• Прикладные модели машинного обучения •
Активное обучение (Active Learning)

Воронцов Константин Вячеславович
k.v.vorontsov@phystech.edu
http://www.MachineLearning.ru/wiki?title=User:Vokov

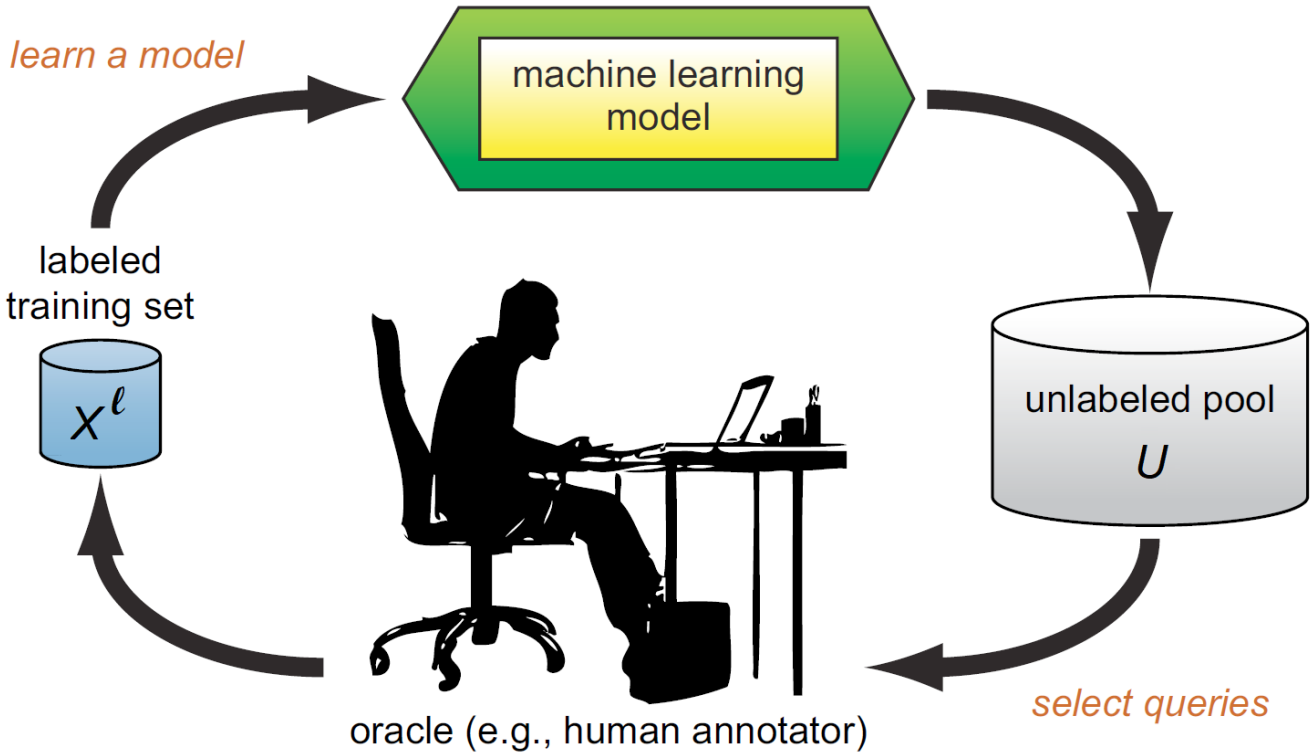
Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

МФТИ • 18 апреля 2025

Содержание

Постановка задачи активного обучения

Задача: обучение модели $a\colon X \to Y$ 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 $a\colon X \to Y$ 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;
$U = (u_i)_{i=1}^K$ — пул неразмеченных объектов;
Выход: модель a и размеченная выборка $(u_i, y_i^*)_{i=1}^k$, $k \leqslant K$;

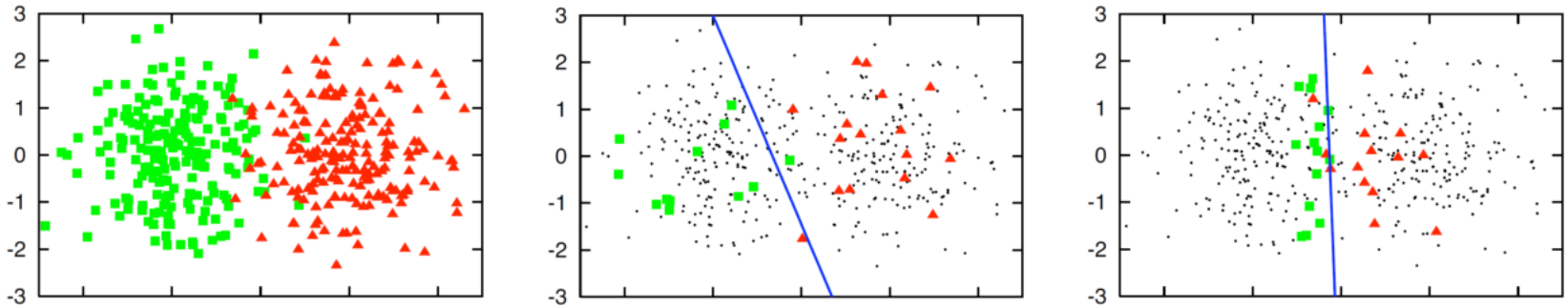
обучить модель a по начальной выборке $(x_i, y_i)_{i=1}^\ell$;
пока есть неразмеченные объекты и модель не обучилась
- $u_i = \arg\max\limits_{u \in U} \phi(u)$ — максимум оценки перспективности;
- узнать для него $y_i^* = y(u_i)$;
- дообучить модель $a(x)$ ещё на одном примере (u_i, y_i^*);

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

Почему активное обучение быстрее пассивного

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

(a) два гауссовских класса;
логистическая регрессия по 30 объектам:
(b) случайным;
(c) отобранным по максимуму неуверенности классификации.



(a) (b) (c)

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

Burr Settles. Active Learning Literature Survey. 2010.

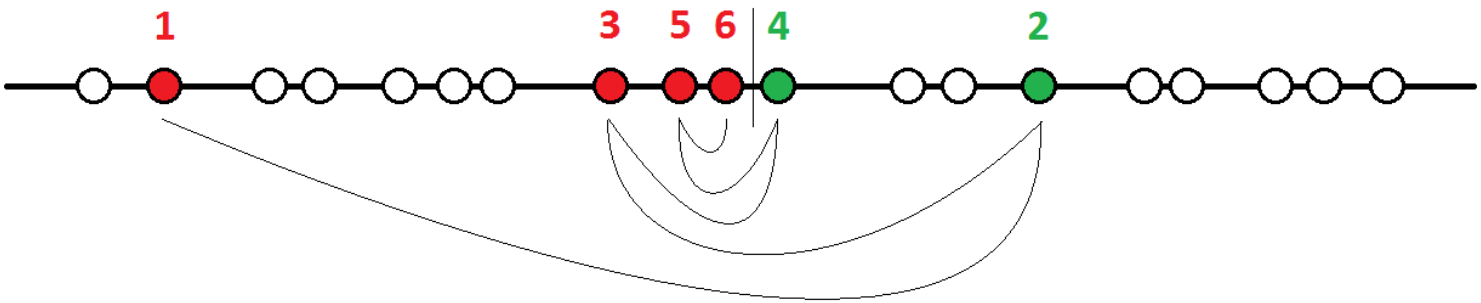
Почему активное обучение быстрее пассивного

Пример 2. Одномерная задача с пороговым классификатором:

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

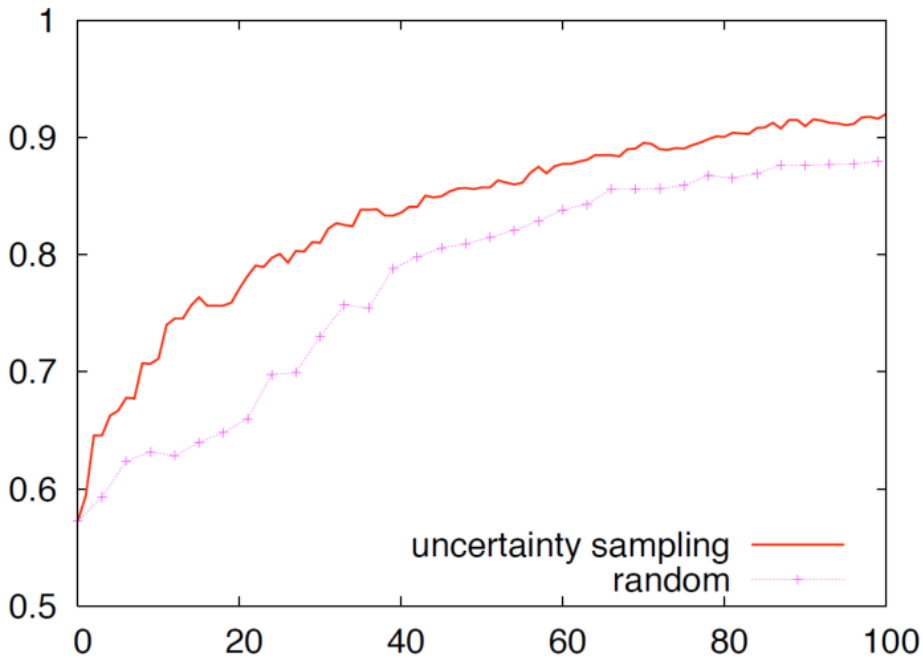
Оценим число шагов для определения θ с точностью $\frac{1}{k}$.

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



Оценивание качества активного обучения

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



Идея: строить кривые обучения отдельно по классам, чтобы размечать объекты, вероятнее лежащие в проблемных классах

Burr Settles. Active Learning Literature Survey. 2010.

Стратегии активного обучения

- Отбор объектов из выборки (pool-based sampling):
  какой следующий u_i выбрать из пула $U = \{u_i\}_{i=1}^{K}$
- 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^*

Методы, чувствительные к стоимости (cost-sensitive):

$$\sum_{i=1}^{\ell} \mathscr{L}(x_i, y_i; \theta) + \sum_{u_i \in U_k} \big(C_i + \mathscr{L}(u_i, y_i^*; \theta)\big) \;\to\; \min_{\theta, U_k},$$

$\mathscr{L}(x, y; \theta)$ — функция *стоимости* потерь для модели $a(x, \theta)$,
C_i — *стоимость* получения разметки $y_i^* = y(u_i)$,
$U_k \subset U$ — подмножество из k размечаемых объектов

Примеры приложений активного обучения

- сбор асессорских данных для информационного поиска, анализа текстов, сигналов, речи, изображений, видео
- в том числе на платформах краудсорсинга
- *планирование экспериментов* в естественных науках или на производстве (пример — комбинаторная химия)
- оптимизация трудно вычислимых функций (пример — оптимизация гиперпараметров, AutoML)

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 неуверенности (uncertainty sampling)

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

Задача многоклассовой классификации:

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

$p_m(u)$, $m = 1...|Y|$ — 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* (minimum margin):

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

- Принцип *максимума энтропии* (maximum entropy):

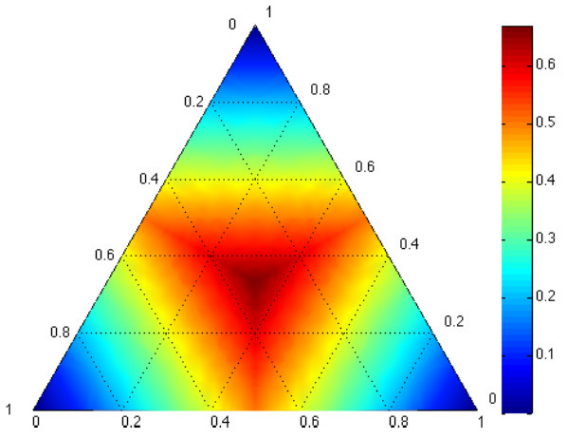
$$u_i = \arg\min_{u \in U} \sum_m p_m(u) \ln p_m(u)$$

Сэмплирование по неуверенности (uncertainty sampling)

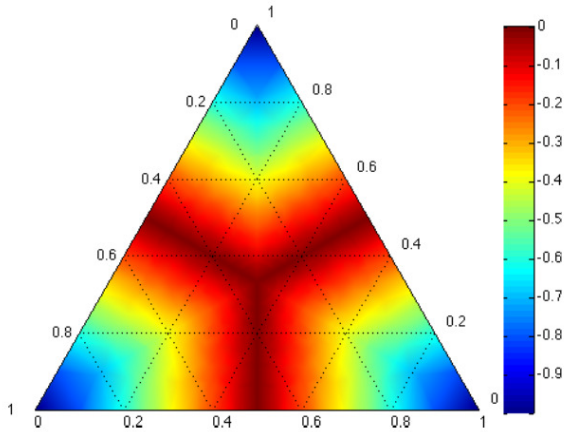
В случае двух классов эти три принципа эквивалентны.

В случае многих классов появляются различия.

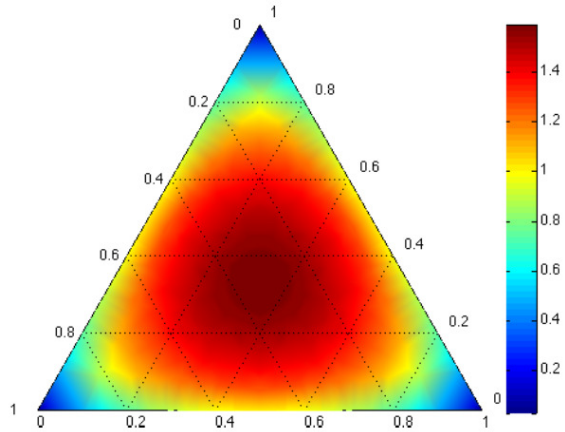
Пример. Три класса, $p_1 + p_2 + p_3 = 1$.
Показаны линии уровни трёх критериев выбора объекта:



$\min p_1$



$\min(p_1 - p_2)$



$\min \sum_m p_m \ln p_m$

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 $a_t(u_i) = \arg\max\limits_{y \in Y} P_t(y|u_i), \quad t = 1, \ldots, T.$

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

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

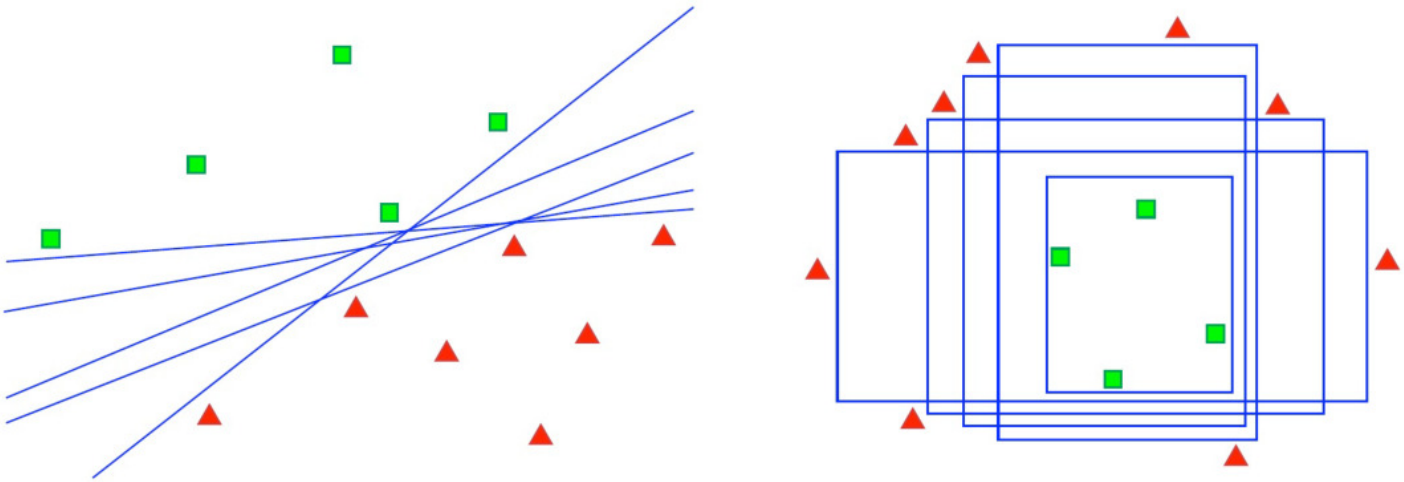
$$u_i = \arg\max_{u \in U} \sum_{t=1}^{T} \mathrm{KL}\big(P_t(y|u) \,\big\|\, \bar{P}(y|u)\big),$$

  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 θ при дообучении модели на (u, y); пусть $\nabla_\theta \mathscr{L}(u, y; \theta)$ — вектор градиента функции потерь.

Принцип *максимума ожидаемой длины градиента:*

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

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^ℓ пример (u, y):

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

- Принцип *максимума уверенности на неразмеченных данных:*

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

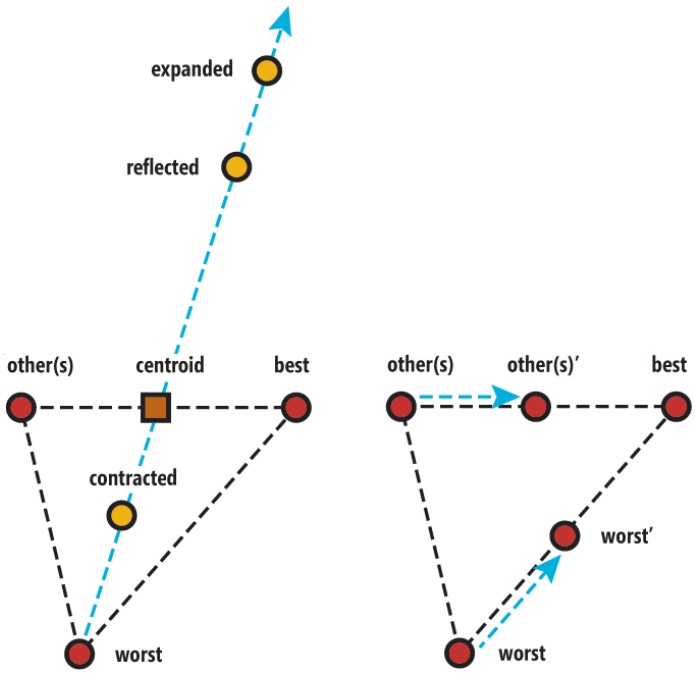
- Принцип *минимума энтропии неразмеченных данных*:

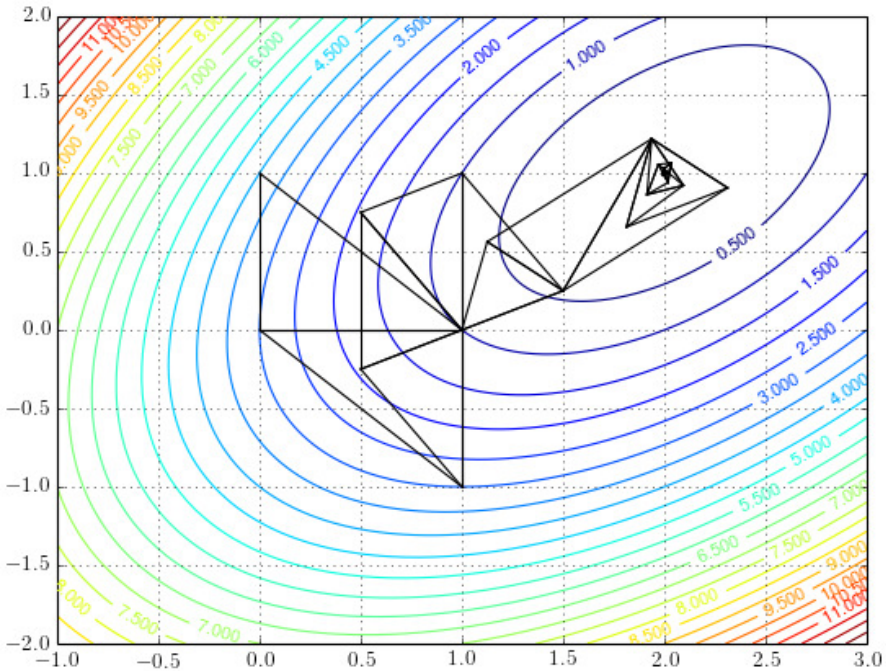
$$u_i = \arg\max_{u \in U} \sum_{y \in Y} P(y|u) \sum_{u_j \in U \setminus u} \sum_{z \in Y} P_{uy}(z|u_j) \log P_{uy}(z|u_j).$$

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





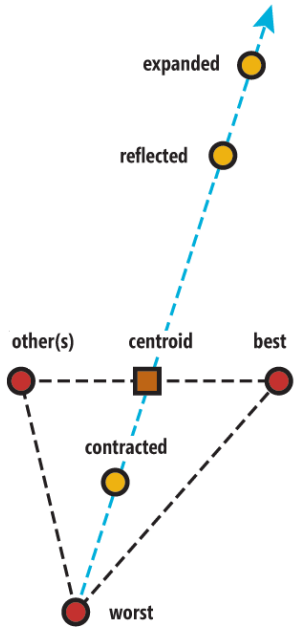
J.A.Nelder, R.Mead. A simplex method for function minimization. 1965.

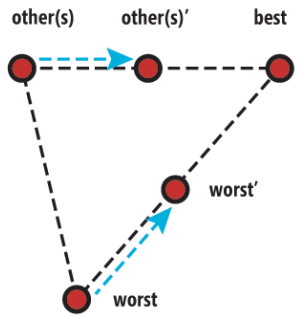
Метод Нелдера–Мида: «отражение–растяжение–сжатие»

повторять

- сортировка $n+1$ точек: $\phi(x_w) < \cdots < \phi(x_b)$;
- центроид грани 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** $\phi(x_w) < \phi(x_r) < \phi(x_b)$ **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: $x_i := x_b + \sigma(x_i - x_b)$;

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 $\mathsf{E}_x \sigma^2_{a(x)}$ после дообучения модели $a(x,\theta)$ на u.

Метод наименьших квадратов, линейная регрессия $a(x,\theta) = x^\mathsf{T}\theta$:

$$S^2(\theta) = \frac{1}{\ell}\sum_{i=1}^{\ell}\big(a(x_i,\theta) - y_i\big)^2 = \frac{1}{\ell}\|F\theta - y\|^2 \to \min_\theta$$

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

$$\sigma^2_{a(x)} \approx S^2\left(\frac{\partial a(x)}{\partial\theta}\right)^\mathsf{T}\left(\frac{\partial S^2}{\partial\theta^2}\right)^{-1}\left(\frac{\partial a(x)}{\partial\theta}\right) = S^2 x^\mathsf{T}(F^\mathsf{T}F)^{-1}x$$

Принцип *сокращения дисперсии* для линейной регрессии:

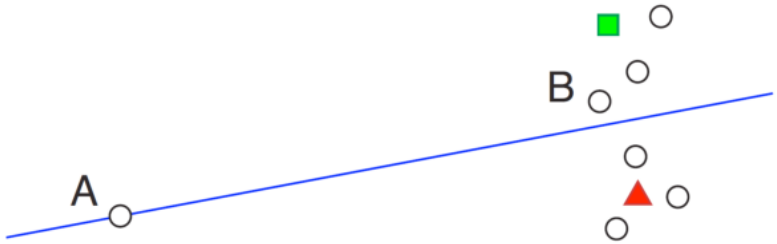
$$u = \arg\min_{u\in X} \textstyle\sum_i x_i^\mathsf{T}(F^\mathsf{T}F + u^\mathsf{T}u)^{-1}x_i$$

Burr Settles. Active learning. Morgan & Claypool Publ., 2012.

Взвешивание по плотности (density-weighted methods)

Идея: понижать вес нерепрезентативных объектов.

Пример. Объект A более пограничный, но менее репрезентативный, чем B.



Любой критерий выбора объектов, имеющий вид

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

может быть уточнён локальной оценкой плотности:

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

$\mathrm{sim}(u, u')$ — 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 ε
- адаптировать параметр ε — уменьшать со временем, в зависимости от успешности изучающих действий
- можно 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 ε-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 $\begin{cases} u_i \text{ неразмеченный}, & \text{с вероятностью } \varepsilon; \\ u_i = \arg\max\limits_{u \in U} \phi(u), & \text{с вероятностью } 1-\varepsilon; \end{cases}$

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 ε исследовательских действий?
как её адаптировать (уменьшать) со временем?

Экспоненциальный градиент (Exponential Gradient)

$\varepsilon_1, \ldots, \varepsilon_H$ — сетка значений параметра ε;
$p_1, \ldots, p_H$ — вероятности использовать значения $\varepsilon_1, \ldots, \varepsilon_H$;
β, τ, κ — параметры метода.

Идея алгоритма EG-active: аналогично алгоритму AdaBoost, экспоненциально увеличивать p_h в случае успеха ε_h:

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

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

- 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

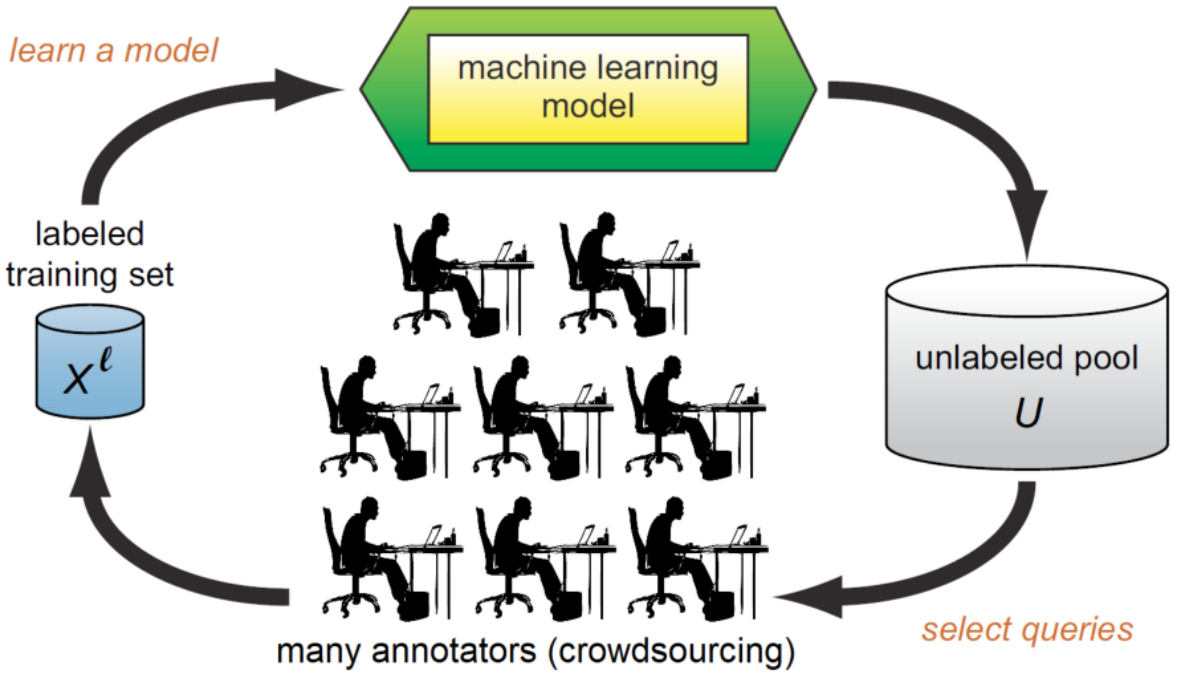
Вход: $X^\ell = (x_i, y_i)_{i=1}^\ell$, $U = (u_i)_{i=1}^K$, параметры $\varepsilon_1, \ldots, \varepsilon_H$, β, τ, κ;
Выход: модель a и размеченная выборка $(u_i, y_i^*)_{i=1}^k$;

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 $(p_1, \ldots, p_H)$;
- выбрать $\begin{cases} u_i \text{ неразмеченный}, & \text{с вероятностью } \varepsilon_h; \\ u_i = \arg\max\limits_{u \in U} \phi(u), & \text{с вероятностью } 1 - \varepsilon_h; \end{cases}$
- 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 ответ (консенсус) $\hat{y}_i$ 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

Консенсус — взвешенное голосование аннотаторов:

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

w_t — вес аннотатора при голосовании, например,
$w_t = 1$ при голосовании по большинству (majority voting, MV)
$w_t = \log \frac{q_t}{1-q_t}$ при предположении, что аннотаторы независимы

EM-like алгоритм согласования аннотаций объекта u_i:

пока оценки не сойдутся
  оценить консенсус $\hat{y}_i$;
  надёжности $q_t := P[y_{it} = \hat{y}_i]$ 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 $|Y|\times|Y|$:

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

- Моделирование трудности объектов:

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

α_t — частотная оценка надёжности аннотатора t;
β_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:

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

Обучение вероятностной модели уверенности аннотатора $q_t(u_i, \theta_t) = \sigma(\theta_t^{\mathsf{T}} u_i)$ на размеченных им объектах U_t:

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

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, компетенции, доступности, загруженности, плане работ.

Пример. Разметка текстов в конкурсе ПРО//ЧТЕНИЕ

Задача: поиск смысловых ошибок в сочинениях ЕГЭ по русскому, литературе, обществознанию, истории, английскому

Алгоритм должен выделять ошибки и давать их объяснения

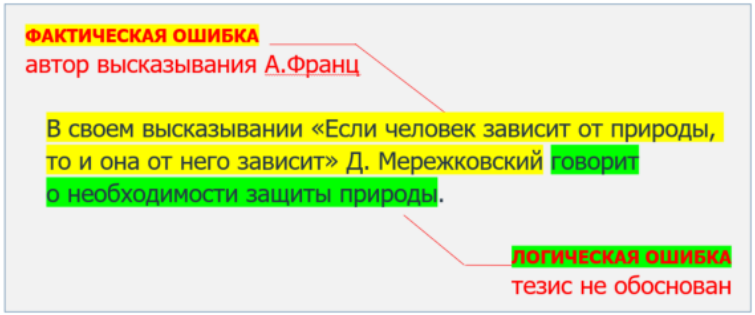
Типов ошибок: 152
(р:70 л:16 о:23 и:20 а:23)
Подтипов ошибок: 236
(р:112 л:19 о:29 и:26 а:50)

Призовой фонд:
— 100M руб. русский язык
— 100M руб. английский язык

Период: дек 2019 – дек 2022

Up Great technology contest

READ // ABLE
comprehension technology



Технический регламент конкурса ПРО//ЧТЕНИЕ: http://ai.upgreat.one

Многокритериальное сравнение двух разметок

1. отличие оценки за сочинение
2. соответствие фрагментов ошибок и смысловых блоков
3. отличие границ сопоставленных фрагментов
4. отличие типов и подтипов ошибок
5. качество сгенерированных алгоритмом пояснений

Оценивание модели разметки по нескольким аннотаторам

$\mathrm{Con}_k(a_i, y_{it})$ — меры согласованности разметок алгоритма a_i и экспертов y_{it} после оптимального сопоставления их фрагментов

$\mathrm{Con}(a_i, y_{it})$ — средневзвешенная согласованность разметок

- *Средняя Точность Алгоритмической Разметки*:
  $$\mathrm{СТАР}(a) = \operatorname*{mean}_{x_i \in X^\ell} \operatorname*{mean}_{t \in T_i} \mathrm{Con}(a_i, y_{it})$$
- *Средняя Точность Экспертных Разметок*:
  $$\mathrm{СТЭР} = \operatorname*{mean}_{x_i \in X^\ell} \operatorname*{mean}_{t,s \in T_i} \mathrm{Con}(y_{it}, y_{is})$$
- *Относительная Точность Алгоритмической Разметки*:
  $$\mathrm{ОТАР}(a) = \frac{\mathrm{СТАР(a)}}{\mathrm{СТЭР}}$$

Условие преодоления технологического барьера в конкурсе:
$\mathrm{ОТАР} \geqslant 1$ означает, что алгоритм работает не хуже экспертов

Технический регламент конкурса ПРО//ЧТЕНИЕ (http://ai.upgreat.one)

Резюме

- Активное обучение используется для уменьшения обучающей выборки, когда размеченные данные дóроги
- 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 быстрее обследовать пространство объектов
- Модели надёжности аннотаторов и трудности заданий позволяют лучше распределять задания в краудсорсинге
- Разметка каждого объекта несколькими аннотаторами позволяет оценить «естественный уровень шума» в задаче

Роберт (Манро) Монарх. Машинное обучение с участием человека. 2022
Burr Settles. Active learning literature survey. 2010
P.Kumar, A.Gupta. Active learning query strategies for classification, regression, and clustering: a survey. 2020
C.C.Aggarwal et al. Active learning: a survey. 2014
Pengzhen Ren et al. A survey of deep active learning. 2020