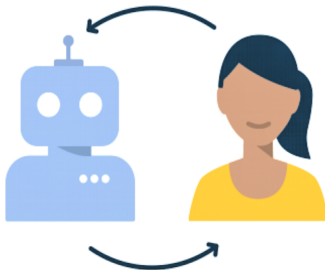


# Активное обучение

Виктор Китов

[v.v.kitov@yandex.ru](mailto:v.v.kitov@yandex.ru)

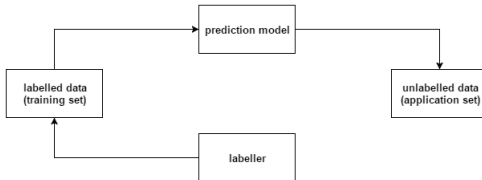


# Содержание

- 1 Постановка задачи
- 2 Кластерный подход к активному обучению
- 3 Активное обучение, основанное на модели

# Типовая последовательность действий

- Последовательность действий в обучении с учителем:



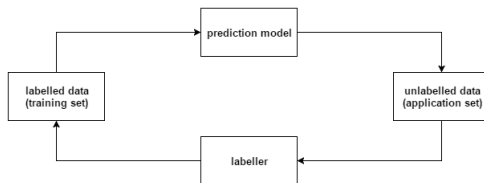
- Проблемы с точностью могут быть не из-за модели, а из-за обучающей выборки.
  - часто мы можем управлять созданием этой выборки.

# Мотивация активного обучения

- Обычно доступно много неразмеченных данных:
  - документы, изображения, видео, речь в интернете
- Получить объекты легко, но дорого размечать.
  - хотим небольшую но высокоинформативную выборку
- Активное обучение (active learning, AL) - процесс оптимального построения обучающей выборки
  - за минимальное число разметок лучше всего настроить модель.

## Схема активного обучения

- В активном обучении возвращаемся к расширению обучающей выборки много раз.



## Мотивационный пример

- $x \in [0, 1]$ ,  $y = \text{sign}(x - \theta)$
- Хотим оценить  $\theta$ , используя  $N$  оценок  $y(x)$ .

---

<sup>1</sup>Какая будет точность, если объекты выбирать случайно из равномерного распределения?

## Мотивационный пример

- $x \in [0, 1]$ ,  $y = \text{sign}(x - \theta)$
- Хотим оценить  $\theta$ , используя  $N$  оценок  $y(x)$ .
- Объекты выбираем равномерно по сетке:<sup>1</sup>
  - точность  $O\left(\frac{1}{N}\right)$ .

---

<sup>1</sup>Какая будет точность, если объекты выбирать случайно из равномерного распределения?

## Мотивационный пример

- $x \in [0, 1]$ ,  $y = \text{sign}(x - \theta)$
- Хотим оценить  $\theta$ , используя  $N$  оценок  $y(x)$ .
- Объекты выбираем равномерно по сетке:<sup>1</sup>
  - точность  $O\left(\frac{1}{N}\right)$ .
- Выбираем каждый раз  $x$  в середине интервала  $[x_i, x_{i+1}]$ , где

$$i = \arg \max_k y(x_k) = -1$$

$$i + 1 = \arg \min_k y(x_k) = +1$$

- точность  $O\left(\frac{1}{2^N}\right)$

---

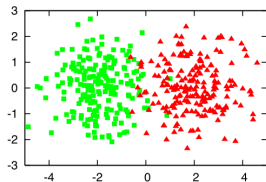
<sup>1</sup>Какая будет точность, если объекты выбирать случайно из равномерного распределения?



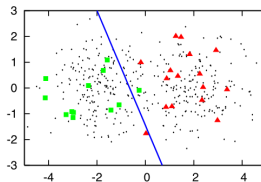
# Мотивационный пример

$$x|y = -1 \sim \mathcal{N}(\mu_1, \Sigma_1), \quad x|y = +1 \sim \mathcal{N}(\mu_2, \Sigma_2)$$

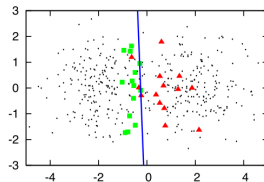
Точность модели значительно повышается, если выбирать объекты вблизи разделяющей гиперплоскости.



реальные классы

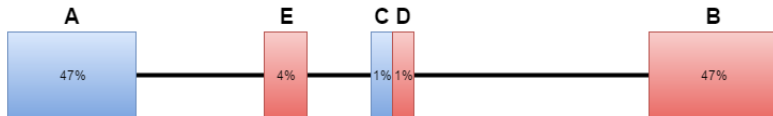


30% случайных объектов



30% выбранных объектов

## Когда активное обучение хуже



- Предположим, объекты исходной выборки оказались из блоков A,C,D,B.
- Активное обучение будет уточнять границу между C,D
  - точность 4%
- Но лучшее решение - граница между A,E
  - точность 1%

## Решение проблемы

- Чтобы такого не возникало,  
нужна представительная начальная обучающая выборка.
- Также можно смешивать обычное и активное обучение  
( $\epsilon$ -active стратегия):

$$(x, y) \text{ выбирается } \begin{cases} \text{случайно} & \text{с вероятностью } \epsilon \\ \text{направленно} & \text{с вероятностью } 1 - \epsilon \end{cases}$$

- $\epsilon \in [0, 1]$  контролирует exploration-exploitation tradeoff.
  - можно динамически менять  $\epsilon$  (exponential gradient):
- Валидационная выборка AL - нерепрезентативна  
(сэмплируются более сложные объекты)
  - оценивать качество нужно на  
случайной валидационной выборке.

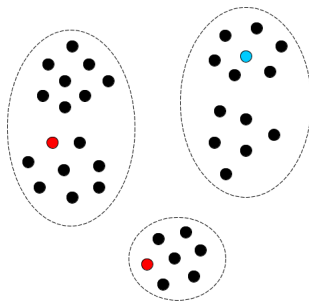
# Категоризация постановок задач и методов

- Объекты для разметки могут:
  - выбираться из готовой коллекции (в лекции)
  - выбираться из динамического потока
    - экспертная интерпретация изменения цены на отдельные акции
  - генерироваться вручную
    - в робототехнике можем посылать любое управляющее воздействие и смотреть результат.
- Основные подходы активного обучения
  - основанные на кластеризации
  - основанные на предиктивной модели

# Содержание

- 1 Постановка задачи
- 2 Кластерный подход к активному обучению
- 3 Активное обучение, основанное на модели

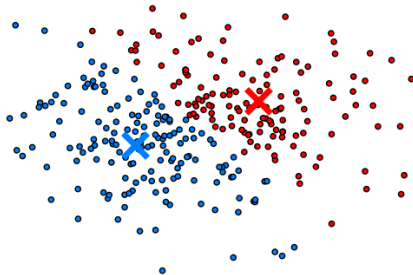
# Кластерный подход к активному обучению



- Идея: более репрезентативны и разнообразны объекты из разных кластеров.
  - выбираем объекты стратифицированно по кластерам
- Результирующая выборка не привязана к целевой модели
  - поэтому результат может быть хуже
  - зато - применимость к разным моделям

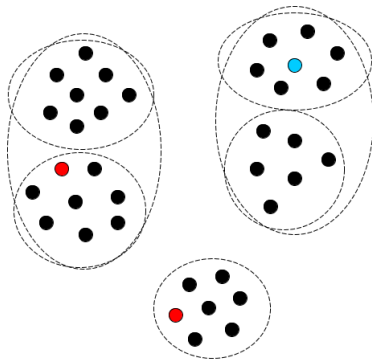
## Проблемы кластерного подхода

Кластерная структура может отсутствовать:



# Проблемы кластерного подхода

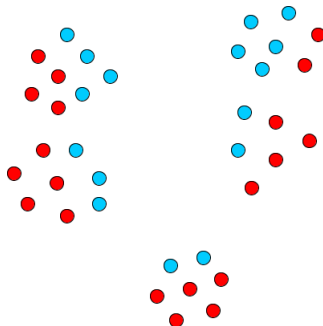
Кластеры могут быть разного уровня детализации:





## Проблемы кластерного подхода

Отклики могут быть различные внутри кластеров:



# Содержание

- 1 Постановка задачи
- 2 Кластерный подход к активному обучению
- 3 Активное обучение, основанное на модели

## Выбор по степени неуверенности: 2 класса

- **Сэмплирование по минимальной уверенности (least confident sampling):**
  - выбираем  $x$ , в прогнозе которого модель более всего не уверена
- В случае бинарной классификации:

$$x^* = \arg \min_x |P(y = +1|x) - 0.5|$$

- Если классификатор не умеет считать вероятности:

## Выбор по степени неуверенности: 2 класса

- **Сэмплирование по минимальной уверенности (least confident sampling):**
  - выбираем  $x$ , в прогнозе которого модель более всего не уверена
- В случае бинарной классификации:

$$x^* = \arg \min_x |P(y = +1|x) - 0.5|$$

- Если классификатор не умеет считать вероятности:
  - выбираем объекты, максимально близкие к разделяющей поверхности

для линейного классификатора:

$$x^* = \arg \min_x \frac{|w_0 + w^T x|}{\|w\|} = \arg \min_x |w_0 + w^T x|$$

## Выбор по степени неуверенности: C классов

- Сэмплирование по минимальной уверенности (least confident sampling)<sup>2</sup>
  - объект с максимальной вероятностью ошибки:

$$x^* = \arg \max_x \left| 1 - \max_y p(y|x) \right|$$

---

<sup>2</sup>Предложите модификацию этого метода, работающую для регрессии.

## Выбор по степени неуверенности: C классов

- **Сэмплирование по минимальной уверенности (least confident sampling)**<sup>2</sup>

- объект с максимальной вероятностью ошибки:

$$x^* = \arg \max_x \left| 1 - \max_y p(y|x) \right|$$

- **Сэмплирование по отсупу (margin sampling)**

- объект с минимальным зазором в лидирующем классе:

$$x^* = \arg \min_x p(\hat{y}_1|x) - p(\hat{y}_2|x)$$

$$\hat{y}_1 = \arg \max_y p(y|x), \quad \hat{y}_2 = \arg \max_{y \neq \hat{y}_1} p(y|x)$$

---

<sup>2</sup>Предложите модификацию этого метода, работающую для регрессии.

## Выбор по степени неуверенности: C классов

- **Сэмплирование по минимальной уверенности (least confident sampling)<sup>2</sup>**

- объект с максимальной вероятностью ошибки:

$$x^* = \arg \max_x \left| 1 - \max_y p(y|x) \right|$$

- **Сэмплирование по отсупу (margin sampling)**

- объект с минимальным зазором в лидирующем классе:

$$x^* = \arg \min_x p(\hat{y}_1|x) - p(\hat{y}_2|x)$$

$$\hat{y}_1 = \arg \max_y p(y|x), \quad \hat{y}_2 = \arg \max_{y \neq \hat{y}_1} p(y|x)$$

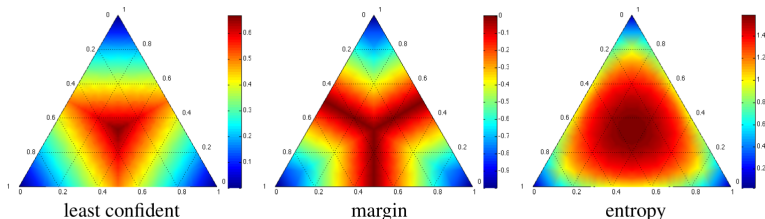
- **Сэмплирование по энтропии (entropy sampling)**

- объект с максимальной энтропией класса:

$$x^* = \arg \max_x \{ Entropy(y|x) \} \quad Entropy(y|x) = - \sum_y p(y|x) \ln p(y|x)$$

<sup>2</sup>Предложите модификацию этого метода, работающую для регрессии.

Приоритетность от  $p(y = 1|x)$ ,  $p(y = 2|x)$ ,  $p(y = 3|x)$





## Отбор по несогласию в ансамбле моделей

- Классификация ансамблем моделей  $f_1, \dots, f_M$ ; выбираем  $x$ , для которого  $f_1(x), \dots, f_M(x)$  сильнее всего рассогласованы.
- **Энтропия голосования (vote entropy)<sup>3</sup>**

$$x^* = \arg \max_x \{ Entropy(y|x) \}$$

$$Entropy(y|x) = - \sum_y \frac{\#[f_i(x) = y]}{M} \ln \left( \frac{\#[f_i(x) = y]}{M} \right)$$

---

<sup>3</sup>Предложите модификацию этого метода, работающую для регрессии.

## Отбор по несогласию в ансамбле моделей

- Классификация ансамблем моделей  $f_1, \dots, f_M$ ; выбираем  $x$ , для которого  $f_1(x), \dots, f_M(x)$  сильнее всего рассогласованы.
- **Энтропия голосования (vote entropy)<sup>3</sup>**

$$x^* = \arg \max_x \{ Entropy(y|x) \}$$

$$Entropy(y|x) = - \sum_y \frac{\#[f_i(x) = y]}{M} \ln \left( \frac{\#[f_i(x) = y]}{M} \right)$$

- **Рассогласованность распределений (spread in probabilities)**

$$x^* = \arg \max_x \frac{1}{M} \sum_{m=1}^M \rho(P_m(x), P(x)) \quad P(x) = \frac{1}{M} \sum_{m=1}^M P_m(x)$$

$P_m(x)$  — распределение классов согласно  $f_m$ ,  $\rho$ -напр.  $KL$

<sup>3</sup>Предложите модификацию этого метода, работающую для регрессии.

## Максимизация ожидаемого влияния на модель

- **Максимизация ожидаемого влияния на модель (expected model change)**
  - выберем  $x$  сильнее всего изменяющий модель.

---

<sup>4</sup>Важна нормализация признаков, т.к. их масштаб влияет на  $\nabla_{\theta} \mathcal{L}(f_{\theta}(x), y)$ .

## Максимизация ожидаемого влияния на модель

- **Максимизация ожидаемого влияния на модель (expected model change)**
  - выберем  $x$  сильнее всего изменяющий модель.
- После добавления  $(x, y)$ :

$$\nabla L(\theta) = \underbrace{\sum_{n=1}^N \nabla \mathcal{L}(f_{\theta}(x_n), y_n)}_{\approx 0 \text{ т.к. оптимизация сошлась}} + \nabla \mathcal{L}(f_{\theta}(x), y) \approx \nabla \mathcal{L}(f_{\theta}(x), y)$$

$\|\nabla_{\theta} \mathcal{L}(f_{\theta}(x), y)\| \approx$  влияние  $(x, y)$  на модель

---

<sup>4</sup>Важна нормализация признаков, т.к. их масштаб влияет на  $\nabla_{\theta} \mathcal{L}(f_{\theta}(x), y)$ .

# Максимизация ожидаемого влияния на модель

- Максимизация ожидаемого влияния на модель (expected model change)
  - выберем  $x$  сильнее всего изменяющий модель.
- После добавления  $(x, y)$ :

$$\nabla L(\theta) = \underbrace{\sum_{n=1}^N \nabla \mathcal{L}(f_{\theta}(x_n), y_n)}_{\approx 0 \text{ т.к. оптимизация сошлась}} + \nabla \mathcal{L}(f_{\theta}(x), y) \approx \nabla \mathcal{L}(f_{\theta}(x), y)$$

$\|\nabla_{\theta} \mathcal{L}(f_{\theta}(x), y)\| \approx$  влияние  $(x, y)$  на модель

- Выберем  $x$  максимизирующий ожидаемое изменение<sup>4</sup>:

$$x^* = \arg \max_x \sum_y p(y|x) \|\nabla_{\theta} \mathcal{L}(f_{\theta}(x), y)\|$$

<sup>4</sup>Важна нормализация признаков, т.к. их масштаб влияет на  $\nabla_{\theta} \mathcal{L}(f_{\theta}(x), y)$ .

## Ожидаемое сокращение ошибки

- Предыдущие методы основывались на неопределенности отдельного объекта.
- **Ожидаемое сокращение ошибки (expected error reduction)**
  - выберем  $x$ , максимально снижающий неопределенность классов в неразмеченной выборке
  - наилучший эффект, но самый медленный метод
- Обозначим  $U$  - множество неразмеченных объектов.
- $f^{+(x,y)}$ : модель  $f$ , дообученная на  $(x, y)$

## Ожидаемое сокращение ошибки

- Снижение частоты ошибок (certainty improvement):

$$x^* = \arg \max_{x \in U} \sum_y p_f(y|x) \left( \sum_{u \in U} p_{f+(x,y)}(\hat{y}_u | x_u) \right),$$

$$\hat{y}_u = \arg \max_y p_{f+(x,y)}(y | x_u)$$

- Снижение энтропии классов (entropy minimization):

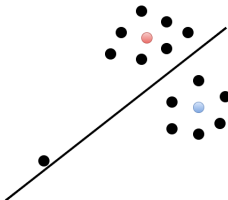
- эквивалентно максимизации логарифма правдоподобия

$$x^* = \arg \min_{x \in U} \sum_y p_f(y|x) \sum_{u \in U} \underbrace{- \sum_{\tilde{y}} p_{f+(x,y)}(\tilde{y} | x_u) \ln p_{f+(x,y)}(\tilde{y} | x_u)}_{\text{Entropy}(\tilde{y} | x_u, f^{+(x,y)})}$$

## Обработка выбросов

Выбросы по умолчанию предпочитаются методами AL:

- лежат в отдельных "кластерах"
- модели ансамбля экстраполируют зависимость по-разному
- сильно влияют на модель и ее прогнозы





## Учет выбросов

- Решение - отфильтровать выбросы либо занижать их влияние:

$$x^* = \arg \max_x \left\{ \text{score}(x) \times \text{typicalness}(x)^\beta \right\}$$

$\text{typicalness}(x)^\beta \in [0, 1]$  - типичность объекта

$\beta \geq 0$  - гиперпараметр (сила фильтрации выбросов)

- Примеры:

$$\text{typicalness}(x) = p(x)$$

$$\text{typicalness}(x) = \frac{1}{\rho(x, x_{i(K)})}$$

$x_{i(K)}$  -  $K$ -й ближайший сосед

## Заключение

- Задача активного обучения:  $\downarrow$ размер и  $\uparrow$ информативность обучающей выборки.
  - обучающая выборка: баланс между случайной и направленной генерацией
  - валидационная выборка: только случайная.
- Подходы активного обучения:
  - основанные на кластеризации
  - основанные на модели:
    - по степени неуверенности прогнозов
    - по степени рассогласования между базовыми моделями ансамбля
    - по влиянию на модель
    - по влиянию на однозначность разметки неразмеченной выборки
- Выбросы должны отбрасываться или учитываться с меньшим весом.