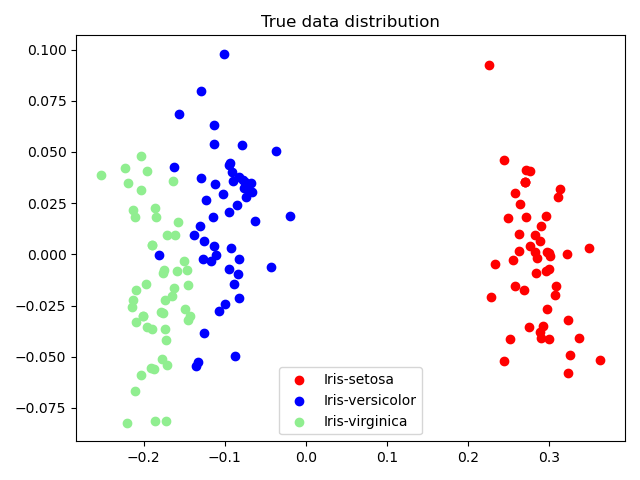
**Clustering**

Для решения задачи кластеризации мною был реализован алгоритм dbscan. Работа алгоритма тестировалась на датасете «Ирисы Фишера». Для уменьшения размерности данных для последующей визуализации был использован алгоритм PCA из пакета sklearn.

Для начала посмотрим, как выглядит реальное распределение датасета:

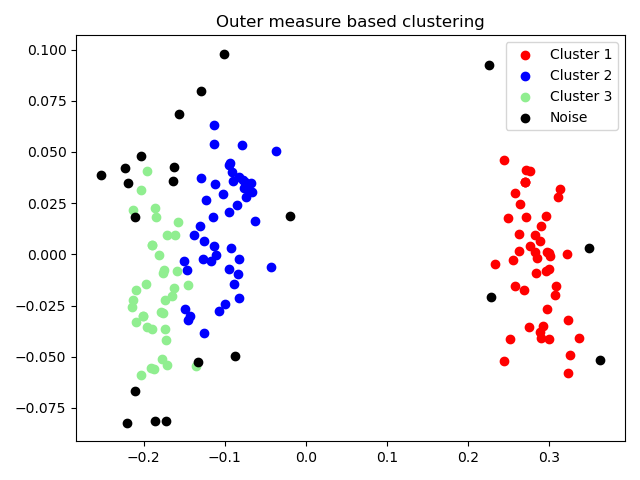


Посмотрим, как отработает наш алгоритм, в том случае, когда мы выбираем параметры на основе внешней меры.

В качестве внешней меры была выбрана Accuracy (или как на викиконспектах Rand). Мы ее максимизировали и по итогу получили такие значения оптимальных параметров:

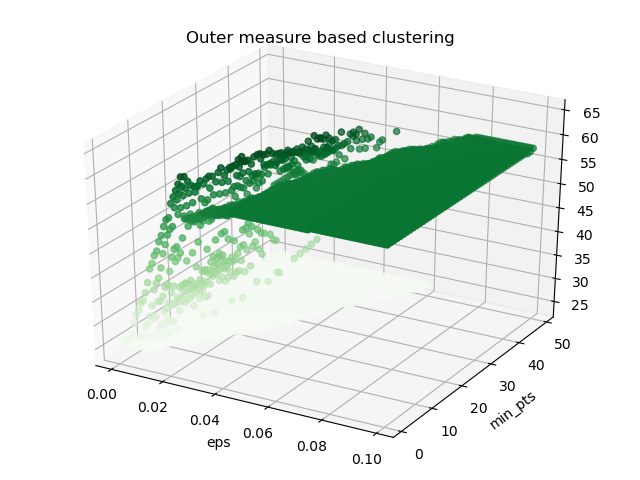
1. eps = 0.046
2. min\_pts = 24

Значение меры равно: 64.13333333333334



Как мы видим первый кластер был выделен практически так же, как в реальных данных. Второй и третий тоже сохранили свою примерную форму, некоторый данные мы пометили как шум.

Посмотрим зависимость нашей меры, от изменения гиперпараметров:

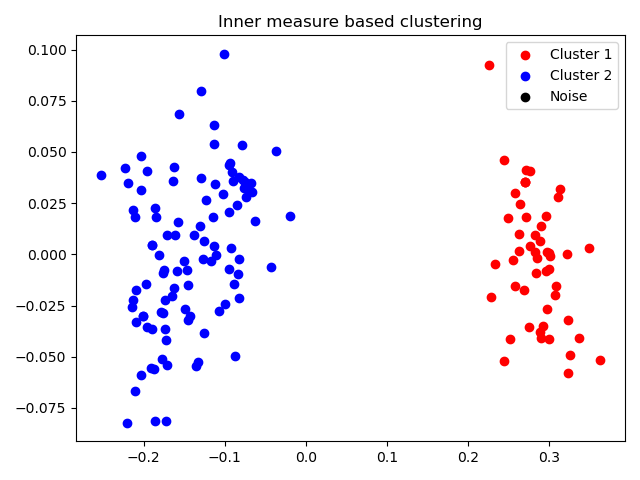


Теперь посмотрим, как отработает наш алгоритм, в том случае, если мы выбираем параметры на основе внутренней меры.

В ее качестве я взял отношение среднего внутрекластерного расстояния и среднего внешнего расстояния в метрическом пространстве. Мы ее минимизировали и по итогу получили такие значения оптимальных параметров:

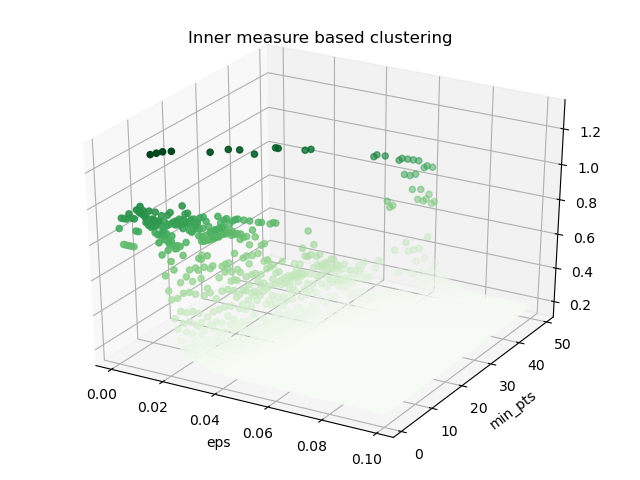
1. eps = 0.072
2. min\_pts = 1

И получили, по итогу, меру равную: 0.18699297460993755



Как мы видим, Первый кластер полностью соответствует исходному распределению. Второй же и третий слились в один, как оказалось, с точки зрения меры это оптимальный вариант. Шума выявлено не было.

Посмотрим зависимость нашей меры, от изменения гиперпараметров:



Подводя итоги, мы видим, что лучший результат показало использование внешней меры, что собственно и не удивительно.