Активное обучение (Active Learning)

К. В. Воронцов
vokov@forecsys.ru

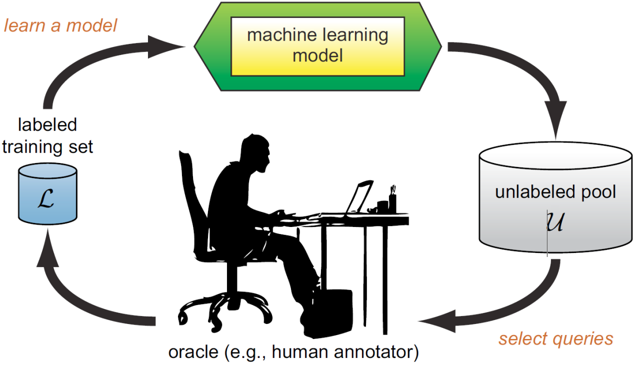
Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

1 ноября 2017

Содержание

Постановка задачи активного обучения

Задача: обучение предсказательной модели $a\colon X \to Y$ по выборке (x_i, y_i), когда получение ответов y_i стоит дорого.



Burr Settles. Active Learning Literature Survey. 2010.

Постановка задачи активного обучения

Задача: обучение предсказательной модели $a\colon X \to Y$
по выборке (x_i, y_i), когда получение ответов y_i стоит дорого.

Вход: начальная размеченная выборка $X^\ell = (x_i, y_i)_{i=1}^{\ell}$;
Выход: модель a и размеченная выборка $(x_i, y_i)_{i=\ell+1}^{\ell+k}$;

обучить модель a по начальной выборке $(x_i, y_i)_{i=1}^{\ell}$;
пока остаются неразмеченные объекты
  выбрать неразмеченный объект x_i;
  узнать для него y_i;
  дообучить модель a ещё на одном примере (x_i, y_i);

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

Примеры приложений активного обучения

- сбор асессорских данных для информационного поиска, анализа текстов, сигналов, речи, изображений, видео
- планирование экспериментов в естественных науках (пример — комбинаторная химия)
- оптимизация трудно вычислимых функций (пример — поиск в пространстве гиперпараметров)
- управление ценами и ассортиментом в торговых сетях
- выбор товара для проведения маркетинговой акции

Стратегии активного обучения

- Отбор объектов из выборки (pool-based sampling):
  какой следующий x_i выбрать из множества $X^k = \{x_i\}_{i=\ell+1}^{\ell+k}$
- Синтез объектов (query synthesis):
  на каждом шаге построить оптимальный объект x_i
- Отбор объектов из потока (selective sampling):
  для каждого приходящего x_i решать, стоит ли узнавать y_i

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

$$\sum_{i=1}^{\ell+k} C_i \mathscr{L}(\theta; x_i, y_i) \to \min_{\theta},$$

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

Сэмплирование по неуверенности (uncertainty sampling)

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

Задача многоклассовой классификации:

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

$p_k(x)$, $k = 1 \dots |Y|$ — ранжированные по убыванию $P(y|x)$, $y \in Y$.

- Принцип *наименьшей достоверности* (least confidence):

$$x_i = \arg\min_{u \in X^k} p_1(u)$$

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

$$x_i = \arg\min_{u \in X^k} \bigl(p_1(u) - p_2(u)\bigr)$$

- Принцип *максимума энтропии* (maximum entropy):

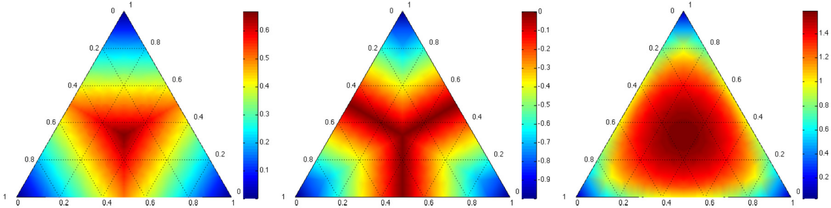
$$x_i = \arg\min_{x \in X^k} \sum_k p_k(u) \ln p_k(u)$$

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 x_i:



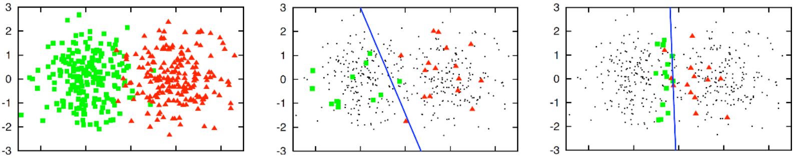
$\min p_1$ $\min(p_1 - p_2)$ $\min \sum\limits_k p_k \ln p_k$

Burr Settles. Active Learning Literature Survey. 2010.

Почему активное обучение быстрее пассивного

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

(a) два гауссовских класса;
(b) логистическая регрессия по 30 случайным объектам;
(c) 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.

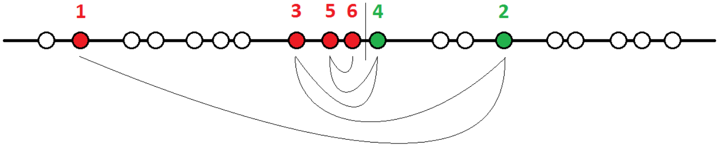
Почему активное обучение быстрее пассивного

Пример 2. Одномерная задача с пороговым классификатором:

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

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

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



Сэмплирование по несогласию в комитете (query by committee)

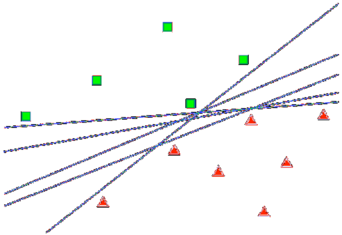
Идея: выбирать x_i с наибольшей несогласованностью решений комитета моделей $a_t(x_i) = \arg\max\limits_{y \in Y} P_t(y|x)$, $\; t = 1, \ldots, T$.

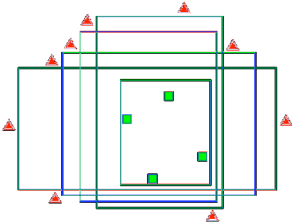
- Принцип *максимума энтропии*:
  выбираем x_i, на котором $a_t(x_i)$ максимально различны:
  $$x_i = \arg\min_{u \in X^k} \sum_{y \in Y} \hat{p}(y|u) \ln \hat{p}(y|u),$$
  где $\hat{p}(y|u) = \frac{1}{T} \sum_{t=1}^{T} \bigl[a_t(u) = y\bigr]$.
- Принцип *максимума средней KL-дивергенции*:
  выбираем x_i, на котором $P_t(y|x_i)$ максимально различны:
  $$x_i = \arg\max_{u \in X^k} \sum_{t=1}^{T} \mathrm{KL}\bigl(P_t(y|u) \,\|\, \bar{P}(y|u)\bigr),$$
  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 x_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 x_i, который в методе стохастического градиента привёл бы к наибольшему изменению модели.

Параметрическая модель многоклассовой классификации:

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

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

Принцип *максимума ожидаемой длины градиента:*

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

Ожидаемое сокращение ошибки (expected error reduction)

Идея: выбрать x_i, который после обучения даст наиболее уверенную классификацию неразмеченной выборки X^k.

Для каждого $u \in X^k$ 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:*

$$x_i = \arg\max_{u \in X^k} \sum_{y \in Y} P(y|u) \sum_{j=\ell+1}^{\ell+k} P_{uy}\bigl(a_{uy}(x_j)|x_j\bigr).$$

- Принцип *минимума энтропии неразмеченных данных*:

$$x_i = \arg\max_{u \in X^k} \sum_{y \in Y} P(y|u) \sum_{j=\ell+1}^{\ell+k} \sum_{z \in Y} P_{uy}(z|x_j) \log P_{uy}(z|x_j).$$

Сокращение дисперсии (variance reduction)

Идея: выбрать x, который после дообучения модели $a(x,\theta)$ даст наименьшую оценку дисперсии $\sigma_a^2(x)$.

Задача регрессии, метод наименьших квадратов:

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

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

$$x = \arg\min_{x\in X} \sigma_a^2(x), \quad \sigma_a^2(x) \approx S^2 \left(\frac{\partial a(x)}{\partial\theta}\right)^{\mathsf{T}} \left(\frac{\partial S^2}{\partial\theta^2}\right)^{-1} \left(\frac{\partial a(x)}{\partial\theta}\right).$$

В частности, для линейной регрессии

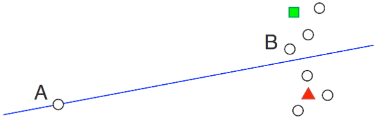
$$\sigma_a^2(x) \approx S^2 x^{\mathsf{T}}(F^{\mathsf{T}}F)^{-1}x,$$

где F — матрица объекты–признаки.

Взвешивание по плотности (density-weighted methods)

Идея: понижать вес нерепрезентативных объектов.

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



Любой критерий сэмплирования объектов, имеющий вид

$$x_i = \arg\max_x \phi(x),$$

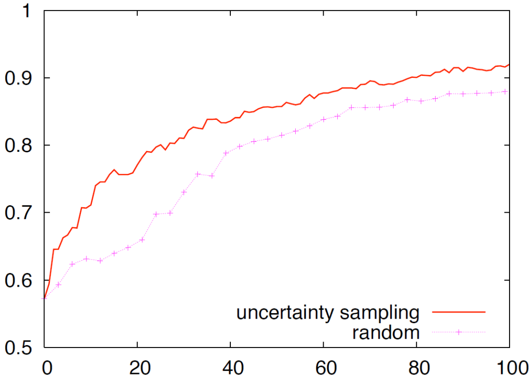
может быть уточнён локальной оценкой плотности:

$$x_i = \arg\max_x \phi(x) \left(\sum_{j=\ell+1}^{\ell+k} \mathrm{sim}(x, x_j) \right)^\beta,$$

$\mathrm{sim}(x, x_j)$ — оценка близости x и x_j (чем ближе, тем больше).

Оценивание качества активного обучения

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



Burr Settles. Active Learning Literature Survey. 2010.

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 в зависимости от успешности изучающих действий
- использовать обучение с подкреплением (contextual MAB)

Djallel Bouneffouf et al. Contextual bandit for active learning: active Thompson sampling. 2014.
Djallel Bouneffouf. Exponentiated Gradient Exploration for Active Learning. 2016.

Алгоритм ε-active

Алгоритм — обёртка над любой стратегией активного обучения

Вход: начальная размеченная выборка $X^\ell = (x_i, y_i)_{i=1}^{\ell}$;
Выход: модель a и размеченная выборка $(x_i, y_i)_{i=\ell+1}^{\ell+k}$;

обучить модель a по начальной выборке $(x_i, y_i)_{i=1}^{\ell}$;
пока остаются неразмеченные объекты
- выбрать неразмеченный x_i случайно с вероятностью ε,
- либо $x_i = \arg\max\limits_{x} \phi(x)$ с вероятностью $1 - \varepsilon$;
- узнать y_i для объекта x_i;
- дообучить модель a ещё на одном примере (x_i, y_i);

Проблема:
как подбирать вероятность ε исследовательских действий?
как её адаптировать (уменьшать) со временем?

Экспоненциальный градиент (Exponential Gradient)

$\varepsilon_1, \dots, \varepsilon_K$ — сетка значений параметра ε;
$p_1, \dots, p_K$ — вероятности использовать значения $\varepsilon_1, \dots, \varepsilon_K$;
β, τ, κ — параметры метода.

Идея алгоритма EG-active: аналогично алгоритму AdaBoost, экспоненциально увеличивать p_k в случае успеха ε_k:

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

$$w_k := w_k \exp\left(\frac{\tau}{p_k}(\phi(x_i) + \beta)\right);$$

- перенормировка вероятностей:

$$p_k := (1 - \kappa)\frac{w_k}{\sum_j w_j} + \kappa\frac{1}{K}.$$

Djallel Bouneffouf. Exponentiated Gradient Exploration for Active Learning. 2016.

Алгоритм EG-active

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

инициализация: $p_k := \frac{1}{K}$, $w_k := 1$;
обучить модель a по начальной выборке $(x_i, y_i)_{i=1}^\ell$;
пока остаются неразмеченные объекты
- выбрать k из дискретного распределения $(p_1, \ldots, p_K)$;
- выбрать неразмеченный x_i случайно с вероятностью ε_k,
- либо $x_i = \arg\max\limits_x \phi(x)$ с вероятностью $1 - \varepsilon_k$;
- узнать y_i для объекта x_i;
- дообучить модель a ещё на одном примере (x_i, y_i);
- $w_k := w_k \exp\left(\frac{\tau}{p_k}(\phi(x_i) + \beta)\right)$;
- $p_k := (1 - \kappa)\frac{w_k}{\sum_j w_j} + \kappa\frac{1}{K}$;

Применение обучения с подкреплением для активного обучения

Недостатки стратегий активного обучения:

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

Идеи применения контекстного бандита (contextual MAB):

- *действия* (ручки) — это кластеры объектов,
- *контекст* кластера — его векторное признаковое описание,
- *премия* поощряет изменение модели $a(x, \theta)$,
- *линейная модель* используется для выбора действий.

Djallel Bouneffouf et al. Contextual bandit for active learning: active Thompson sampling. 2014.
William R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. 1933.

Томпсоновское сэмплирование (Thompson sampling)

C — множество действий (ручек, кластеров объектов),
$b_{tc} \in \mathbb{R}^n$ — вектор признаков кластера $c \in C$ на шаге t,
$w \in \mathbb{R}^n$ — вектор коэффициентов линейной модели.

Игра агента и среды (contextual bandit with linear payoff):

инициализация априорного распределения $p_1(w)$;
для всех $t = 1, \dots, T$
- среда сообщает агенту контексты b_{tc} для всех $c \in C$;
- агент сэмплирует вектор линейной модели $w_t \sim p_t(w)$;
- агент выбирает действие $c_t = \arg\max\limits_{c \in C} \langle b_{tc}, w_t \rangle$;
- среда генерирует премию r_t;
- агент корректирует распределение по формуле Байеса: $p_{t+1}(w) \propto p(r_t|w)p_t(w)$;

Томпсоновское сэмплирование (гауссовский случай)

Априорные и апостериорные распределения — гауссовские.

Игра агента и среды (contextual bandit with linear payoff):

инициализация: $B = I_{n\times n}$; $w = 0_n$; $f = 0_n$;

для всех $t = 1, \dots, T$

- среда сообщает агенту контексты b_{tc} для всех $c \in C$;
- агент сэмплирует вектор линейной модели $w_t \sim \mathcal{N}(w, \sigma^2 B^{-1})$;
- агент выбирает действие $c_t = \arg\max\limits_{c\in C} \langle b_{tc}, w_t \rangle$;
- среда генерирует премию r_t;
- агент корректирует распределение по формуле Байеса: $B := B + b_{tc} b_{tc}^{\mathsf{T}}$; $f := f + b_{tc} r_t$; $w := B^{-1} f$;

Рекомендуемое значение константы $\sigma^2 = 0.25$.

Активное томпсоновское сэмплирование

Игра агента и среды (встраиваем активное обучение)

$C :=$ кластеризация неразмеченной выборки X^k;
инициализация: $B = I_{n\times n}$; $w = 0_n$; $f = 0_n$;
для всех $t = 1, \dots, T$, пока остаются неразмеченные объекты
- вычислить контексты b_{tc} для всех кластеров $c \in C$;
- сэмплировать вектор линейной модели $w_t \sim \mathcal{N}(w, \sigma^2 B^{-1})$;
- выбрать кластер $c_t = \arg\max\limits_{c\in C} \langle b_{tc}, w_t\rangle$;
- выбрать случайный неразмеченный x_i из кластера c_t;
- узнать для него y_i;
- дообучить модель a ещё на одном примере (x_i, y_i);
- вычислить премию r_t (формула на следующем слайде);
- скорректировать распределение по формуле Байеса:
  $B := B + b_{tc} b_{tc}^{\mathsf{T}}$; $f := f + b_{tc} r_t$; $w := B^{-1} f$;

Как вычисляются премии

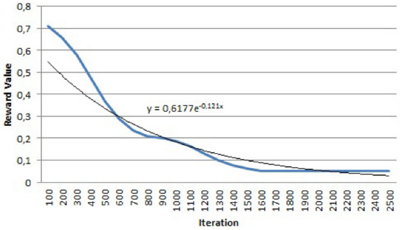
Идея: премия поощряет изменение модели $a(x,\theta)$.

$H_t = \bigl(a(x_i,\theta_t)\bigr)_{i=1}^{\ell+k}$ — вектор ответов на выборке $X^\ell \cup X^k$

Премия — угол между векторами H_t и H_{t-1}:

$$r_t := e^{\beta t} \arccos \frac{\langle H_t, H_{t-1}\rangle}{\|H_t\|\,\|H_{t-1}\|},$$

где экспоненциальный множитель компенсирует убывание расстояний;
$\beta = 0.121$ — эмпирически подобранный параметр.



Djallel Bouneffouf et al. Contextual bandit for active learning: active Thompson sampling. 2014.

Как вычисляются признаки контекстов (кластеров)

$b_{tc} = \bigl(\mathrm{Mdis}_c, \mathrm{Vdis}_c, |c|, \mathrm{plb}_{tc}, \mathrm{MixRate}_{tcy}\bigr)$

- Mdis_c — среднее внутрикластерное расстояние;
- Vdis_c — дисперсия внутрикластерных расстояний;
- $|c|$ — число объектов в кластере;
- plb_{tc} — доля размеченных объектов в кластере;
- $\mathrm{MixRate}_{tcy}$ — доля объектов класса $y \in Y$ в кластере.

Всего признаков: $4 + |Y|$.

Djallel Bouneffouf et al. Contextual bandit for active learning: active Thompson sampling. 2014.

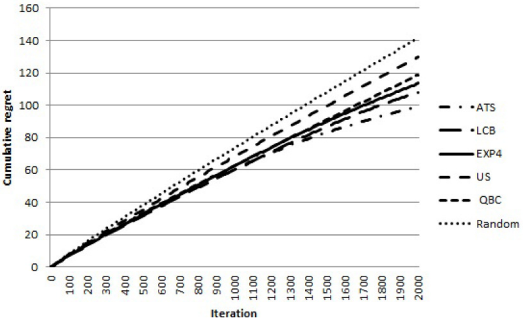
Как оценивается качество

Накопленные потери (cumulative regret):

$$R(T) = \sum_{t=1}^{T} \bigl(\langle b_{tc_t^*}, w_t\rangle - \langle b_{tc}, w_t\rangle\bigr),$$

c_t^* — оптимальное действие ($R = 0$, если все действия оптимальны)

Сравнение накопленных потерь для различных алгоритмов:



Резюме

- Активное обучение используется для уменьшения обучающей выборки, когда размеченные данные до́роги
- Активное обучение быстрее пассивного
- При малом объёме размеченных данных оно достигает того же качества, что пассивное при полной разметке
- Введение изучающих действий в активном обучении позволяет ещё быстрее обследовать пространство X
- Для этого в последние годы стали применяться адаптивные стратегии или обучение с подкреплением