Стабилизация и процессы Дирихле в решении MLBootCamp IV

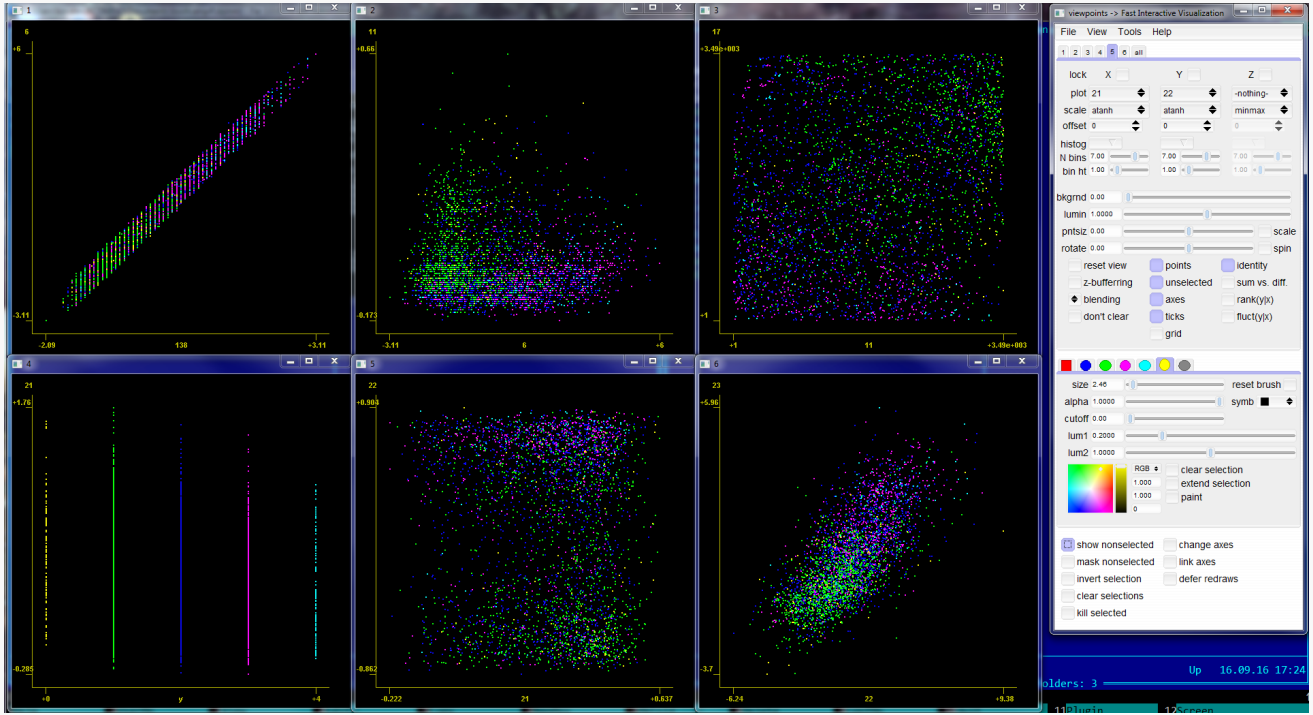
Святослав Ковалёв

Второе место из 563

Данные

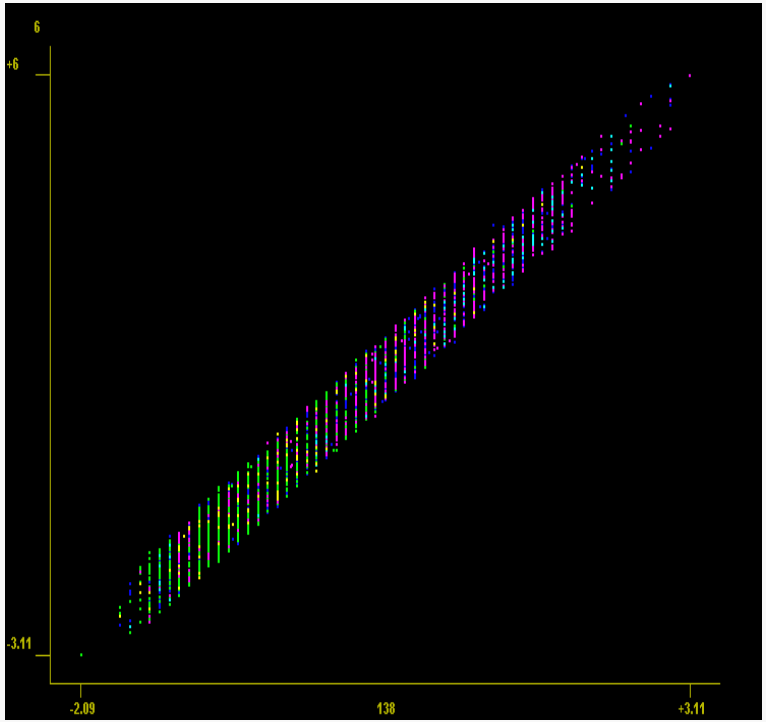
- Датасет с обфусцированными данными на 223 признака
- 3489 точек в train, 2327 точек в test
- Целевая переменная – пять классов
- Метрика – Accuracy

Наблюдения по данным



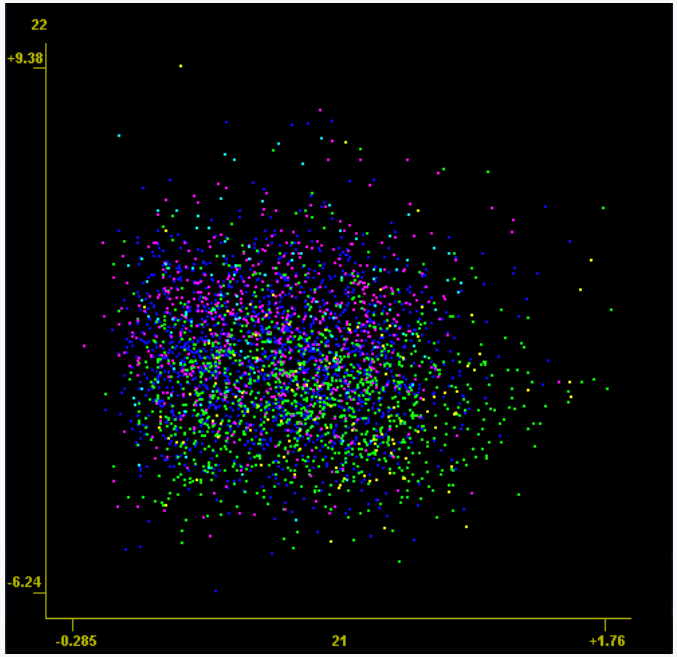
https://software.nasa.gov/software/ARC-16019-1

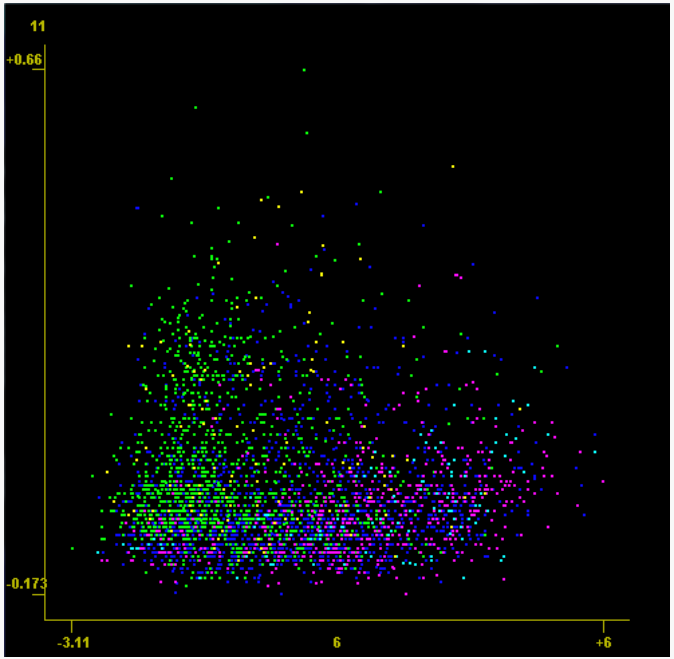
Наблюдения по данным



- Много колонок с высокой корреляцией
- Есть колонки с малым числом уникальных значений
- Классы распределены «последовательно» друг за другом

Наблюдения по данным

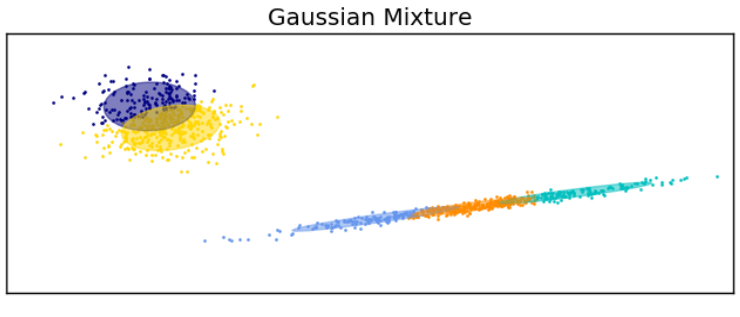




Классы группируются в пересекающиеся «облака»

Можно ли определить параметры этих облаков?

Процессы Дирихле (DPGMM)



Картинка из документации sklearn.mixture.BayesianGaussianMixture

DPGMM – Dirichlet Process Gaussian Mixture Model

Способ представить данные как сэмпл из смеси распределений. Результат DPMM – взвешенная смесь распределений с известными параметрами.

Хороши на мелких датасетах

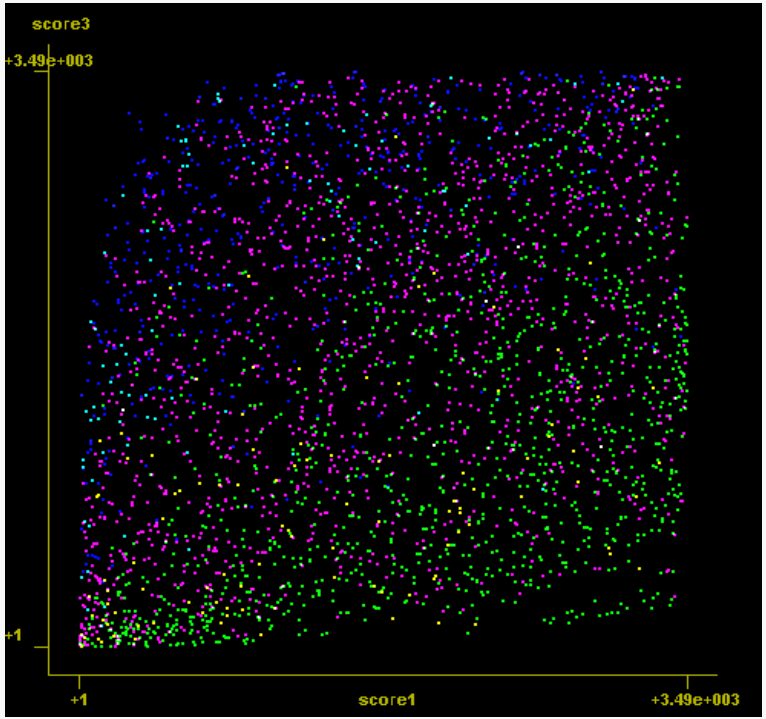
Процессы Дирихле (DPGMM)

- Берём точки принадлежащие одному классу
- Обучаем на них DPGMM
- Оцениваем для каждой точки loglikelihood по обученной модели
- Повторяем для всех классов
- Получаем пять колонок новых признаков

Важные замечания

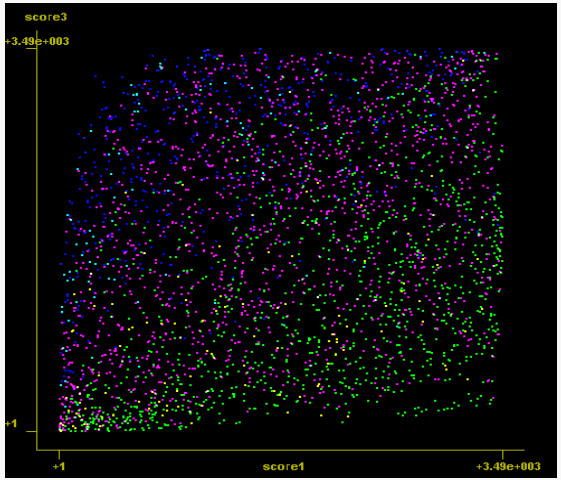
- Output модели – это **не вероятность** принадлежности к классу
- OOF не нужен
  - И без того сложно переобучиться
  - С OOF на малых данных будут получаться слишком разные модели и предсказания разных масштабов
  - Я всё равно проверил

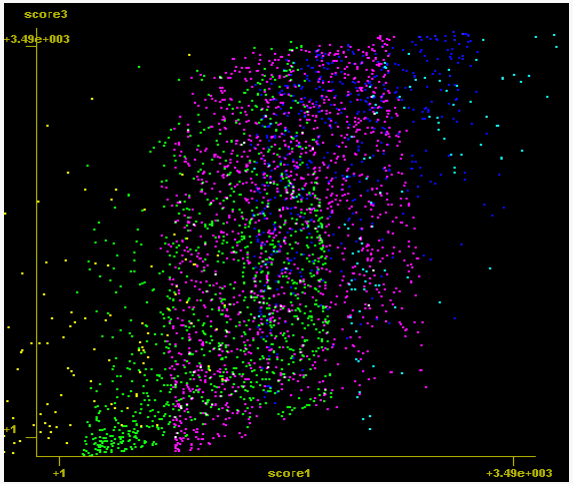
Сгенерированные из DPGMM признаки

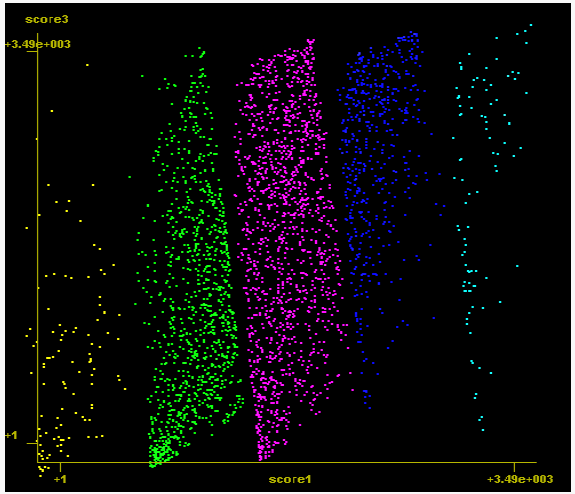


Если строить попарные распределения правдоподобия «противолежащих» классов, то кажется, что классы разделяются.

Сгенерированные из DPGMM признаки







Не тут то было! Класс 2 везде.

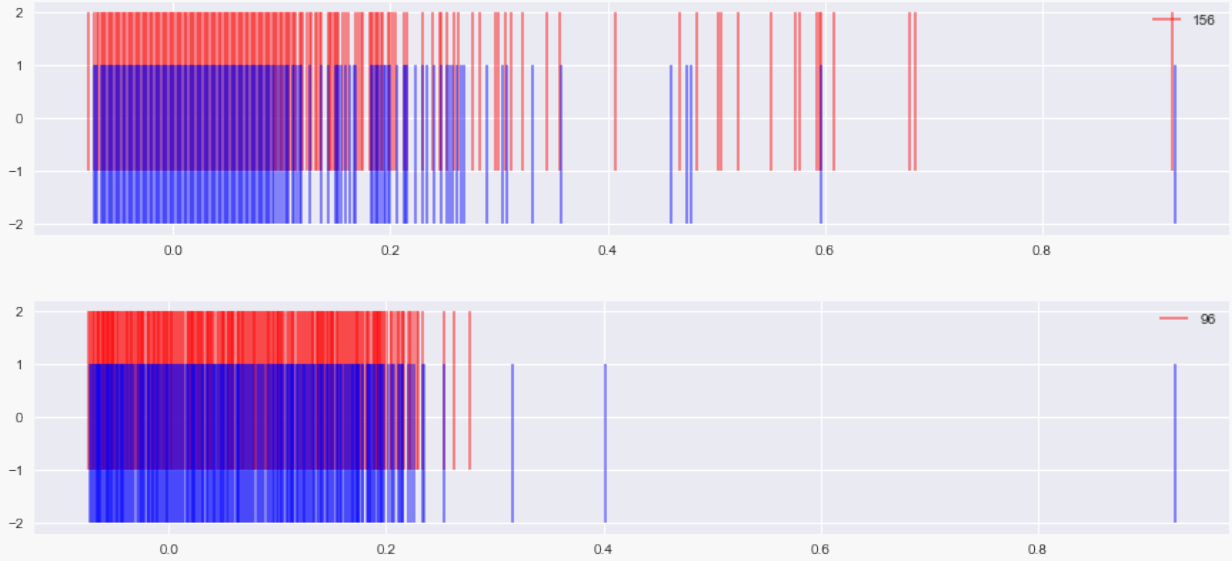
Что дальше

- Только на признаках из DPGMM точность 0.63 (logreg, rf)
- Вместе с исходными признаками 0.65-0.67 (rf, et, voting)
- Если брать просто максимум правдоподобия 0.57-0.6 (для мультикласса так делать нельзя)

Другие признаки

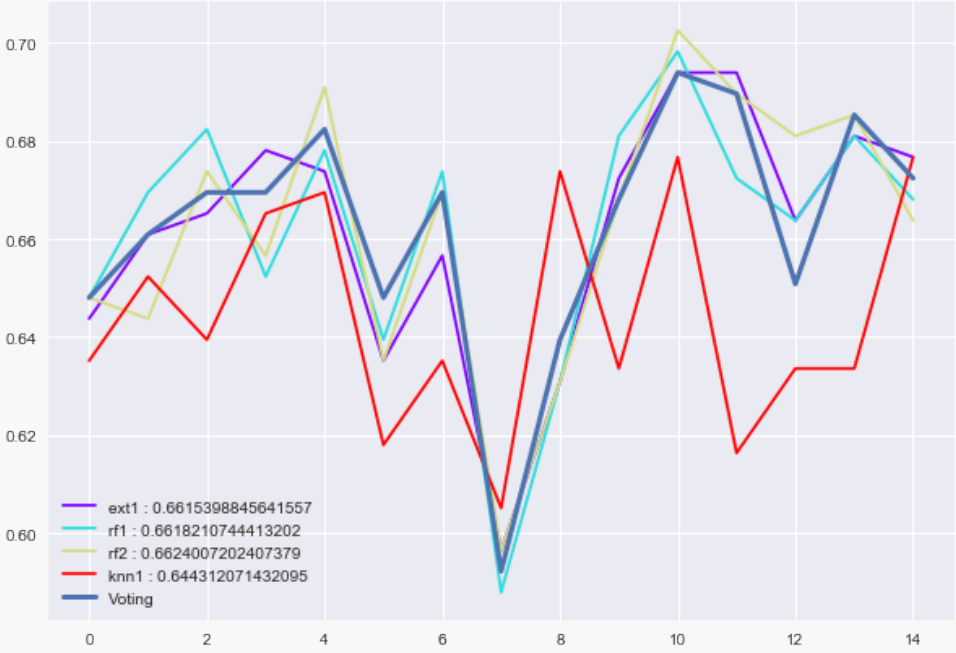
- DPGMM (по колонке на каждый класс)
- Isolation Forest (outlier score)
- Колонка с шумом
- Полином второй степени

Isolation Forest и ранжирование



Хвосты распределений слишком разрежены
Сильно отличаются на обучающей и тестовой выбрках
Accuracy заставляет бороться за каждую точку

Валидация



Распределение признаков на разных фолдах даёт информацию о том, чего можно ожидать на private LB.

Много сил ушло на укрощение дисперсии на фолдах.

Стабилизация модели

Скор ужасно скачет от изменения сидов?

- Голосование моделей с разными сидами
- Или поиск золотого сида?

```
from sklearn.model_selection import GridSearchCVSearchCV

param_map = {
    'random_state': np.arange(1, 2000, 1)
}
```

Прочие фокусы

- Ранжирование признаков
- Удаление классов (0 => 1, 4 => 3)
- Сглаженное голосование
- Стэкинг

Финальная модель

Сглаженное голосование:

Пара лучших решений на исходных признаках (разные наборы моделей) + всё вышеописанное + решение на стэкинге

0.683 public -> 0.661 private

Лучшая модель (голосование голосований + шум)

0.666 public -> 0.668 private

Заключение

- Фиксируйте сиды
- Смотрите на данные
- Не переобучайтесь

@iggisv9t