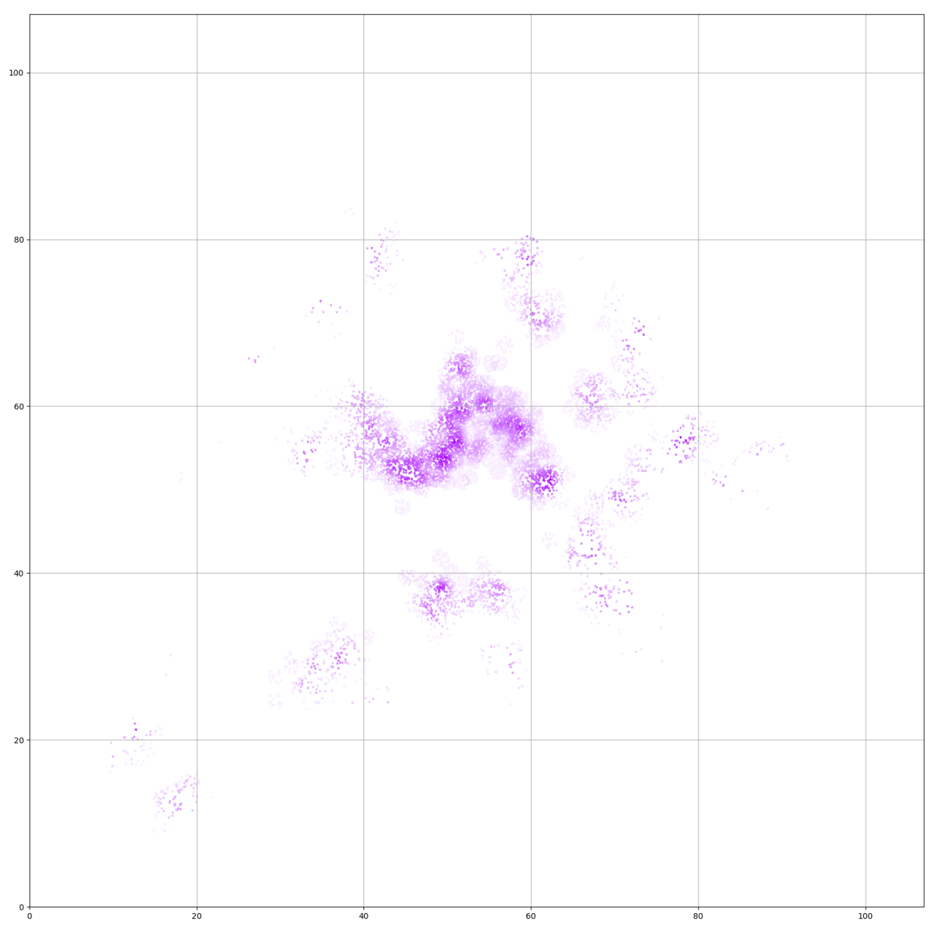

Fig. 1 Original Data

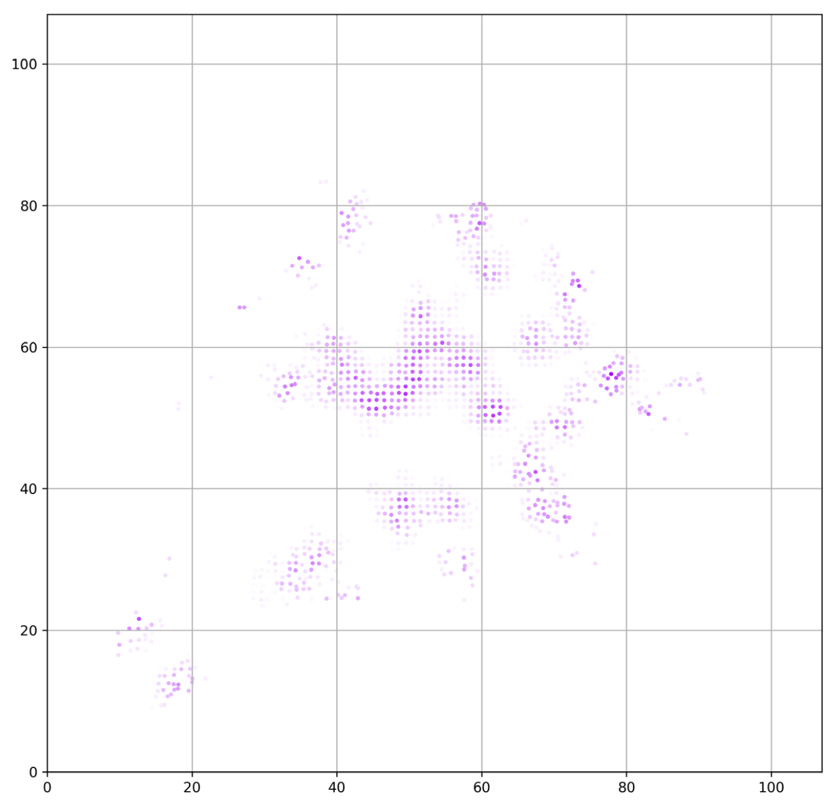

Fig. 2 Preprocessed Data

Design Description

Hierarchical clustering

Hierarchical clustering is one of the most frequently used methods in unsupervised learning. Given a set of data points, the output is a binary tree (dendrogram) whose leaves are the data points and whose internal nodes represent nested clusters of various sizes. The algorithm is iterative: starting from the leaves, it merges a pair of trees at every setp based on some specific criteria.

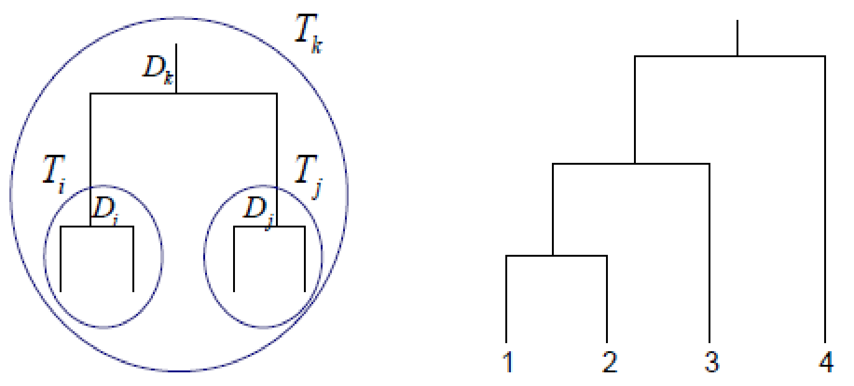

Fig.3 Hierarchical Structure[1]

Bayesian Hierarchical Clustering

H_0^k: all the data in D_k are i.i.d; H_1^k is the alternative hypothesis.

The posterior probability of the merged hypothesis is calculated:

$$r_k = p(H_0^k|D_k) = \frac{p(H_0^k)p(D_k|H_0^k)}{p(H_0^k)p(D_k|H_0^k) + p(H_1^k)p(D_k|H_1^k)}$$

This quantity is used to decide greedily which two trees to merge and is also used to determine which merges in the final hierarchy structure were justified.

Dirichlet Process Mixture Model

We want to simulate the process to draw *Xn* for a distribution *H* and a scaling parameter *α*: With probability *α/(α+n-1)*, *Xn* is drawn from *H*; with probability *nx/(α+n-1)*, *Xn=x* where *nx* is the number of *x* occurring in previous *n-1* times. This is equivalent to draw a distribution *P* from *DP(H,α) and then* draw all *Xi's* from *P* independently. Here, *DP* is a distribution over distributions, called a Dirichlet process. We define $\pi_k \overset{\text{def}}{=} p(H_0^k)$ and it can be derived based on the characteristic of DPM model

```
initialize each leaf $i$ to have $d_i = \alpha$, $\pi_i = 1$
for each internal node $k$ do
    $d_k = \alpha\Gamma(n_k) + d_{\text{left}_k}\, d_{\text{right}_k}$
    $\pi_k = \frac{\alpha\Gamma(n_k)}{d_k}$
end for
```

Validation

We plot our result together with the dataset which is preprocessed by us. And here are the figure. We can find that our tree location has almost the same distribution of the dataset which means our method predict the number and location of trees well.

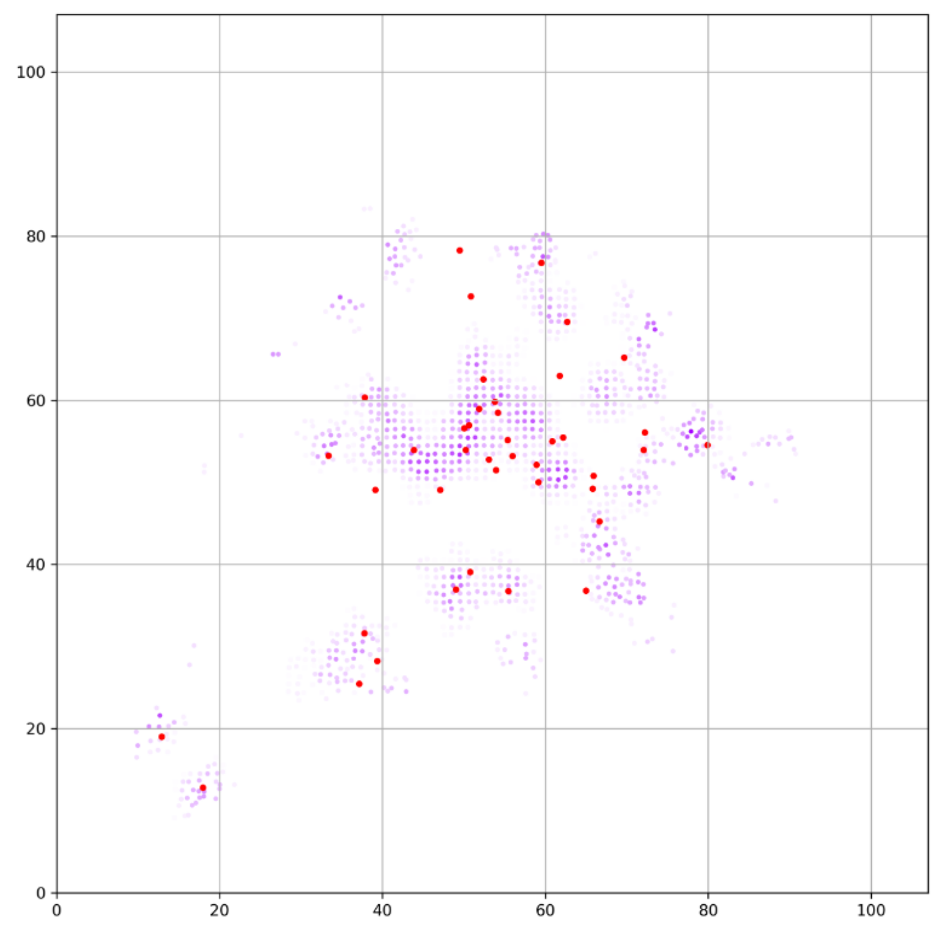

Fig.4 Predicted Location of Jiuling

Conclusion

Bayesian Hierarchical Clustering based on Dirichlet process mixtures is a useful method in unsupervised learning. The key to achieve the goal is to find the suitable r_k. And we have a relatively good result as well as the compute speed.

Reference

[1]https://www2.stat.duke.edu/~kheller/bhcnew.pdf