Задача частичного обучения

Евгений Борисов

Задача частичного обучения

способы организации данных и типы задач ML

supervised learning
- размеченный датасет {*X,target*}, (регрессия, классификация)

unsupervised learning
- НЕразмеченный датасет *{X}* (кластеризация)

semi-supervised learning
- частично размеченный датасет {*X,target,X'*}, (трансдуктивные модели)

Задача частичного обучения

Дано: множество объектов X, множество классов Y ;

из них только ℓ размеченных

$X\ell$ = {x1, . . . , xℓ } ; { y1, . . . , yℓ } - размеченная выборка (labeled data);

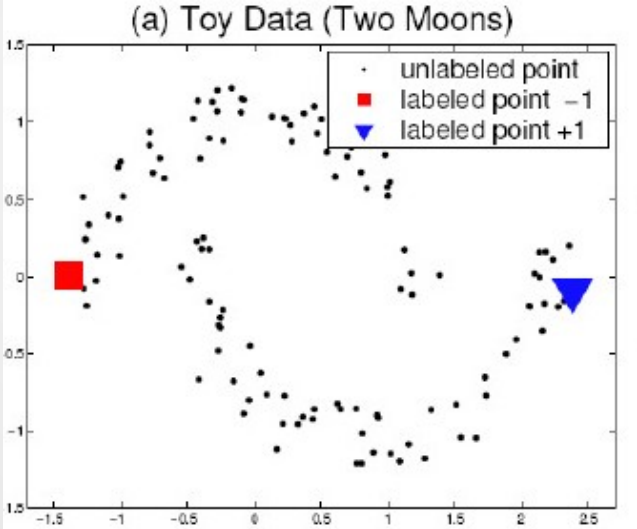
Xk = { xℓ+1, . . . , xℓ+k }— неразмеченная выборка (unlabeled data).

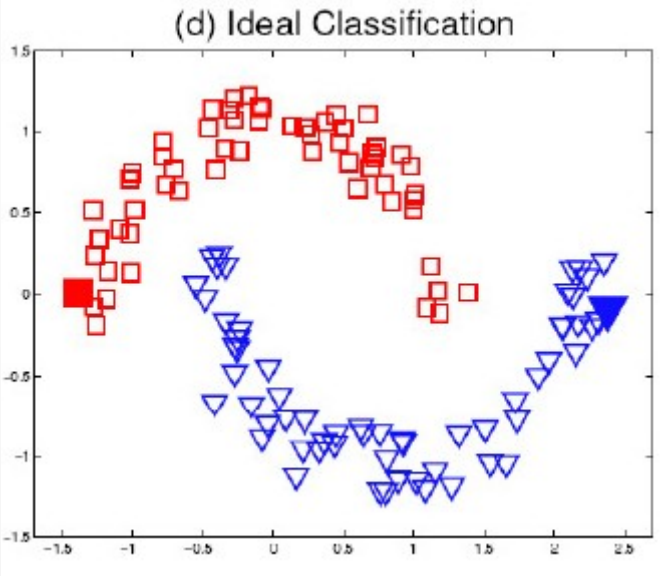
Два варианта постановки задачи:

- **Частичное обучение** (semi-supervised learning):
  построить алгоритм классификации a : X → Y .
- **Трансдуктивное обучение** (transductive learning):
  доразметить данные,
  т. е. зная все { xℓ+1, . . . , xℓ+k}, получить метки { yℓ+1, . . . , yℓ+k }

Задача частичного обучения

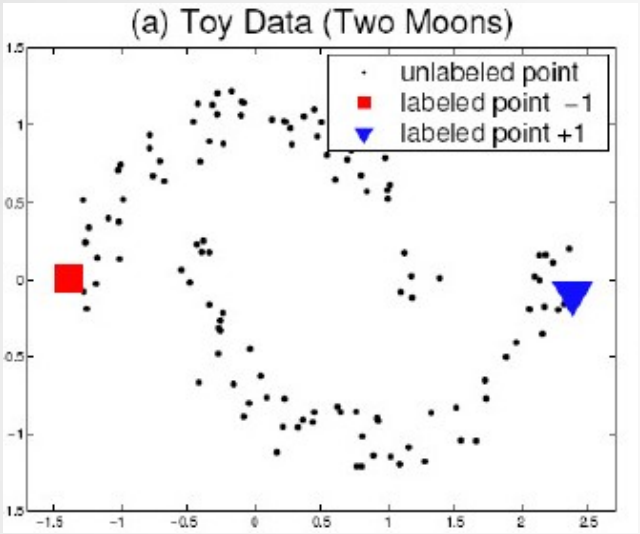
Рассмотрим частично размеченный датасет

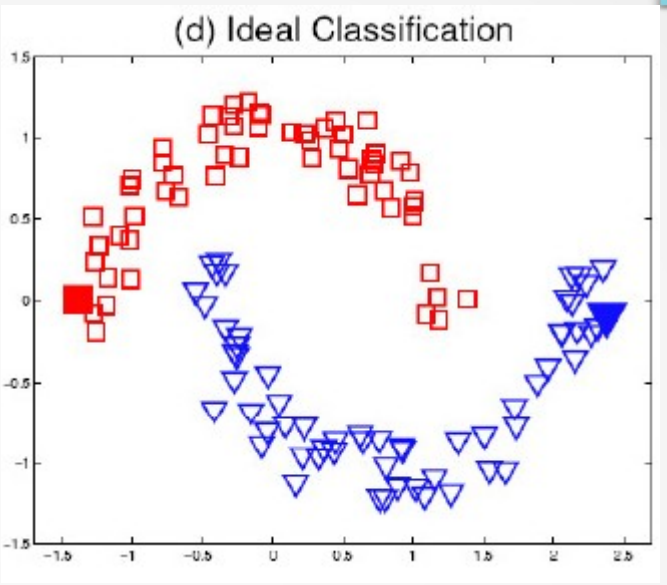


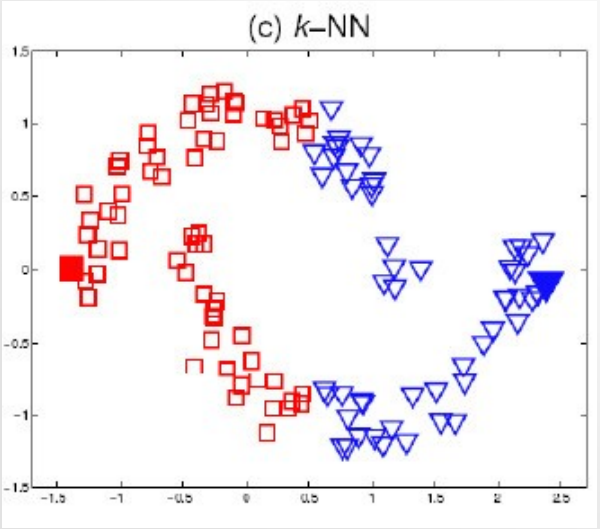


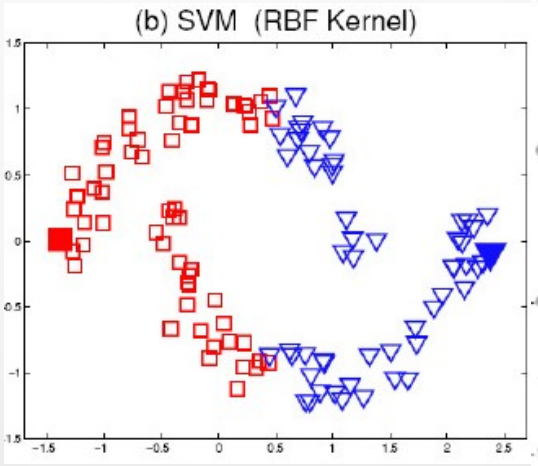
Задача частичного обучения

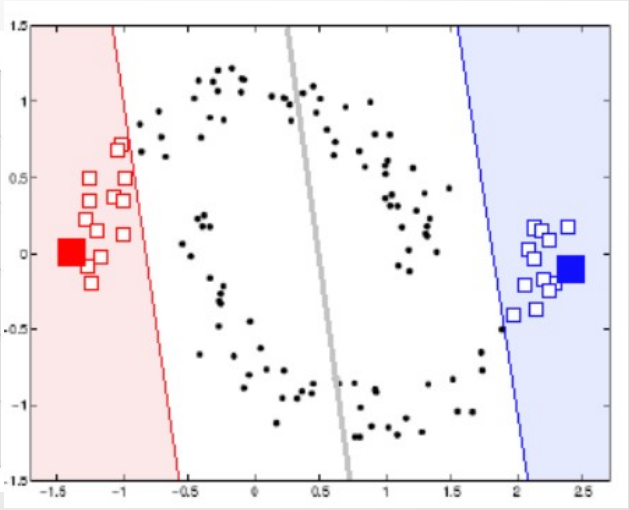
SSL не сводится к классификации





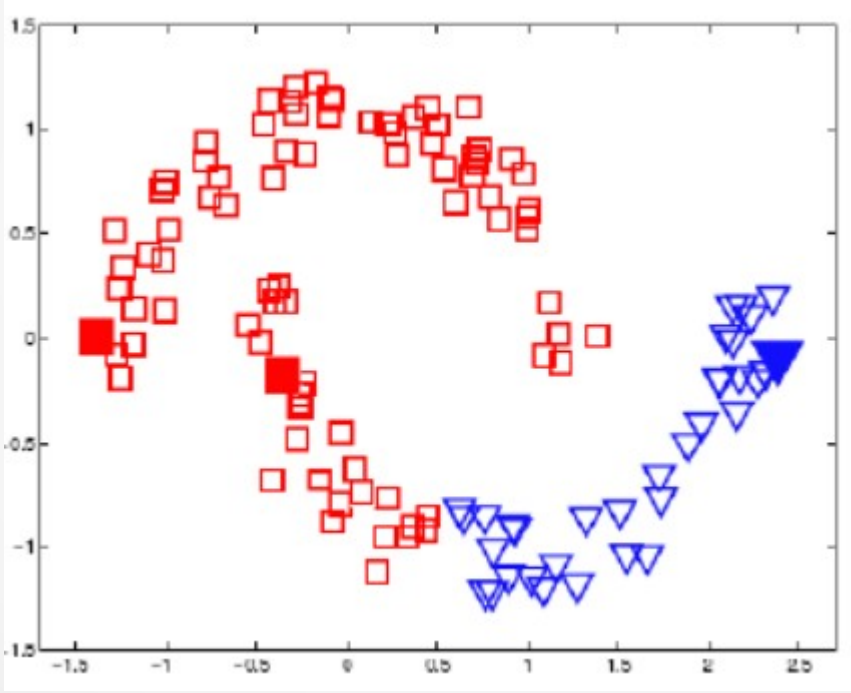






Задача частичного обучения

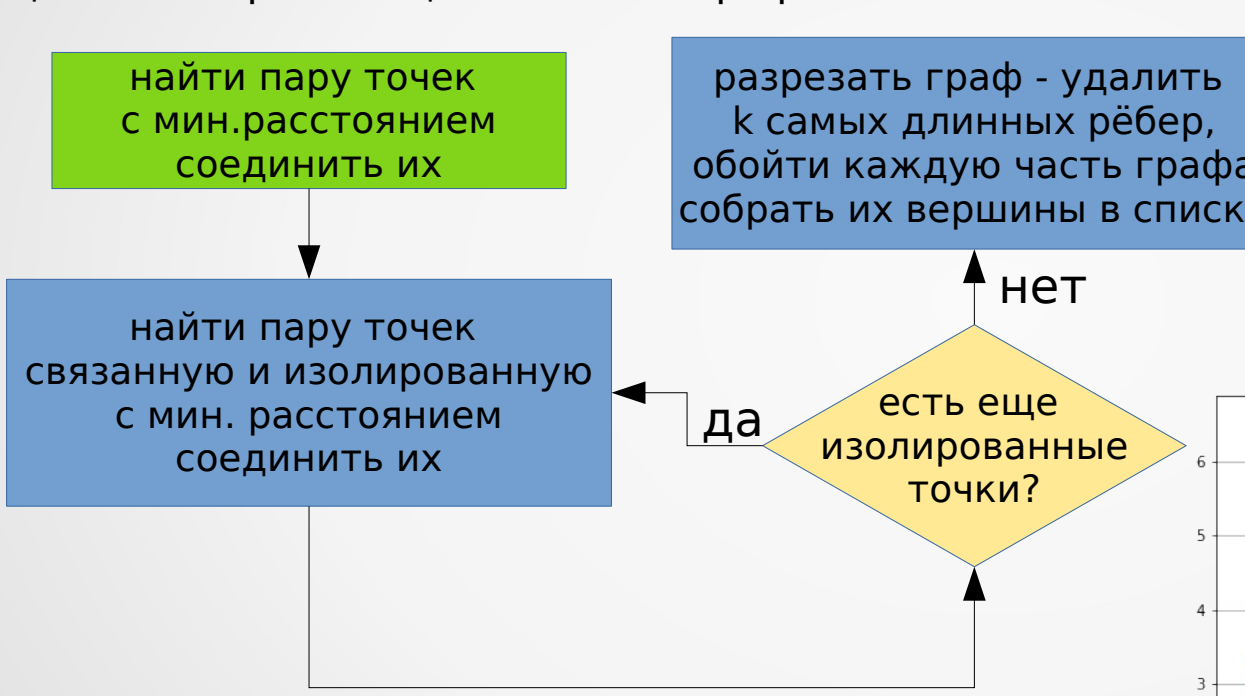
к кластеризации SSL тоже не сводится

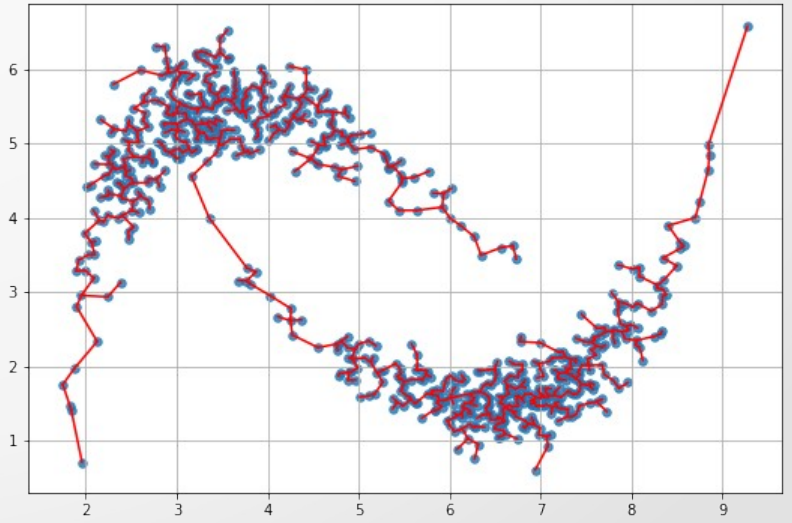


Задача частичного обучения

метод кластеризации КНП (Кратчайший Незамкнутый Путь)

параметр - количество кластеров k
цель - построить ациклический граф на точках

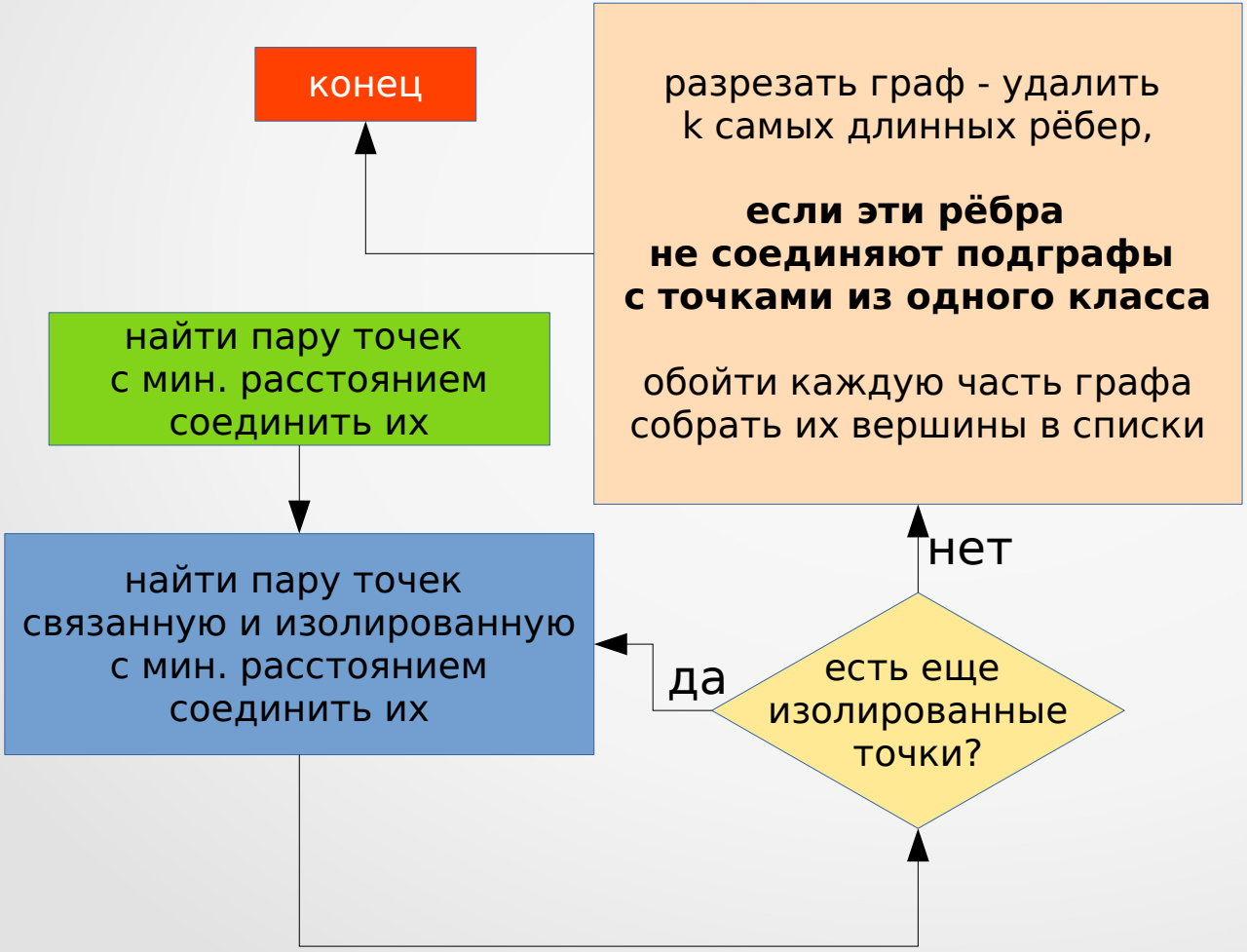


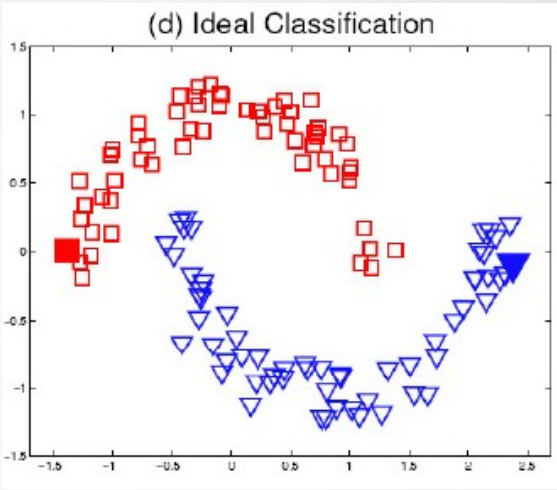


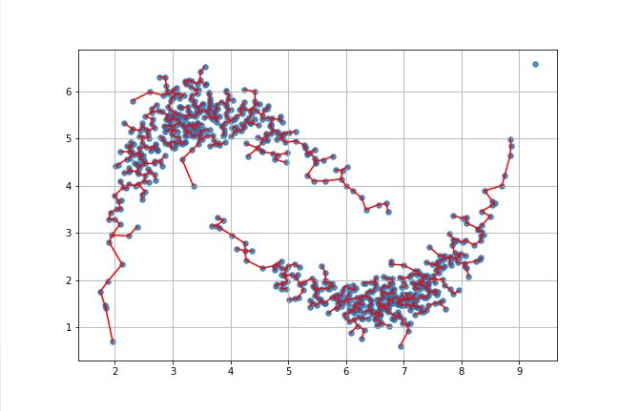
Задача частичного обучения

метод Т-КНП (Трансдуктивный Кратчайший Незамкнутый Путь)

параметр - количество кластеров k
цель - построить ациклический граф на точках



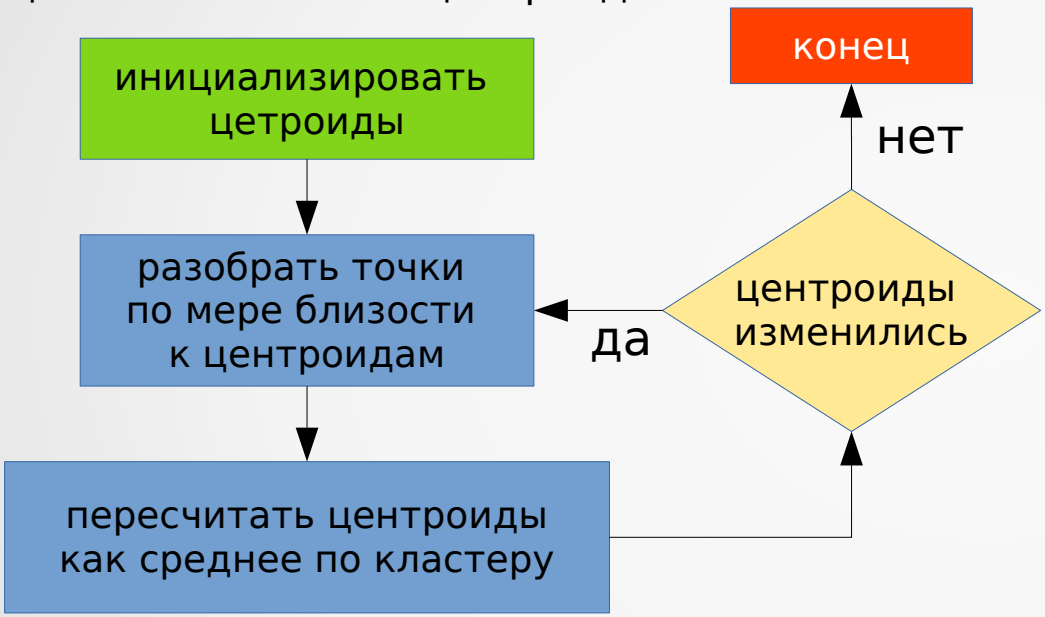


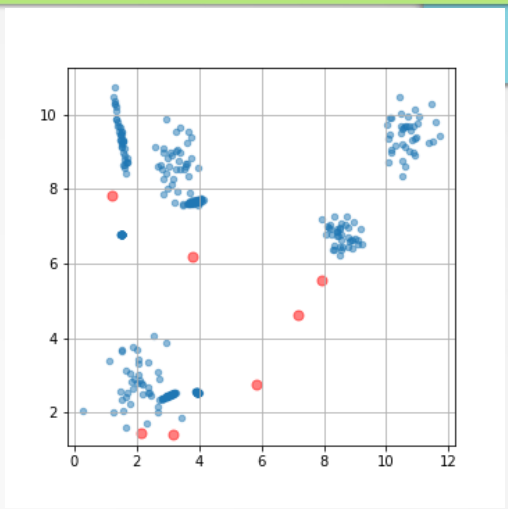


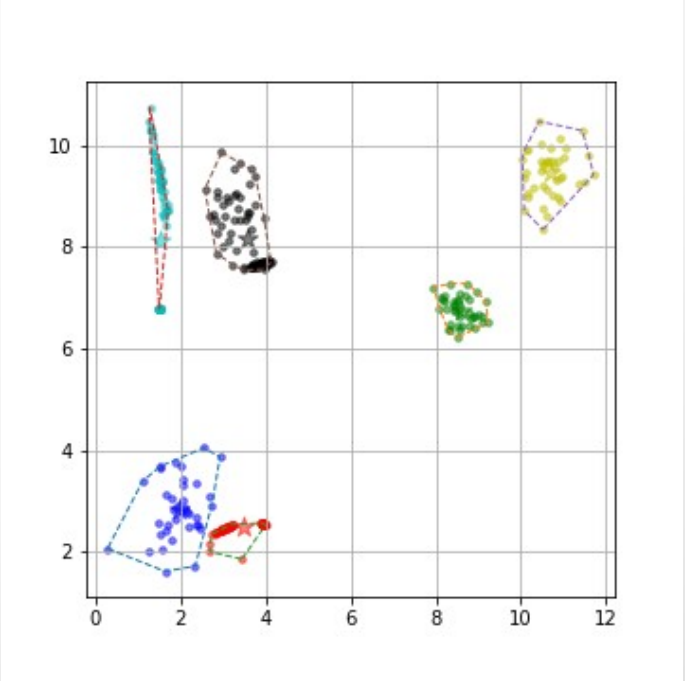
Задача частичного обучения

метод кластеризации к-средних (k-means)

параметр - количество кластеров
цель - найти точки-центроиды



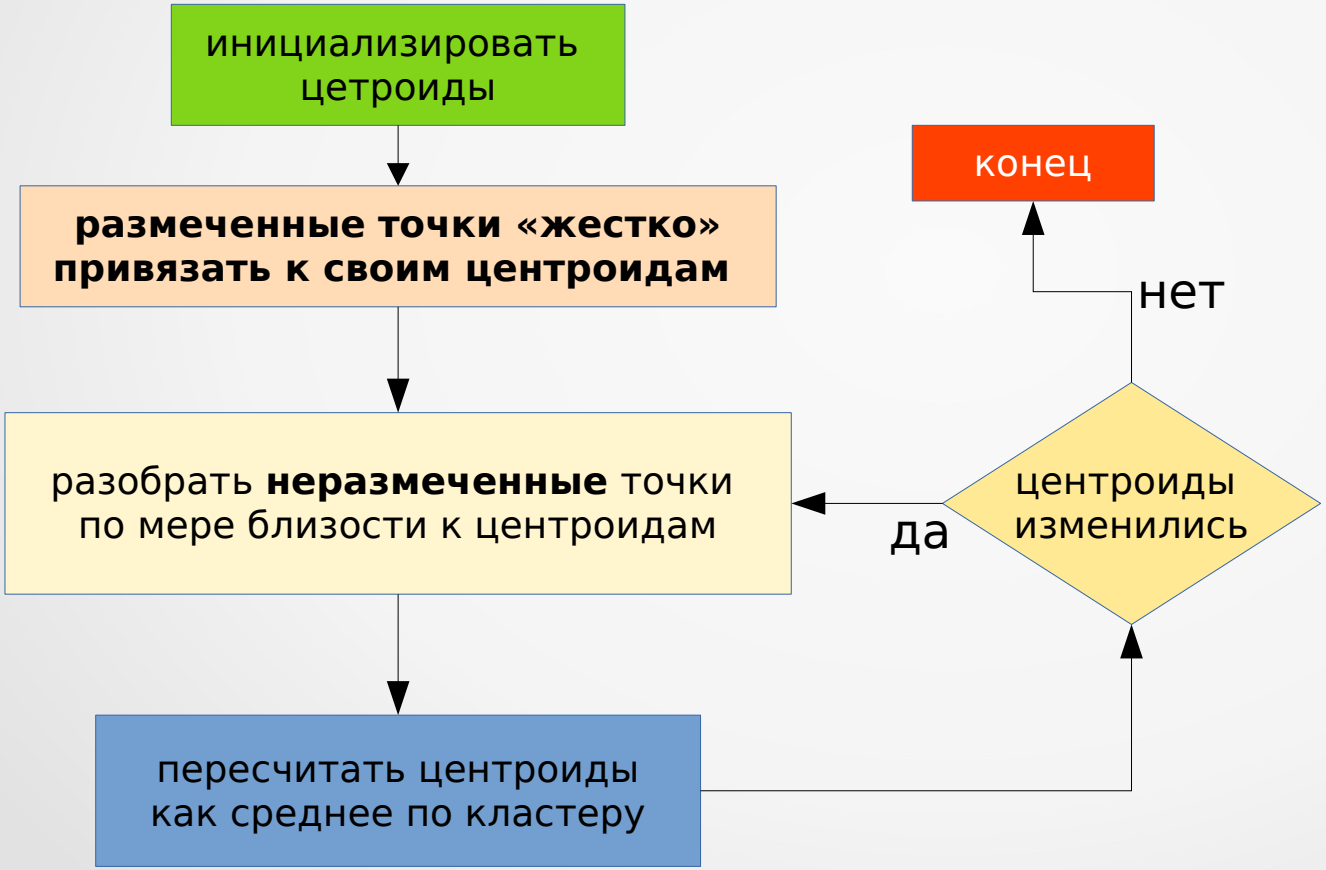




Задача частичного обучения

трансдуктивный метод k-means

параметр - количество кластеров
цель - найти точки-центроиды



Задача частичного обучения

недостаток трансдуктивных методов кластеризации

- это медленно, проблематично исполнять для больших датасетов

Задача частичного обучения

Классификация частично размеченного датасета

X, Y - учебный набор, частично размеченный;
$X^\ell = \{x_1, \dots, x_\ell\}, \{y_1, \dots, y_\ell\}$ - размеченная часть выборки (labeled data);
$X^k = \{x_{\ell+1}, \dots, x_{\ell+k}\}$ - неразмеченная часть выборки (unlabeled data);

$b_y(x)$ - оценка принадлежности объекта x к классу y

классификатор - выбираем для объекта x класс y с наилучшей оценкой b ;

$$a(x) = \arg\max_{y \in Y} b_y(x)$$

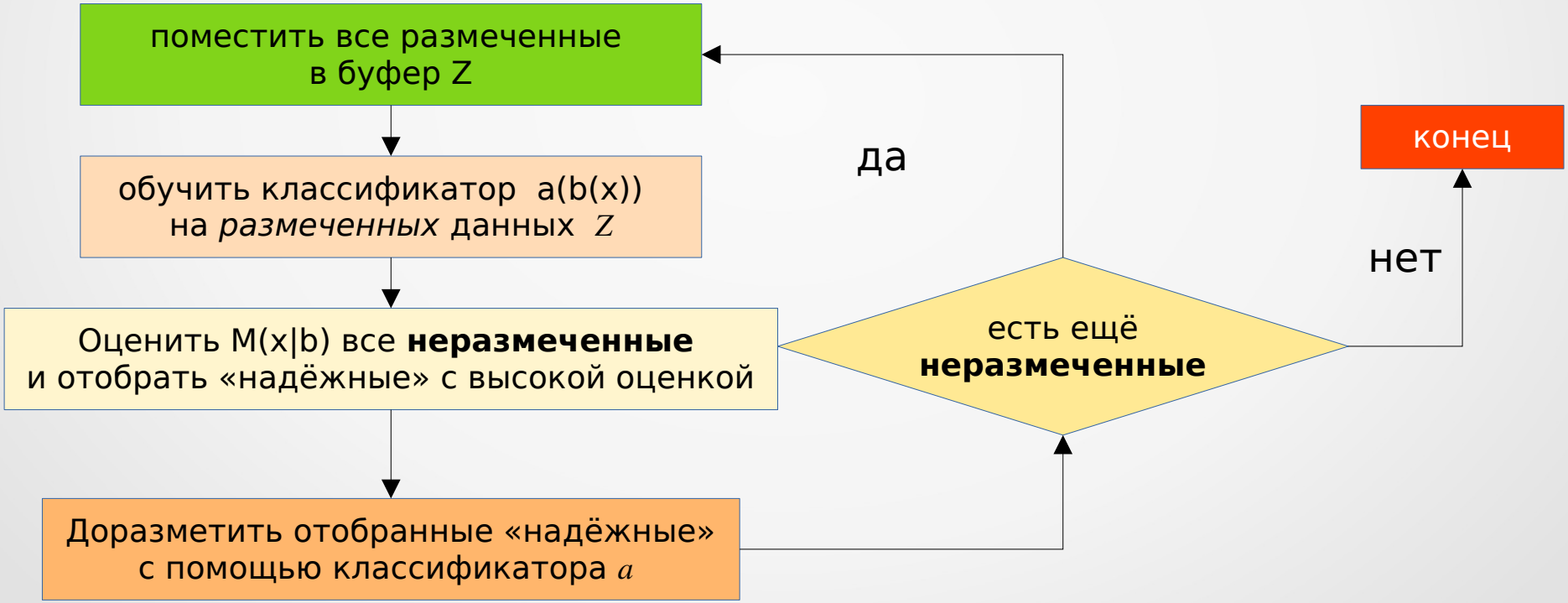
Как обучить классификатор на частично размеченном датасете?

Задача частичного обучения

Алгоритм Self-Training — обёртка (wrapper) над произвольным методом обучения классификатора

оценка "степень доверия" классификации,
насколько оценка класса-победителя лучше оценок остальных классов

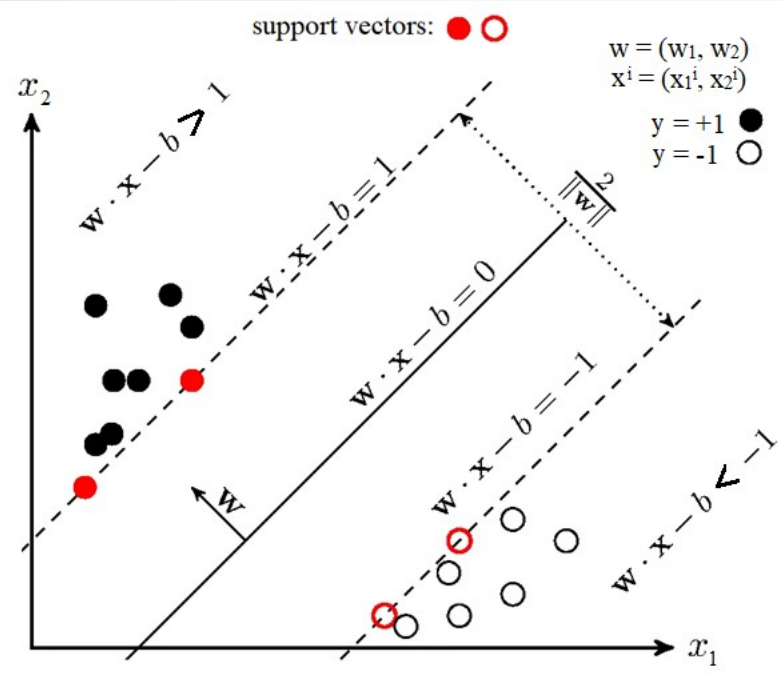
$$y = a(b(x))$$
$$M(x|b) = \max_{y \in Y} b_y(x) - \max_{q \in Y \setminus y} b_q(x)$$



Задача частичного обучения

Классификатор SVM

цель - разделительная полоса максимальной ширины



ширина полосы

$$\left\langle x_s^+ - x_s^- , \frac{w}{\|w\|} \right\rangle = \frac{2}{\|w\|}$$

метки классов

$$y \in \{-1, +1\}$$

модель классификатора

$$a(x) = sign(x \cdot w - b)$$

отступ (margine)

$$M = y \cdot (x \cdot w - b)$$

разделяющая гиперплоскость

$$\langle x, w \rangle - b = 0$$

$\langle x^+, w \rangle - b > 0$ позитивные

$\langle x^-, w \rangle - b < 0$ негативные

https://habr.com/ru/company/ods/blog/484148/

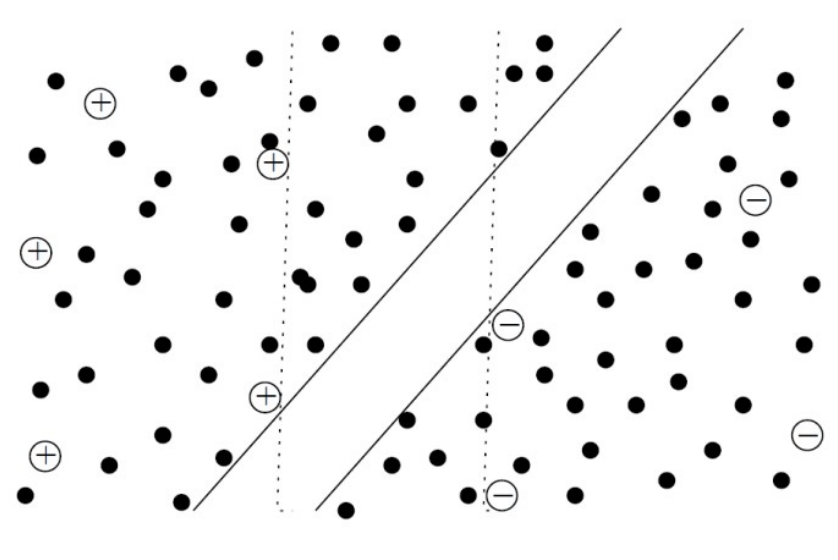
Задача частичного обучения

T-SVM (Трансдуктивный SVM)

цель - разделительная полоса максимальной ширины

строим разделяющую полосу с двумя условиями
1. полоса максимальной ширины ограничивается размеченными объектами
2. минимизировать количество (неразмеченных) объектов попадающих на полосу

Функция потерь L(M) = (1−|M|)+
штрафует за попадание неразмеченных объектов внутрь разделяющей полосы.
где |M| - абсолютное значение отступа (модуль)



Обучение весов w, w_0 по частично размеченной выборке:

$$Q(w, w_0) = \sum_{i=1}^{\ell} \big(1 - M_i(w, w_0)\big)_+ + \frac{1}{2C}\|w\|^2 + \\ + \gamma \sum_{i=\ell+1}^{\ell+k} \big(1 - |M_i(w, w_0)|\big)_+ \to \min_{w, w_0}.$$

Пунктирная линия — решение обычного SVM

Сплошная линия — решение T-SVM

https://habr.com/ru/company/ods/blog/484148/

Задача частичного обучения

Борисов Е.С. Методы машинного обучения. 2024
https://github.com/mechanoid5/ml_lectorium_2024_I

К.В. Воронцов Методы частичного обучения.
https://www.youtube.com/watch?v=DwA91VHydXU
http://www.machinelearning.ru/wiki/images/archive/9/9f/20151223000522%21Voron-ML-SSL.pdf

К.В.Воронцов Машинное обучение. курс лекций.
http://www.machinelearning.ru/wiki/index.php?title=Машинное_обучение_%28курс_лекций%2C_К.В.Воронцов%29