Иллюстрации Positive-Unlabeled learning

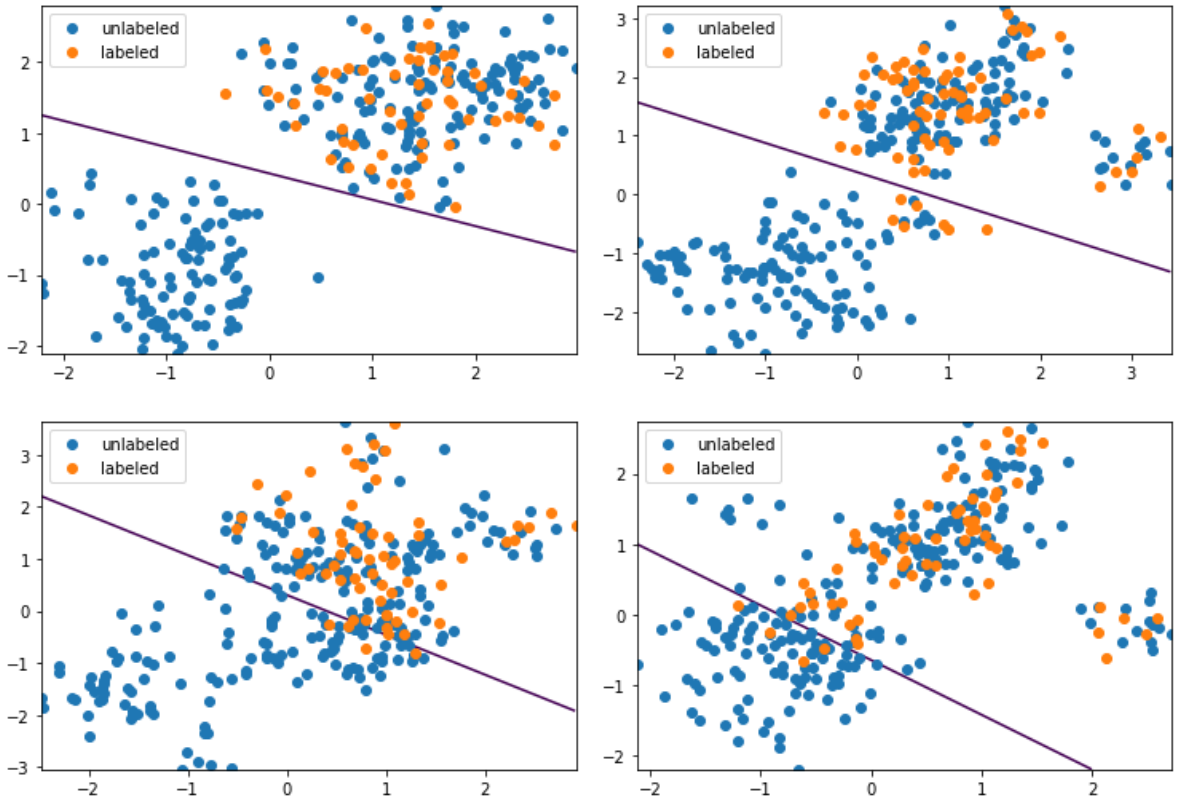
Мария Давыденкова

5 апреля 2020 г.

Датасет здесь представляет собой смесь гауссиан positive и negative точек. Некоторые positive точки отмечены. Предполагается, что вероятность того, что positive точка отмечена, постоянна. В данном случае она равна 0,3.

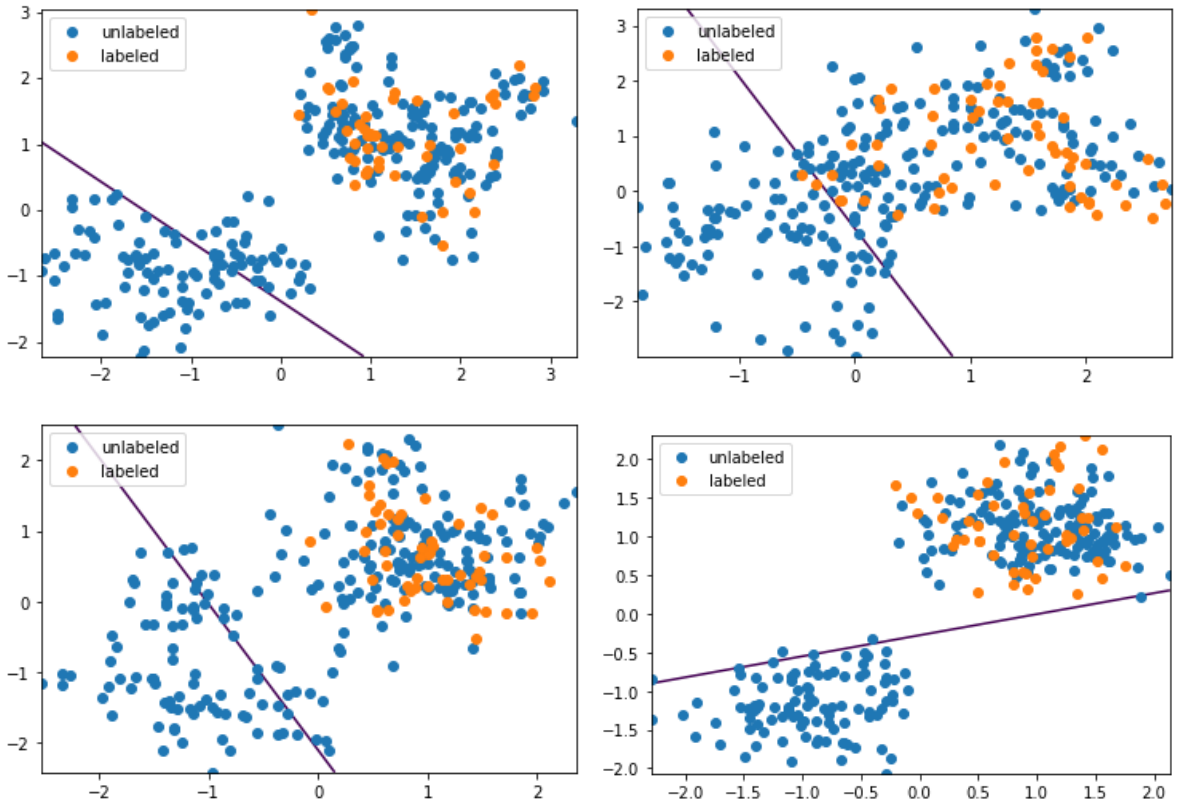
1 Алгоритм, использующий классический классификатор на нестандартных данных

Некоторые изображения разделяющей прямой, полученные алгоритмом:



2 Алгоритм, использующий взвешенные неотмеченные точки

Некоторые изображения разделяющей прямой, полученные алгоритмом:

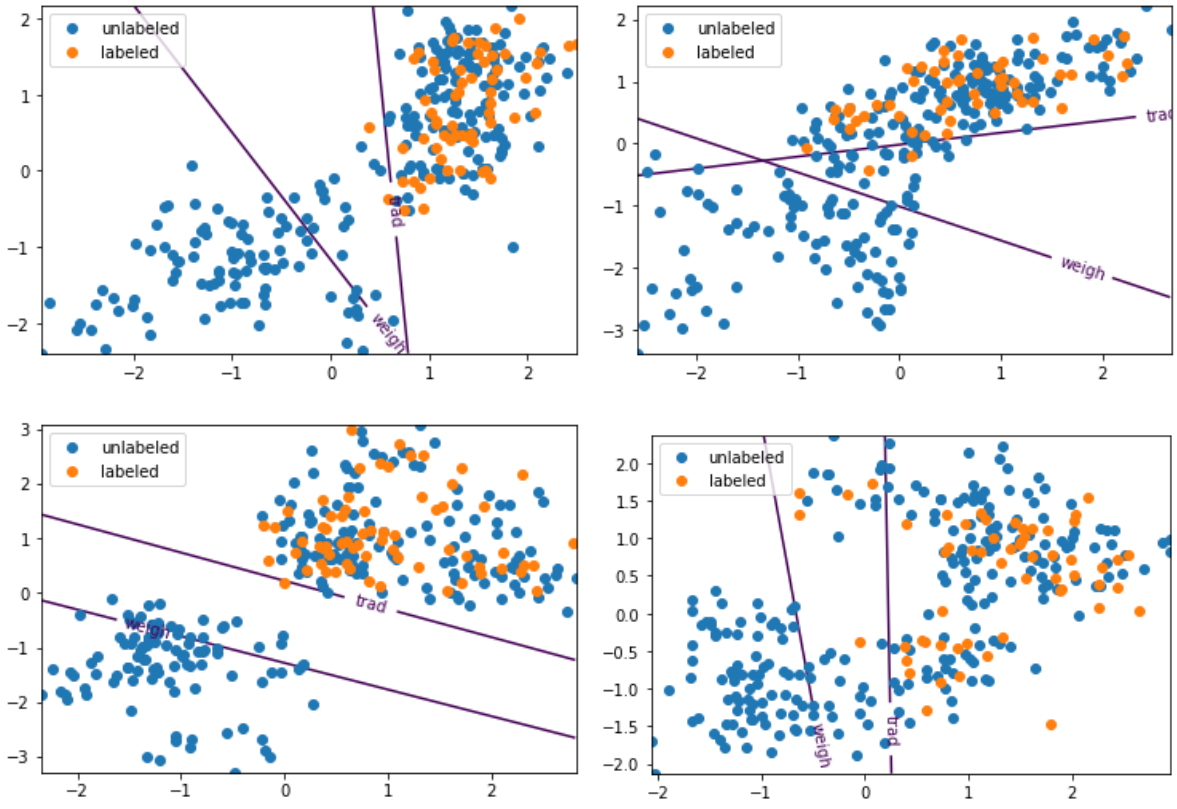


Хочется отметить, что иногда алгоритм выдает неожиданные результаты. Кажется, что в целом он считает точку положительной с большей вероятностью, чем предыдущий алгоритм. Например, на левой верхней картинке прямая проходит примерно через центр синего скопления, хотя поблизости нет отмеченных точек.

Также интересно отметить эффект, наблюдаемый на правой нижней картинке: данные разделены хорошо, но нам хочется верить, что угол наклона прямой должен быть другим, проходить дальше от нарисованных точек.

3 Сравнение двух алгоритмов

Сравним эти алгоритмы на одинаковых датасетах. Здесь trad обозначает прямую, построенную первым алгоритмом, а weigh – прямую, построенную вторым алгоритмом.



Интересно обратить внимание на области, заключенные между двумя прямыми. Наблюдателю зачастую непонятно, к какому классу стоит их отнести. Кажется, что истина кроется где-то посередине. Возможно, даже среднее арифметическое увеличит точность предсказания. А можно найти параметры выпуклой комбинации двух прямых, при которых картинка получится оптимальной.