Индустриальный семинар

Look-alike моделирование

Элен Теванян

Руководитель направления алгоритмического анализа, X5 Tech

x5retail.tech

Nota Bene

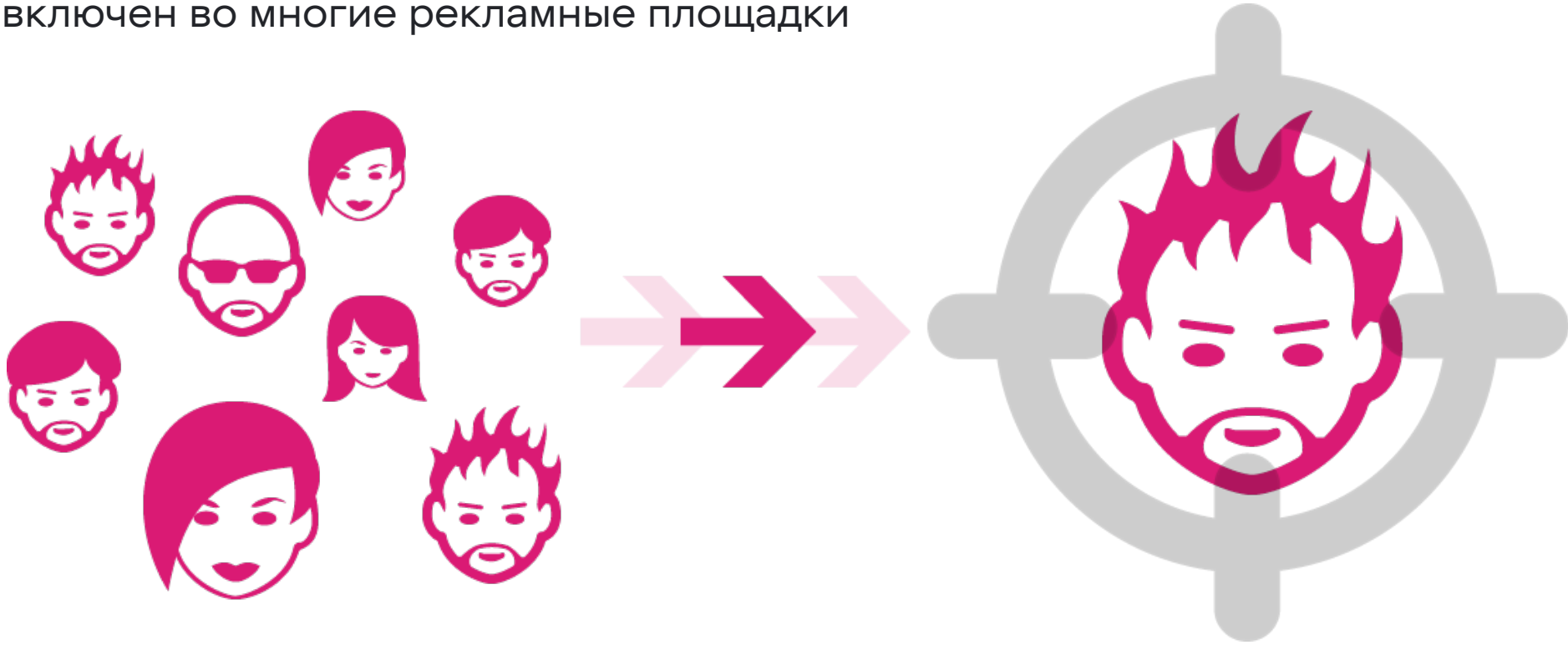
- Обсуждаем третью третью большую сюжетную арку
- Как модель для обучения подразумеваем любую supervised модель
- Таргет, категорийный или численный, определяется задачей и ее постановкой

«Ах, если бы найти похожих»

- Есть узкий сегмент, хочется его увеличить
- Например, Перекресток выходил на кейс-чемпионат с запросом увеличить аудиторию клуба «Здоровых привычек»
- Допустим, всего участников 10 000. Активная клиентская база – несколько млн. Наверняка есть сегмент(-ы), кто мог бы присоединиться к клубу, но, мб, не знает о нем/не думал/ не видел

Look-alike моделирование

- «Можно же лукэлайки построить, это же легко!»
- Задача: найти похожих пользователей из известного сегмента
- Функционал включен во многие рекламные площадки



«Ах, если бы найти похожих»

- Есть узкий сегмент, хочется его увеличить
- Например, Перекресток выходил на кейс-чемпионат с запросом увеличить аудиторию клуба «Здоровых привычек»
- Допустим, всего участников 10 000. Активная клиентская база – несколько млн. Наверняка есть сегмент(-ы), кто мог бы присоединиться к клубу, но, мб, не знает о нем/не думал/ не видел

- Как сделать?

Рабочий вариант 1:

Свести к Supervised-задаче

- Есть X людей с положительными метками
- Можем насэмплировать Y людей с отрицательными метками
- => классическая задача классификации

Сложности?

Рабочий вариант 1:

Свести к Supervised-задаче

- Есть X людей с положительными метками
- Можем насэмплировать Y людей с отрицательными метками
- => классическая задача классификации

Сложности?

- Легко скатиться к несбалансированной задаче
- Отрицательные метки – не факт, что отрицательные

Рабочий вариант 2:

Свести к Uplift-моделированию

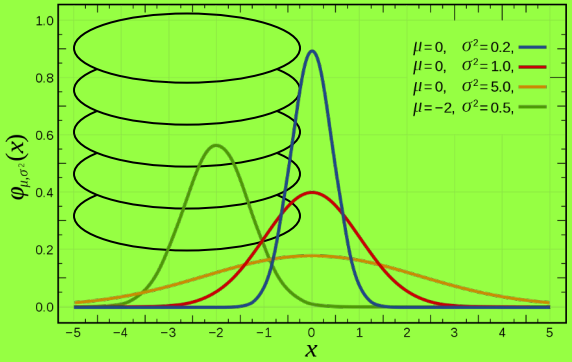
- Мы хотим, чтобы пользователи кликнули на рекламу/активировали участие
- А также хотим, чтобы они участвовали в кампаниях/акциях/клубах
- Сведем-ка задачу к аплифту – будем искать тот сегмент, который положительно откликнется на коммуникацию

Рабочий вариант 3:
One Class Learning

- Классическая задача классификации предполагает наличие двух классов, где модель пытается разделить данные разных классов
- В 1996 году представили понятие One-class classification – обучении модели только на одном класе данных
- Примеры такого подхода не только в LAL, но и:
  - Детекция аномалий
  - Поиск выбросов
  - Обнаружение новизны
- Частные примеры: аварии в двигателях, критические ситуации на АЭС, поломки в нефтяных скважинах и т.п.

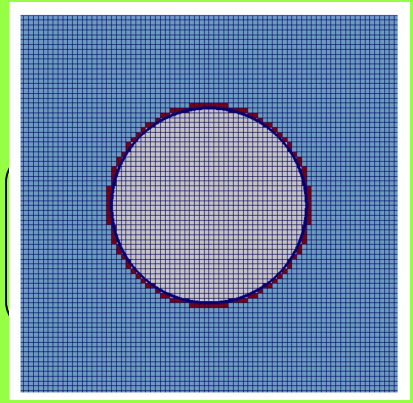
One Class Learning

Три подхода



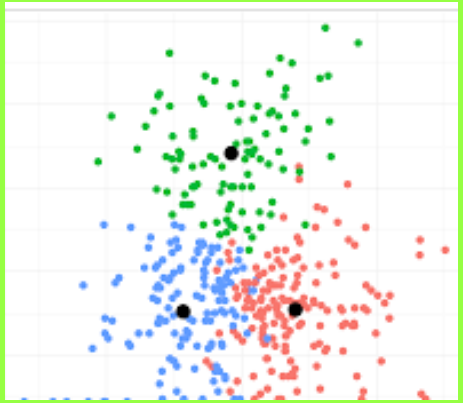
Density Methods

Gaussian Mixtures, Parzen Density



Boundary Methods

K-centers, Nearest Neighbor, SVDD



Reconstruction Methods

K-means Clustering, Learning Vector Quantization, Self-Organizing Maps

One Class Learning

SVM

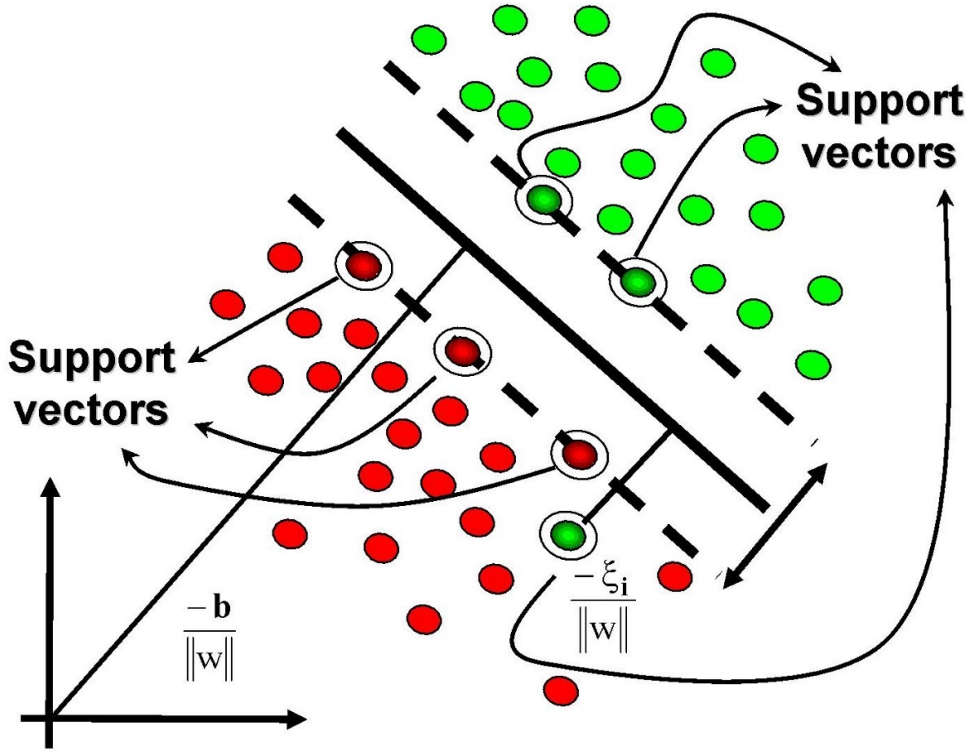
- Классический SVM проводит гиперплоскости, чтобы разделить данные одного класса от другого

$$\min_{w,\,b,\,\xi_i} \frac{\|w\|^2}{2} + C\sum_{i=1}^{n}\xi_i$$

subject to:

$$y_i(w^T\phi(x_i) + b) \geq 1 - \xi_i \qquad \text{for all } i = 1, \ldots, n$$

$$\xi_i \geq 0 \qquad \text{for all } i = 1, \ldots, n$$



One Class Learning

Support Vector Data Description

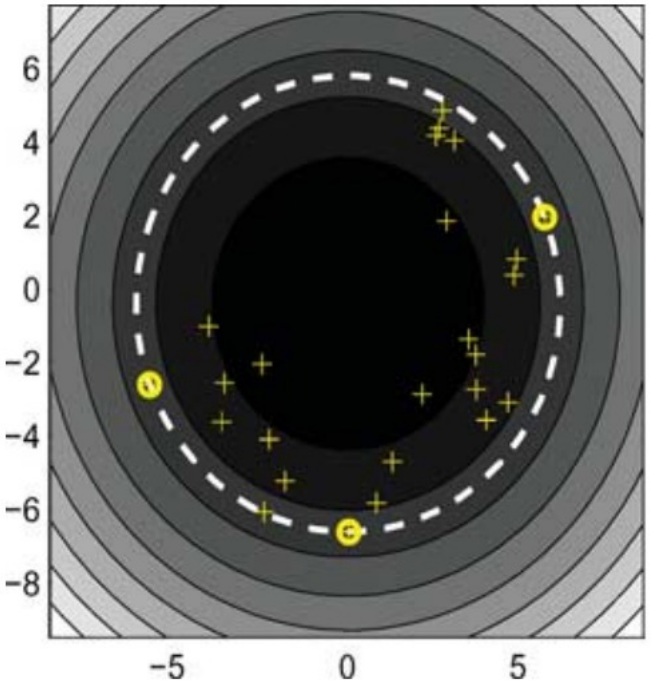
- Алгоритм восстанавливает сферу вокруг данных. Минимизируется объем сферы

$$\min_{R,\,\mathbf{a}} R^2 + C \sum_{i=1}^{n} \xi_i$$

subject to:

$$\|x_i - \mathbf{a}\|^2 \leq R^2 + \xi_i \qquad \text{for all } i = 1, \ldots, n$$

$$\xi_i \geq 0 \qquad \text{for all } i = 1, \ldots, n$$



One Class Learning

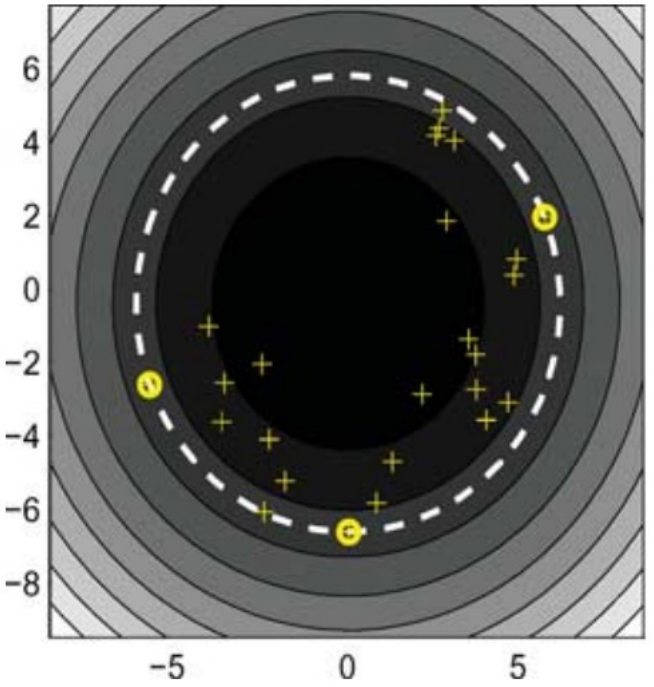
Support Vector Data Description

- Если вдруг есть какое-то количество представителей альтернативного класса, можно применить и другую постановку задачи:

$$F(R, \mathbf{a}, \xi_i, \xi_l) = R^2 + C_1 \sum_i \xi_i + C_2 \sum_l \xi_l$$

and the constraints

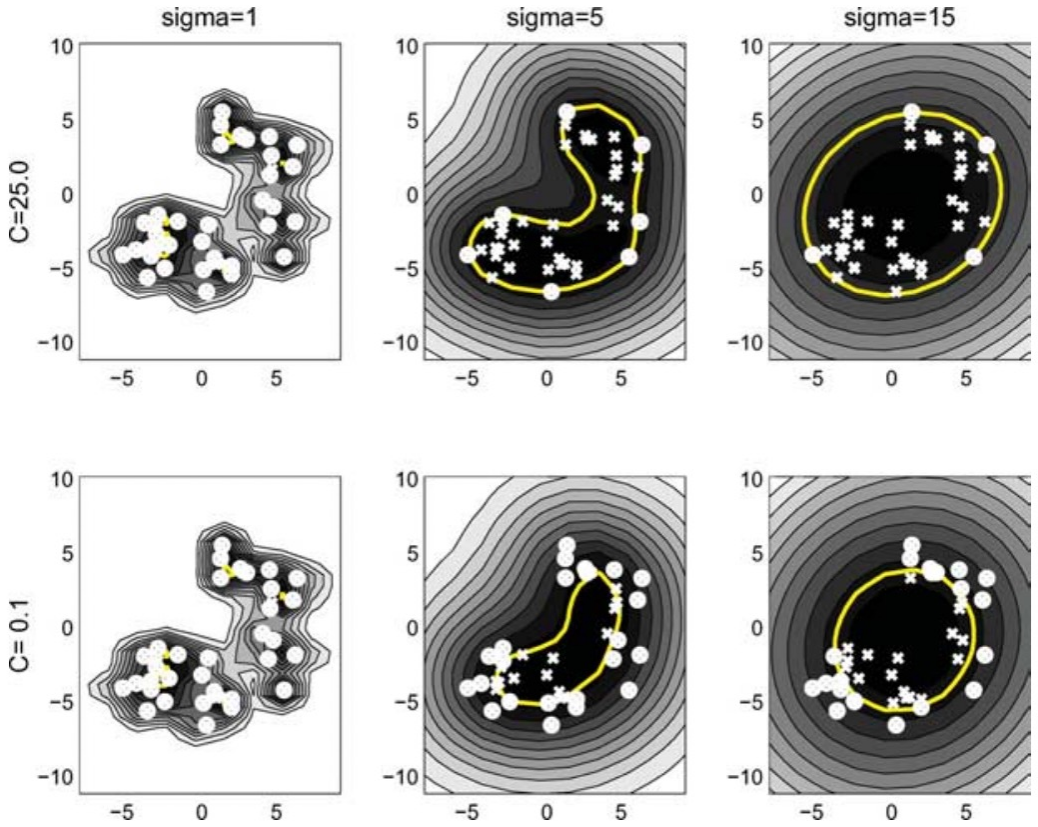
$$\|\mathbf{x}_i - \mathbf{a}\|^2 \leq R^2 + \xi_i, \quad \|\mathbf{x}_l - \mathbf{a}\|^2 \geq R^2 - \xi_l, \quad \xi_i \geq 0, \, \xi_l \geq 0 \quad \forall i, l$$



One Class Learning

Support Vector Data Description

- Kernel Trick тоже имеет место быть



Практические заметки

- Если возможно уйти от задачи LAL к Uplift – уходите
- Если надо делать LAL, выбор чаще в пользу классификации

Ссылки

- Целый диплом по One Class Classification
  https://homepage.tudelft.nl/n9d04/thesis.pdf
- SVDD:
  https://homepage.tudelft.nl/a9p19/papers/ML_SVDD_04.pdf
- Deep One Class Classification
  https://medium.com/analytics-vidhya/paper-summary-deep-one-class-classification-doc-adc4368af75c
- Gaussian Mixture Models

https://towardsdatascience.com/gaussian-mixture-models-explained-6986aaf5a95

X5 Tech

18 октября 2022