Мультимоделирование как универсальный способ описания выборки общего вида

Роман Алексеевич Логинов

Московский физико-технический институт

Курс: Численные методы обучения по прецедентам (практика, В. В. Стрижов), Группа 694, весна 2019
Консультант: А. А. Адуенко

Цель исследования

Задача

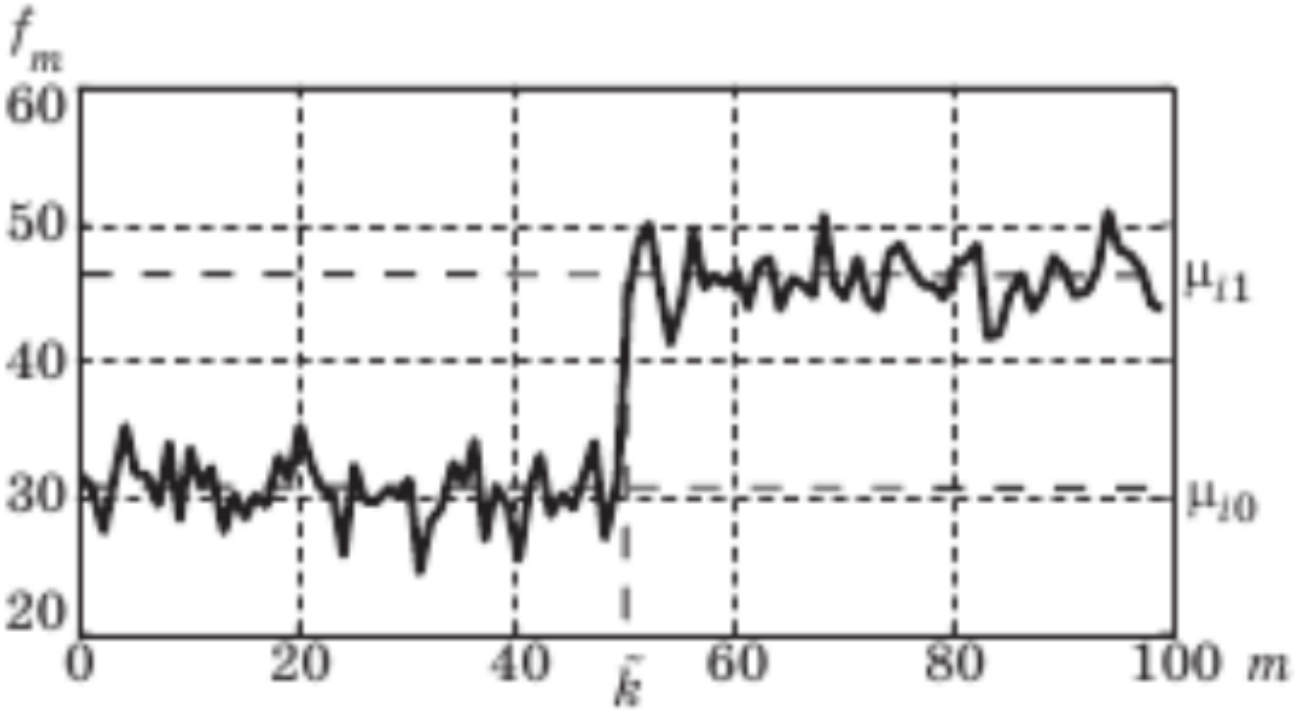
По заданной мультимодели, а также новому набору данных, понять, каким образом дообучать мультимодель.
Детектировать изменение зависимости в данных во времени в задаче бинарной классификации.

Требования к моделям

- Статистически отличимы друг от друга
- Обучены на соответсвующих подвыборках
- Параметры изменяются при получении новых объектов

Примеры временной зависимости

- Разная вероятность одобрения кредита с одинаковой зарплатой 10 лет назад и сейчас. Возможно использование мультимодели.
- Ежедневная прибыль магазина. Закономерность меняется при появлении конкурентов.



Теоретическая база

Ближайшая работа

Адуенко А.А.
Выбор мультимоделей в задачах классификации, МФТИ, 2017

- Алгоритм построения мультимоделей в частных случаях
- Количественная оценка различимости моделей
- Отсутствие временной структуры

Теоретические сведения

Bishop Pattern recognition and machine learning
Motrenko, Strijov Sample size determination for logistic regression
Ge, Jiang On Consistency of Bayesian Inference with Mixtures of logistic regression

Обозначения: одиночный классификатор

Дано:

$$\mathfrak{D} = \{(\mathbf{x}_i, y_i)\}, i \in \mathcal{I} = \{1, ..., m\}$$

- обучающая выборка
$\mathbf{x}_i \in \mathbb{X} \subset \mathbb{R}^n$ - признаковое описание объекта
$y_i \in \{+1, -1\}$ - корректная метка класса

Вероятностный классификатор:

$$p(y, \mathbf{w}|\mathbf{x}, \mu) = \sigma(y\mathbf{w}^\top \mathbf{x})p(\mathbf{w}|\mu),$$

где $\mathbf{w}$ - параметр модели, получаемый при обучении, μ - гиперпараметры

Временная структура:

$$t(X_i) < t(X_j)\ \forall i < j,\ \ i, j \in \mathcal{I}$$

Дополнительные данные:
Разбиение выборки на множества

$$\mathcal{I} = \mathcal{I}_1 \sqcup ... \sqcup \mathcal{I}_K,$$

правила кластеризации объектов:

$$\mathbb{R}^n = \mathbb{X}_1 \sqcup ... \sqcup \mathbb{X}_K,$$

гиперпараметры распределения весов $\mathbf{A_1}, ..., \mathbf{A_k}$.
Сводится к независимому обучению K логистических регрессий.
В итоге построен новый алгоритм именно для многоуровневой модели.

Смесь моделей

Аналогично смеси распределений, для каждого объекта выбирается модель генерации вероятностно.

Распределение на индексах моделей π обучается вместе с весами.

Соответствующий структурный параметр μ - параметр распределения Дирихле.

В основе обучения - вариационная нижняя оценка и **EM-алгоритм**.

Оценка качества модели

Требования

- В каждый момент времени обучена мультимодель зафиксированного вида.
- Предсказания на новом объекте получаются лишь при обучении на предыдущих.
- Модель перестраивается в зависимости от времени.

Пусть f_{mult} – полученный классификатор. Цель – отыскать

$$\arg\max AUCROC(\mathfrak{D}, f_{mult}(\mathbf{w}))$$
$$s.t.\ \mathbf{w} = \mathbf{w}(t).$$

Описание алгоритма

Выборка – последовательные во времени блоки размера C.
C – гиперпараметр, который требуется подбирать отдельно.

Пусть f – имеющийся классификатор на момент времени T, G_i – новый блок данных заданного размера.

Алгоритм

1. Построить мультимодель f_0, обучив её на данных из G_i;
2. Вычислить *s-score* для оценки различимости f и f_0;
3. Если различимость не отвергается, то дообучить f на данных из G_i;
4. Если отвергается, заменить f на f_0.

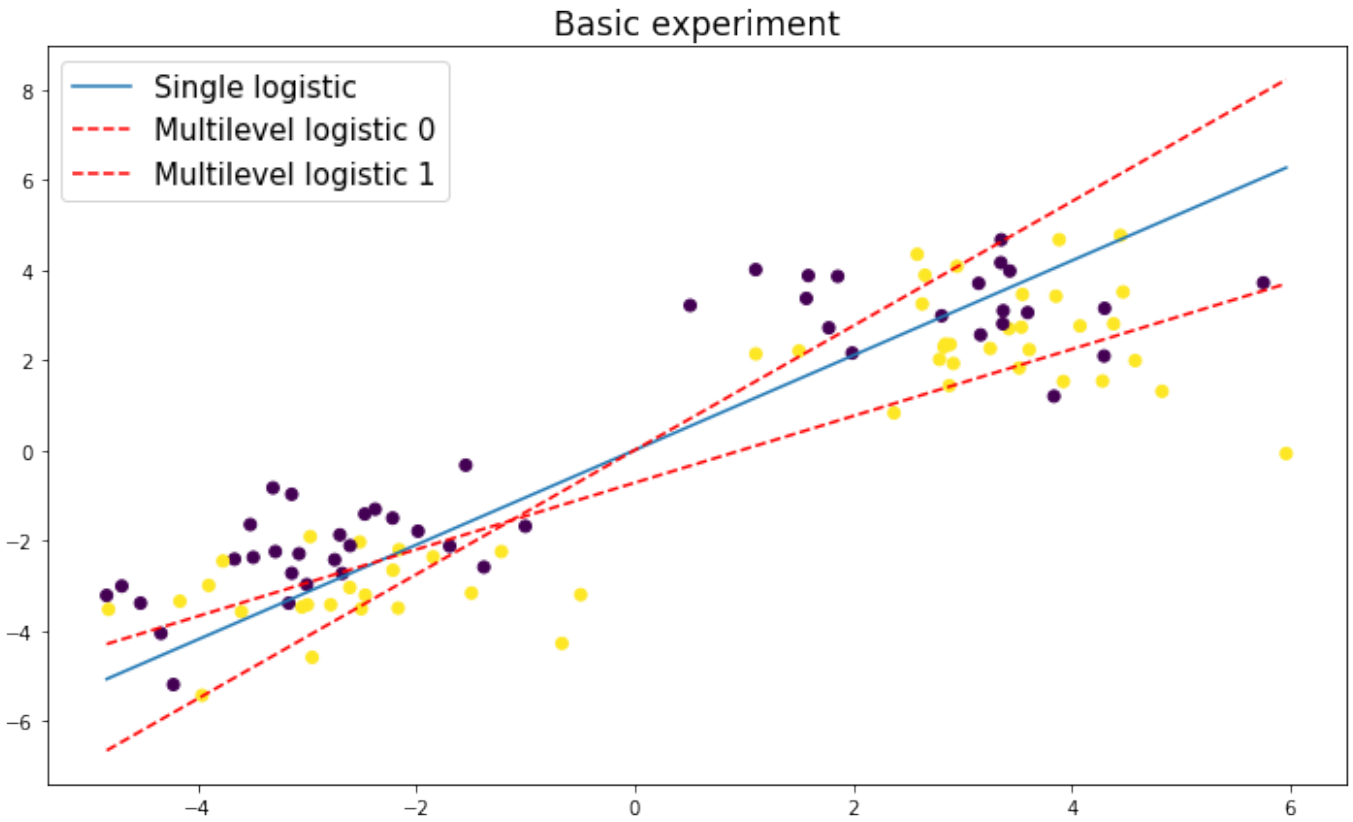
План эксперимента

Для описанных моделей

- Проверить работу базового алгоритма в случае неоднородности данных
- Исследовать поведение критерия различимости моделей
- Понять, какой размер выборки необходим для разлечения моделей на двух разных кластерах
- Сгенерировать датасет с изменяющимися во времени параметрами
- Протестировать на этих данных описанный алгоритм

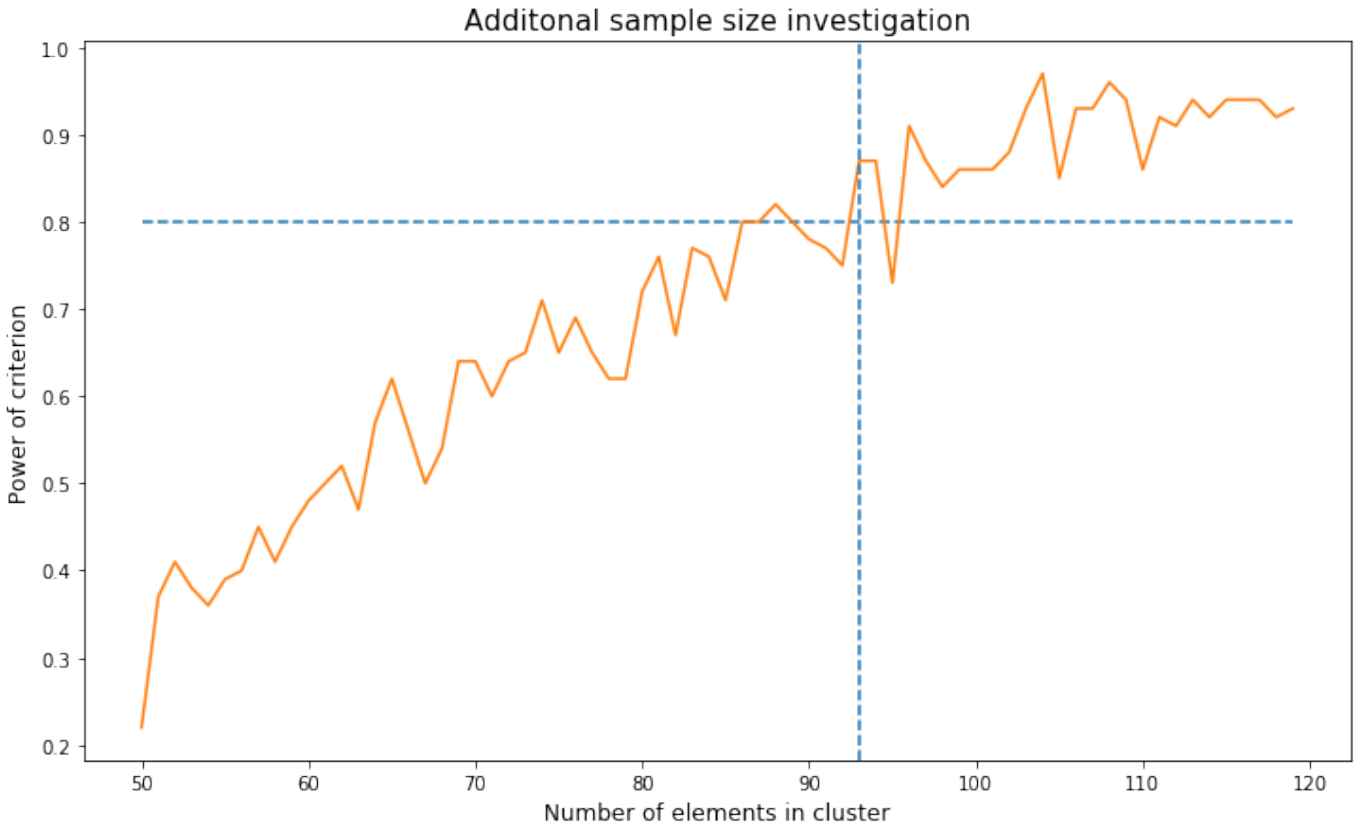
Базовый алгоритм

Простой случай мультимодели с очевидной неоднородностью



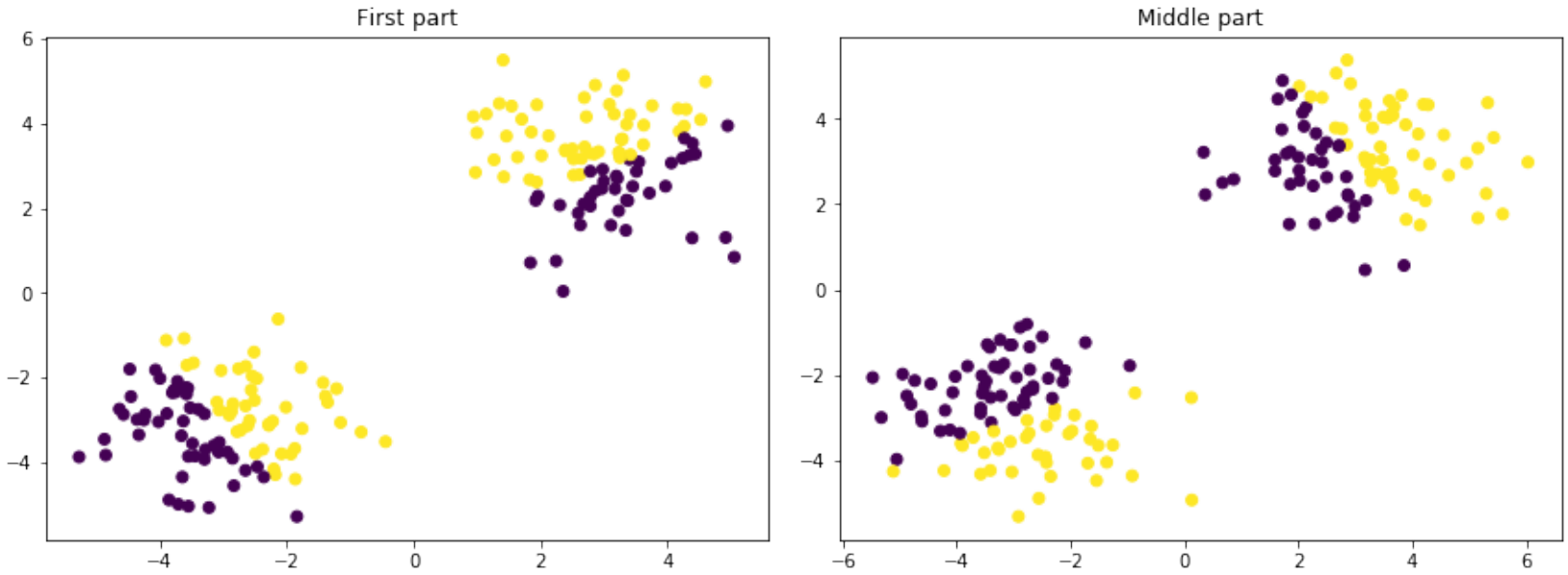
Мощность критерия

Гипотеза H_0: модели, обученные на различных данных, неразличимы между собой (гипотеза близости).
Основная идея – применение s-score как функции близости.

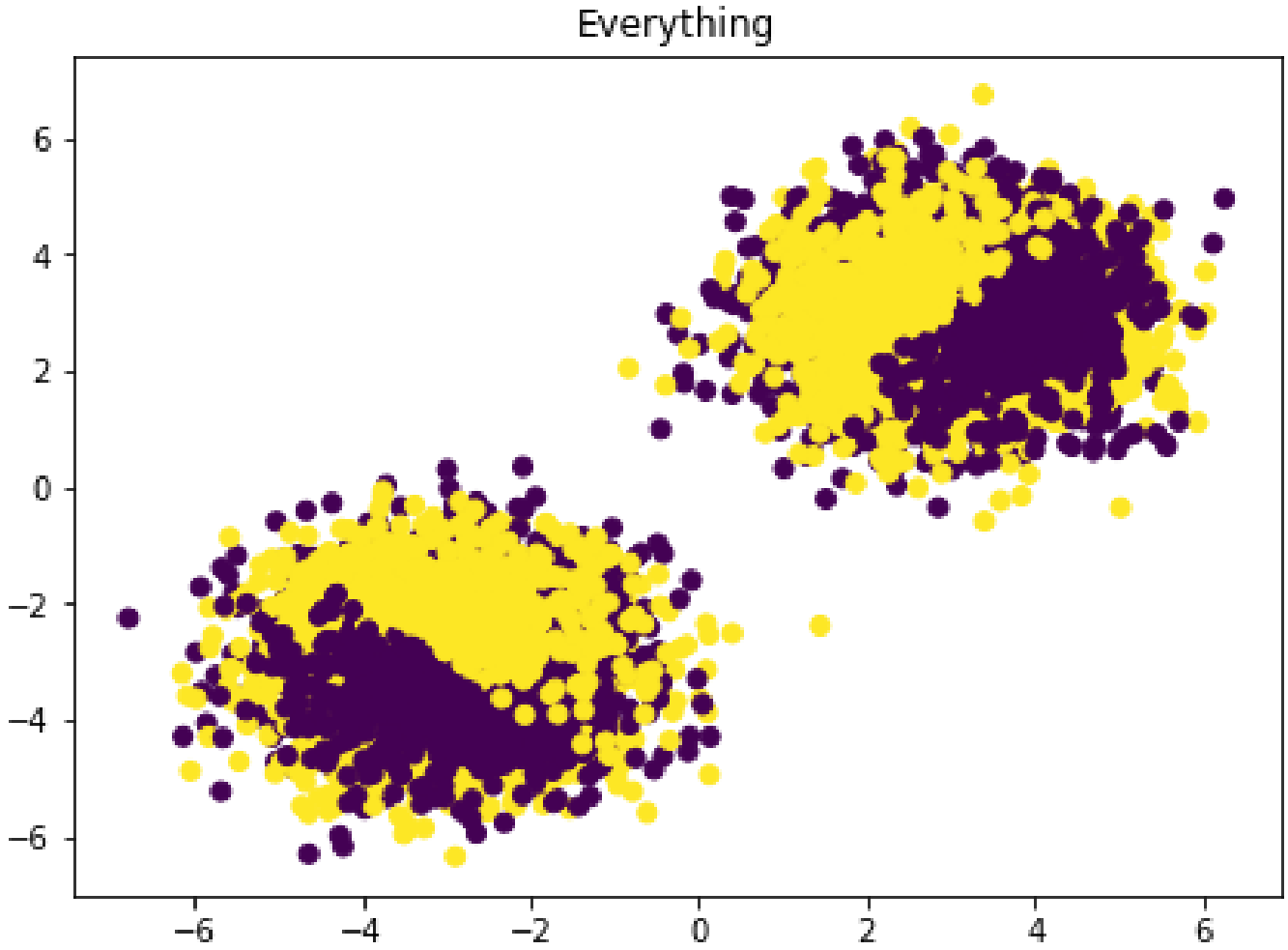


Генерация датасета

2 кластера, в которых разделяющие прямые поворачиваются через равные промежутки времени

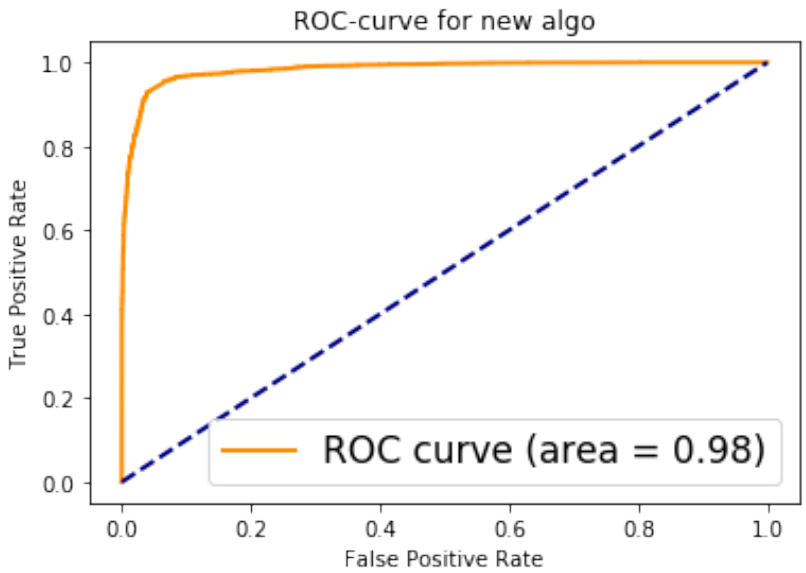


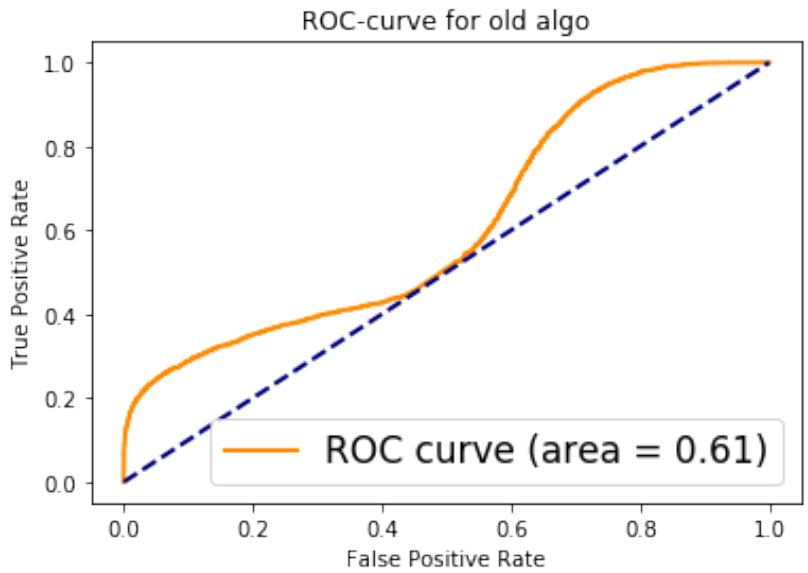
Итоговый датасет



Результат

С учетом того, что параметр C подобран оптимально, получаем значимые улучшения в AUC-ROC.





Заключение

Итоги:

- Подобран размер выборки для разделения моделей на кластерах с заданным различием весов
- Сгенерирован датасет с изменением во времени
- Реализован алгоритм, учитывающий подобную структуру
- Проведено сравнение с имеющимся ранее алгоритмом

Дальнейшее исследование:

- Сравнить работу алгоритмов на реальных данных из UCI
- Найти метод подбора размера блока
- Смоделировать временную структуру при помощи гауссовского процесса