

# Активное обучение (Active Learning)

К. В. Воронцов  
vokov@forecsys.ru

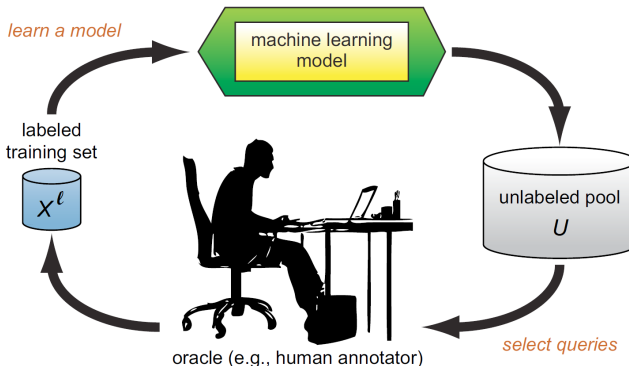
Этот курс доступен на странице вики-ресурса  
<http://www.MachineLearning.ru/wiki>  
«Машинное обучение (курс лекций, К.В.Воронцов)»

ШАД Яндекс • 24 ноября 2020

- 1 Стратегии активного обучения**
  - Постановка задачи активного обучения
  - Отбор объектов из выборки
  - Синтез объектов (планирование экспериментов)
- 2 Активное обучение с изучающими действиями**
  - Компромисс «изучение–применение»
  - Алгоритм  $\varepsilon$ -active
  - Экспоненциальный градиент
- 3 Активное обучение в краудсорсинге**
  - Задача краудсорсинга
  - Согласование оценок аннотаторов
  - Активное распределение заданий по аннотаторам

## Постановка задачи активного обучения

Задача: обучение модели  $a: X \rightarrow Y$  по выборке  $(x_i, y_i)$ , когда получение ответов  $y_i = y(x_i)$  стоит дорого.



Burr Settles. Active Learning Literature Survey. 2010.

## Постановка задачи активного обучения

**Задача:** обучение модели  $a: X \rightarrow Y$  по выборке  $(x_i, y_i)$ , когда получение ответов  $y_i = y(x_i)$  стоит дорого.

**Вход:**  $X^\ell = (x_i, y_i)_{i=1}^\ell$  — выборка размеченных объектов;

$U = (u_i)_{i=1}^K$  — выборка (пул) неразмеченных объектов;

**Выход:** модель  $a$  и размеченная выборка  $(u_i, y_i^*)_{i=1}^k$ ,  $k \leq K$ ;

обучить модель  $a$  по начальной выборке  $(x_i, y_i)_{i=1}^\ell$ ;

**пока** есть неразмеченные объекты и модель не обучилась

$u_i = \arg \max_{u \in U} \phi(u)$  — выбрать неразмеченный объект;

узнать для него  $y_i^* = y(u_i)$ ;

дообучить модель  $a(x)$  ещё на одном примере  $(u_i, y_i^*)$ ;

**Цель:** достичь как можно лучшего качества модели  $a$ , использовав как можно меньше дополнительных примеров  $k$ .

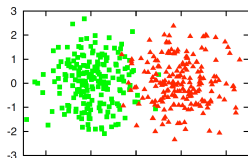
## Почему активное обучение быстрее пассивного

**Пример 1.** Синтетические данные:  $\ell = 30$ ,  $\ell + k = 400$ ;

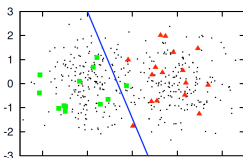
(а) два гауссовских класса;

(б) логистическая регрессия по 30 случайным объектам;

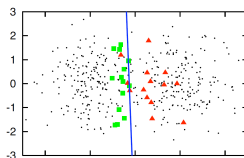
(с) логистическая регрессия по 30 объектам, отобранным с помощью активного обучения.



(a)



(b)



(c)

Обучение по смещённой неслучайной выборке требует меньше данных для построения алгоритма сопоставимого качества.

*Burr Settles.* Active Learning Literature Survey. 2010.

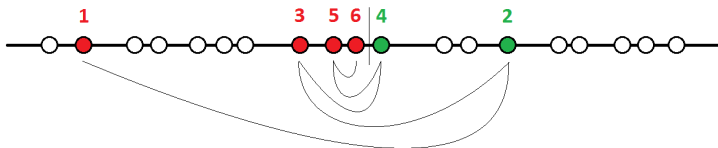
## Почему активное обучение быстрее пассивного

**Пример 2.** Одномерная задача с пороговым классификатором:

$$x_i \sim \text{uniform}[-1, +1], \quad y_i = [x_i > 0], \quad a(x, \theta) = [x > \theta].$$

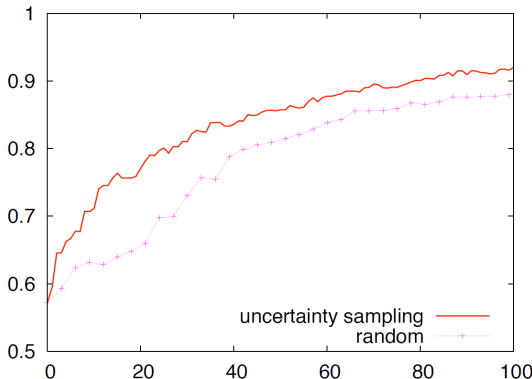
Оценим число шагов для определения  $\theta$  с точностью  $\frac{1}{k}$ .

- Наивная стратегия: выбрать  $u_i \sim \text{uniform}(U)$ ;  
— число шагов  $O(k)$ .
- Бинарный поиск: выбрать  $u_i$ , ближайший к середине зазора между классами  $\frac{1}{2} \left( \max_{y_j=0} (x_j) + \min_{y_j=1} (x_j) \right)$ ;  
— число шагов  $O(\log k)$ .



## Оценивание качества активного обучения

Кривая обучения (learning curve) — зависимость точности классификации на тесте от числа размеченных объектов  $k$ .



*Burr Settles*. Active Learning Literature Survey. 2010.

## Стратегии активного обучения

- **Отбор объектов из выборки (pool-based sampling):**  
какой следующий  $u_i$  выбрать из множества  $U = \{u_i\}_{i=1}^K$
- **Синтез объектов (query synthesis):**  
на каждом шаге синтезировать оптимальный объект  $u_i$
- **Отбор объектов из потока (selective sampling):**  
для каждого приходящего  $u_i$  решать, стоит ли узнавать  $y_i^*$

Функционал качества модели  $a(x, \theta)$  с параметром  $\theta$ :

$$\sum_{i=1}^{\ell} \mathcal{L}(x_i, y_i; \theta) + \sum_{i=1}^k C_i \mathcal{L}(u_i, y_i^*; \theta) \rightarrow \min_{\theta},$$

где  $\mathcal{L}$  — функция потерь,  $C_i$  — стоимость информации  $y(u_i)$  для методов, чувствительных к стоимости (cost-sensitive)



## Примеры приложений активного обучения

- сбор ассессорских данных для информационного поиска, анализа текстов, сигналов, речи, изображений, видео
- в том числе на платформах краудсорсинга
- *планирование экспериментов* в естественных науках (пример — комбинаторная химия)
- оптимизация трудно вычисляемых функций (пример — поиск в пространстве гиперпараметров)

### Применения в бизнесе:

- управление ценами и ассортиментом в торговых сетях
- выбор товара для проведения маркетинговой акции
- проактивное взаимодействие с клиентами
- выборочный контроль качества
- выявление аномалий в данных, случаев мошенничества

## Сэмплирование по неопределённости (uncertainty sampling)

Идея: выбирать  $u_i$  с наибольшей неопределённостью  $a(u_i)$ .

Задача многоклассовой классификации:

$$a(u) = \arg \max_{y \in Y} P(y|u)$$

$p_m(u)$ ,  $m=1 \dots |Y|$  — ранжированные по убыванию  $P(y|u)$ ,  $y \in Y$ .

- Принцип *наименьшей достоверности* (least confidence):

$$u_i = \arg \min_{u \in U} p_1(u)$$

- Принцип *наименьшей разности отступов* (margin sampling):

$$u_i = \arg \min_{u \in U} (p_1(u) - p_2(u))$$

- Принцип *максимума энтропии* (maximum entropy):

$$u_i = \arg \min_{x \in U} \sum_m p_m(u) \ln p_m(u)$$

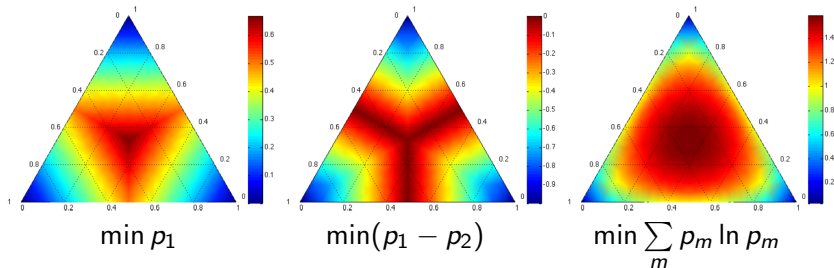
## Сэмплирование по не уверенности (uncertainty sampling)

В случае двух классов эти три принципа эквивалентны.

В случае многих классов появляются различия.

**Пример.** Три класса,  $p_1 + p_2 + p_3 = 1$ .

Показаны линии уровни трёх критериев выбора объекта:



*Burr Settles.* Active Learning Literature Survey. 2010.

## Сэмплирование по несогласию в комитете (query by committee)

**Идея:** выбирать  $u_i$  с наибольшей несогласованностью решений комитета моделей  $a_t(u_i) = \arg \max_{y \in Y} P_t(y|u_i)$ ,  $t = 1, \dots, T$ .

- Принцип *максимума энтропии*:  
выбираем  $u_i$ , на котором  $a_t(u_i)$  максимально различны:

$$u_i = \arg \min_{u \in U} \sum_{y \in Y} \hat{p}(y|u) \ln \hat{p}(y|u),$$

где  $\hat{p}(y|u) = \frac{1}{T} \sum_{t=1}^T [a_t(u) = y]$ .

- Принцип *максимума средней KL-дивергенции*:  
выбираем  $u_i$ , на котором  $P_t(y|u_i)$  максимально различны:

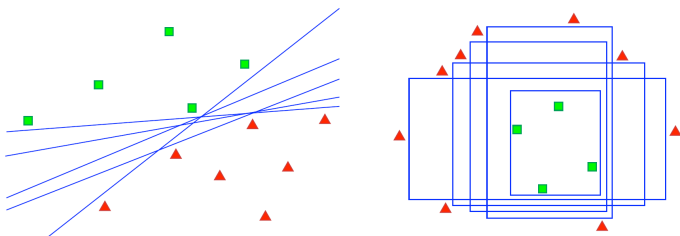
$$u_i = \arg \max_{u \in U} \sum_{t=1}^T \text{KL}(P_t(y|u) \parallel \bar{P}(y|u)),$$

где  $\bar{P}(y|u) = \frac{1}{T} \sum_{t=1}^T P_t(y|u)$  — консенсус комитета.

## Сокращение пространства решений (version space reduction)

**Идея:** выбирать  $u_i$ , максимально сужая множество решений.

**Пример.** Пространства допустимых решений для линейных и пороговых классификаторов (двумерный случай):



Бустинг и бэггинг находят конечные подмножества решений. Поэтому сэмплирование по несогласию в комитете — это аппроксимация принципа сокращения пространства решений.

## Ожидаемое изменение модели (expected model change)

**Идея:** выбрать  $u_i$ , который в методе стохастического градиента привёл бы к наибольшему изменению модели.

Параметрическая модель многоклассовой классификации:

$$a(u, \theta) = \arg \max_{y \in Y} P(y|u, \theta);$$

Для каждого  $u \in U$  и  $y \in Y$  оценим длину градиентного шага в пространстве параметров  $\theta$  при дообучении модели на  $(u, y)$ ; пусть  $\nabla_{\theta} \mathcal{L}(u, y; \theta)$  — вектор градиента функции потерь.

Принцип *максимума ожидаемой длины градиента*:

$$u_i = \arg \max_{u \in U} \sum_{y \in Y} P(y|u, \theta) \|\nabla_{\theta} \mathcal{L}(u, y; \theta)\|.$$

## Ожидаемое сокращение ошибки (expected error reduction)

**Идея:** выбирать  $u_i$ , который после дообучения даст наиболее уверенную классификацию неразмеченной выборки  $U \setminus u_i$ .

Для каждого  $u \in U$  и  $y \in Y$  обучим модель классификации, добавив к размеченной обучающей выборке  $X^\ell$  пример  $(u, y)$ :

$$a_{uy}(x) = \arg \max_{z \in Y} P_{uy}(z|x).$$

- Принцип *максимума уверенности на неразмеченных данных*:

$$u_i = \arg \max_{u \in U} \sum_{y \in Y} P(y|u) \sum_{u_j \in U \setminus u} P_{uy}(a_{uy}(u_j)|u_j).$$

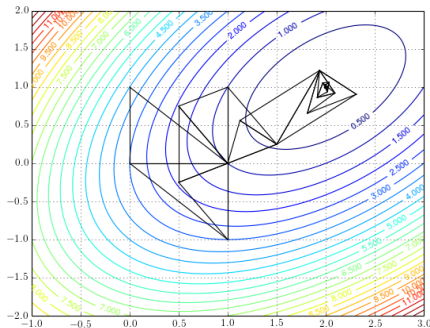
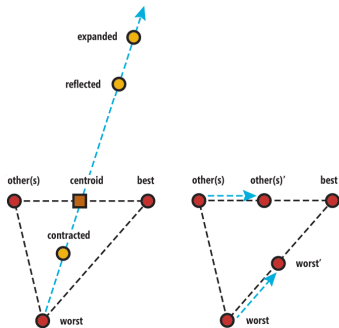
- Принцип *минимума энтропии неразмеченных данных*:

$$u_i = \arg \max_{u \in U} \sum_{y \in Y} P(y|u) \sum_{u_j \in U \setminus u} \sum_{z \in Y} P_{uy}(z|u_j) \log P_{uy}(z|u_j).$$

## Безградиентная оптимизация. Метод Нелдера–Мида

Идея: выбирать объекты  $u_i$  не из конечного пула, а из всего  $X$ , максимизируя  $\max_{u \in X} \phi(u)$  любым безградиентным методом.

Метод Нелдера–Мида: перемещение и деформирование симплекса из  $n + 1$  точек в пространстве  $X$  размерности  $n$



J.A.Nelder, R.Mead. A simplex method for function minimization. 1965.



# Метод Нелдера–Мида: «отражение–растяжение–сжатие»

## повторять

сортировка  $n + 1$  точек:  $\phi(x_w) < \phi(x_o) < \dots < \phi(x_b)$ ;

центроид  $x_c$ : по всем точкам кроме  $x_w$ ;

отражение:  $x_r := x_c + \alpha(x_c - x_w)$ ;

**если**  $\phi(x_b) < \phi(x_r)$  **то**

растяжение:  $x_{exp} := x_c + \gamma(x_r - x_c)$ ;

$x_w := (\phi(x_r) < \phi(x_{exp})) ? x_{exp} : x_r$ ;

**иначе если**  $\phi(x_o) < \phi(x_r) < \phi(x_b)$  **то**  $x_w := x_r$ ;

**иначе**

сжатие:  $x_{con} := x_c + \beta(x_w - x_c)$ ;

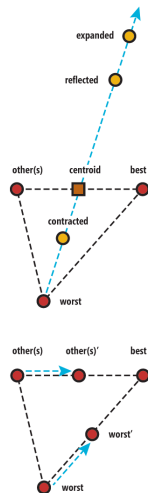
**если**  $\phi(x_w) < \phi(x_{con})$  **то**  $x_w := x_{con}$ ;

**иначе**

сжатие симплекса:  $x_j := x_b + \sigma(x_j - x_b)$ ;

**пока**  $\phi(x_w) \ll \phi(x_b)$ ;

Рекомендуемые параметры:  $\alpha = 1$ ,  $\beta = \frac{1}{2}$ ,  $\gamma = 2$ ,  $\sigma = \frac{1}{2}$



## Сокращение дисперсии (variance reduction)

**Идея:** выбирать  $u \in X$ , который даст наименьшую оценку дисперсии  $\sigma_a^2(u)$  после дообучения модели  $a(x, \theta)$ .

Задача регрессии, метод наименьших квадратов:

$$S^2(\theta) = \frac{1}{\ell} \sum_{i=1}^{\ell} (a(x_i, \theta) - y_i)^2 \rightarrow \min_{\theta}.$$

Из теории *оптимального планирования экспериментов* (OED, optimal experiment design):

$$u = \arg \min_{u \in U} \sigma_a^2(u), \quad \sigma_a^2(u) \approx S^2 \left( \frac{\partial a(u)}{\partial \theta} \right)^{\top} \left( \frac{\partial S^2}{\partial \theta^2} \right)^{-1} \left( \frac{\partial a(u)}{\partial \theta} \right).$$

В частности, для линейной регрессии

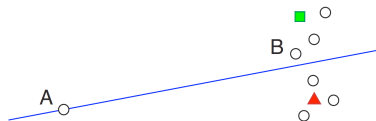
$$\sigma_a^2(u) \approx S^2 u^{\top} (F^{\top} F)^{-1} u,$$

где  $F$  — матрица объекты–признаки.

## Взвешивание по плотности (density-weighted methods)

**Идея:** понижать вес нерепрезентативных объектов.

**Пример.** Объект А более пограничный, но менее репрезентативный, чем В.



Любой критерий выбора объектов, имеющий вид

$$u = \arg \max_{u \in U} \phi(u),$$

может быть уточнён локальной оценкой плотности:

$$u = \arg \max_{u \in U} \phi(u) \left( \sum_{u' \in U} \text{sim}(u, u') \right)^\beta,$$

$\text{sim}(u, u')$  — оценка близости  $u$  и  $u'$  (чем ближе, тем больше).

## Необходимость изучающих действий в активном обучении

**Недостатки** стратегий активного обучения:

- остаются не обследованные области пространства  $X$ ,
- в результате снижается качество обучения,
- увеличивается время обучения.

**Идеи** применения изучающих действий:

- брать случайный объект с вероятностью  $\epsilon$
- адаптировать параметр  $\epsilon$  в зависимости от успешности изучающих действий
- использовать обучение с подкреплением

---

*Djallel Bouneffouf*. Exponentiated gradient exploration for active learning. 2016.  
*Djallel Bouneffouf et al.* Contextual bandit for active learning: active Thompson sampling. 2014.

## Алгоритм $\epsilon$ -active

Алгоритм — обёртка над любой стратегией активного обучения

**Вход:** размеченная выборка  $X^\ell = (x_i, y_i)_{i=1}^\ell$  и пул  $U = (u_i)_{i=1}^K$ ;

**Выход:** модель  $a$  и размеченная выборка  $(u_i, y_i^*)_{i=1}^k$ ;

обучить модель  $a$  по начальной выборке  $(x_i, y_i)_{i=1}^\ell$ ;

**пока** есть неразмеченные объекты и модель не обучилась

    выбрать неразмеченный  $u_i$  случайно с вероятностью  $\epsilon$ ,

    либо  $u_i = \arg \max_{u \in U} \phi(u)$  с вероятностью  $1 - \epsilon$ ;

    узнать  $y_i^* = y(u_i)$  для объекта  $u_i$ ;

    дообучить модель  $a$  ещё на одном примере  $(u_i, y_i^*)$ ;

**Проблема:**

как подбирать вероятность  $\epsilon$  исследовательских действий?

как её адаптировать (уменьшать) со временем?

## Экспоненциальный градиент (Exponential Gradient)

$\varepsilon_1, \dots, \varepsilon_H$  — сетка значений параметра  $\varepsilon$ ;

$p_1, \dots, p_H$  — вероятности использовать значения  $\varepsilon_1, \dots, \varepsilon_H$ ;

$\beta, \tau, \kappa$  — параметры метода.

**Идея** алгоритма EG-active: аналогично алгоритму AdaBoost, экспоненциально увеличивать  $p_h$  в случае успеха  $\varepsilon_h$ :

- экспоненциальное обновление весов  $w_h$  по значению критерия  $\phi(u_i)$  на выбранном объекте  $u_i$ :

$$w_h := w_h \exp\left(\frac{\tau}{p_h}(\phi(u_i) + \beta)\right);$$

- перенормировка вероятностей:

$$p_h := (1 - \kappa) \frac{w_h}{\sum_j w_j} + \kappa \frac{1}{H}.$$

*Djallel Bouneffouf*. Exponentiated gradient exploration for active learning. 2016.

## Алгоритм EG-active

**Вход:**  $X^\ell = (x_i, y_i)_{i=1}^\ell$ ,  $U = (u_i)_{i=1}^K$ , параметры  $\epsilon_1, \dots, \epsilon_H$ ,  $\beta$ ,  $\tau$ ,  $\kappa$ ;

**Выход:** модель  $a$  и размеченная выборка  $(u_i, y_i^*)_{i=1}^k$ ;

инициализация:  $p_h := \frac{1}{H}$ ,  $w_h := 1$ ;

обучить модель  $a$  по начальной выборке  $(x_i, y_i)_{i=1}^\ell$ ;

**пока** есть неразмеченные объекты и модель не обучилась

выбрать  $h$  из дискретного распределения  $(p_1, \dots, p_H)$ ;

выбрать неразмеченный  $u_i$  случайно с вероятностью  $\epsilon_h$ ,

либо  $u_i = \arg \max_{u \in U} \phi(u)$  с вероятностью  $1 - \epsilon_h$ ;

узнать  $y_i^*$  для объекта  $u_i$ ;

дообучить модель  $a$  ещё на одном примере  $(u_i, y_i^*)$ ;

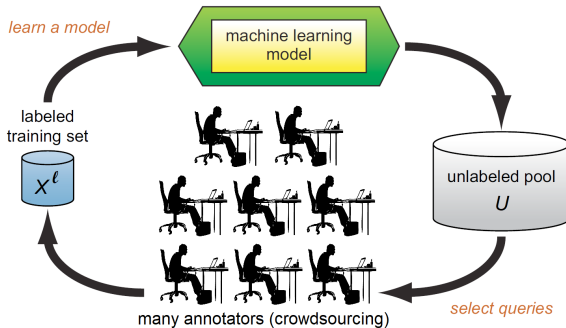
$w_h := w_h \exp\left(\frac{\tau}{p_h}(\phi(u_i) + \beta)\right)$ ;

$p_h := (1 - \kappa) \frac{w_h}{\sum_j w_j} + \kappa \frac{1}{H}$ ;

## Активное обучение, когда аннотаторов много

$y_{it}$  — ответы аннотаторов  $t \in T$  на объекте  $u_i$

**Задача:** сформировать согласованный «правильный» ответ  $\hat{y}_i$   
и оценить надёжность каждого аннотатора  $q_t = P[y_{it} = \hat{y}_i]$



Р.А.Гилязев, Д.Ю.Турдаков. Активное обучение и краудсорсинг: обзор методов оптимизации разметки данных. 2018.



## Согласование оценок аннотаторов

$y_{it} \in Y$  — ответ аннотатора  $t \in T$  на объекте  $u_i$

$T_i \subseteq T$  — множество аннотаторов, разметивших объект  $u_i$

Взвешенное голосование аннотаторов:

$$\hat{y}_i = \arg \max_{y \in Y} \sum_{t \in T_i} w_t [y_{it} = y]$$

$w_t$  — вес аннотатора при голосовании

$w_t = 1$  при голосовании по большинству (majority voting, MV)

$w_t = \log \frac{q_t}{1-q_t}$  при предположении, что аннотаторы независимы

EM-подобный алгоритм согласования аннотаций объекта  $u_i$ :

**пока** оценки не сойдутся

оценить правильный ответ  $\hat{y}_i$ ;

оценить надёжности  $q_t$  и веса  $w_t$  аннотаторов;

**если**  $q_t < \delta$  **то** исключить аннотатора из оценки;

## Варианты моделирования надёжности аннотаторов

- По результатам выполнения тестовых заданий.
- Моделирование матрицы ошибок  $|Y| \times |Y|$ :

$$\pi_{yz}^t = P[\text{аннотатор } t \text{ ставит } z \text{ вместо } y], \quad y, z \in Y$$

- Моделирование трудности объектов:

$$q_t(u_i) = \sigma\left(\frac{\alpha_t}{\beta_i}\right) = \frac{1}{1 + \exp\left(-\frac{\alpha_t}{\beta_i}\right)},$$

$\alpha_t$  — частотная оценка надёжности аннотатора  $t$ ;

$\beta_i$  — оценка трудности объекта  $u_i$  (по большому  $|T_i|$ ).

- Моделирование тематической компетентности аннотаторов:  
 $p(\text{topic} | u_i)$  — тематическое векторное представление объекта  $u_i$ , например, если объект является текстом

---

*Р.А.Гилязев, Д.Ю.Турдаков. Активное обучение и краудсорсинг: обзор методов оптимизации разметки данных. 2018.*

## Задача назначения заданий аннотаторам

Общая схема распределения заданий:

$$\begin{cases} u_i = \arg \max_{u \in U} \phi(u) & \text{— выбор неразмеченного объекта в AL} \\ t = \arg \max_{t \in T} q_t(u_i) & \text{— выбор наиболее уверенного аннотатора} \end{cases}$$

Обучение вероятностной модели уверенности аннотатора

$q_t(u_i, \theta_t) = \sigma(\theta_t^\top u_i)$  на размеченных им объектах  $U_t$ :

$$\sum_{u_i \in U_t} (y_{it} = \hat{y}_i) q_t(u_i, \theta_t) + (y_{it} \neq \hat{y}_i) (1 - q_t(u_i, \theta_t)) \rightarrow \max_{\theta_t}$$

**Недостаток:** одни аннотаторы будут выбираться слишком часто, другие не будут выбираться совсем

**Сэмплирование аннотаторов:**  $t \sim q_t(u_i)p(t)$  с учётом априорной информации  $p(t)$  о средней надёжности  $q_t$ , опыте, текущей доступности, объёме проделанной работы.

- Активное обучение используется для уменьшения обучающей выборки, когда размеченные данные дороги
- При малом объёме размеченных данных оно достигает того же качества, что пассивное при полной разметке
- Два основных типа активного обучения:  
выбор объектов из пула и синтез новых объектов
- Введение изучающих действий в активном обучении позволяет ещё быстрее обследовать пространство  $X$
- В краудсорсинге активное обучение совмещается с оцениванием надёжности аннотаторов и трудности заданий при распределении заданий по аннотаторам

---

*P.Kumar, A.Gupta. Active learning query strategies for classification, regression, and clustering: a survey. 2020*

*Pengzhen Ren et al. A survey of deep active learning. 2020*

*Burr Settles. Active learning literature survey. 2010*

*C.C.Aggarwal et al. Active learning: a survey // Data classification: algorithms and applications. 2014*

*Р.А.Гилязев, Д.Ю.Турдаков. Активное обучение и краудсорсинг: обзор методов оптимизации разметки данных. 2018*