Активное обучение

Понизова Вероника Сергеевна
Федяев Игорь Павлович

Санкт-Петербургский Государственный Университет
Прикладная математика и информатика
Статистическое моделирование.

Санкт-Петербург
2019г.

Пусть $\mathcal{X}$ — множество объектов, $\mathcal{Y}$ — множество ответов, $y : \mathcal{X} \to \mathcal{Y}$ — неизвестная зависимость.

Задача: найти функцию $a : \mathcal{X} \to \mathcal{Y}$, приближающую y на всем множестве $\mathcal{X}$ (обучить предсказательную модель), когда в выборке есть неразмеченные объекты, при условии, что получение ответов y_i стоит дорого.

Примеры:

- сбор асессорских данных для информационного поиска структурно сложных объектов;
- планирование экспериментов в естественных науках (комбинаторная химия);
- оптимизация трудновычислимых функций (поиск в пространстве гиперпараметров);
- управление ценами и ассортиментами в торговых сетях.

Вход: начальная размеченная выборка $X^l = (x_i, y_i)_{i=1}^{l}$;
Выход: модель a и размеченная выборка $(x_i, y_i)_{i=1}^{l+k} = X^k \cup X^l$;
Схема: обучить модель a по начальной выборке $(x_i, y_i)_{i=1}^{l}$;
пока остаются неразмеченные объекты $x_{l+1}, \dots, x_{l+k}$:

1. выбрать неразмеченный объект x_i;
2. узнать для него ответ y_i (спросить у «оракула»);
3. дообучить модель a ещё на одном примере (x_i, y_i).

Цель: за фиксированное число запрошенных у оракула ответов k достичь как можно лучшего качества модели.

Примеры ситуаций, где применимо активное обучения:

- Отбор объектов из пула: какой следующий x_i выбрать из множества X^k ?
- Синтез объектов: на каждой i–м шаге строить оптимальный объект x_i;
- Отбор объектов из потока: для каждого приходящего x_i решать, стоит ли узнавать y_i.

Постановка задачи

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

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

где $\mathscr{L}$ — функция потерь, C_i — стоимость информации y_i.

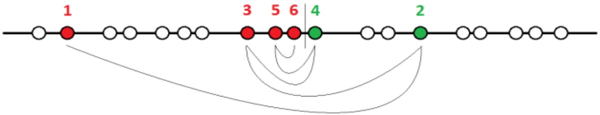
Пример: в планировании эксперимента в естественных науках для получения ответа оракула необходимо провести ту или иную реакцию, реакции имеют разные сложности $\Rightarrow$ величина C_i непосредственно зависит от этой сложности.

Почему активное обучение быстрее пассивного: пример

Рассмотрим задачу с пороговым классификатором:
$x_i \sim U(-1,1), \quad y_i = [x_i > 0], \quad a(x,\theta) = [x > \theta].$

Оценим число шагов для определения θ с точностью $1/k$:

- наивная стратегия: на каждом шаге выбирать $x_i \sim U(X^k)$, трудоемкость — $O(k)$
- бинарный поиск: выбирать x_i, ближайший к середине зазора между классами $\frac{1}{k}\left(\max\limits_{y_j=0}(x_j) + \min\limits_{y_j=1}(x_j)\right)$, трудоемкость $O(\log k)$.



Задача многоклассовой классификации:

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

Идея: выбирать x_i с наибольшей неопределенностью $a(x_i)$.
Для всех $x \in X^k$ обозначим $p_j(x)$, $j = 1, \dots, |Y|$ — ранжированные по убыванию вероятности $P(y|x)$, $y \in Y$.

- Принцип наименьшей достоверности (least confidence):

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

- Принцип наименьшей разности отступов (margin sampling):

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

- Принцип максимума энтропии (maximum entropy):

$$x_i = \arg\min_{u \in X^k} \sum_{j=1}^{|Y|} p_j(u) \ln p_j(u).$$

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

- Принцип максимума энтропии: выбираем x_i, на котором $a_t(x_i)$ дают максимально различные ответы:

$$x_i = \arg\min_{u \in X^k} \sum_{y \in Y} \hat{p}(y|u) \ln \hat{p}(y|u),$$

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

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

$$x_i = \arg\max_{u \in X^k} \sum_{t=1}^{T} \mathsf{KL}(P_t(y|u) || \bar{P}(y|u)),$$

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

Напоминание: $\mathsf{KL}(q||p) = -\int q(x) \log \frac{p(x)}{q(x)} dx$.

Ожидаемое изменение модели (expected model change)

Параметрическая модель многоклассовой классификации:

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

Идея: выбирать x_i, который в методе стохастического градиента привёл бы к наибольшему изменению модели.
Для каждого $u \in X^k$ и $y \in Y$ оценим длину градиентного шага в пространстве параметра θ при дообучении модели на (u, y). Обозначим $\nabla_\theta \mathscr{L}(\theta; u, y)$ — вектор градиента функции потерь.

- Принцип максимума ожидаемой длины градиента:

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

Идея: выбирать x_i, который после дообучения даст выборке наиболее уверенную классификацию неразмеченной выборки X^k.
Для каждого $u \in X^k$ и $y \in Y$ обучим модель, добавив к размеченной обучающей выборке X^l пример (u, y) :

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

Всего таких моделей: $|X^k||Y|$.

- Принцип максимума уверенности на неразмеченных данных:

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

- Принцип минимума энтропии неразмеченных данных:

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

Идея: выбирать x_i, который после дообучения модели даст наименьшую оценку дисперсии предсказания.
Для примера рассмотрим линейную регрессию: $Y = \mathrm{X}B + \mathcal{E}$, $Y \in \mathbb{R}^n$, $B \in \mathbb{R}^m$, $\mathrm{X} \in \mathbb{R}^{n \times m}$, $\mathcal{E} = (\epsilon_1, \ldots, \epsilon_n)^{\mathrm{T}} \in \mathbb{R}^n$, $\forall i\ \mathbb{E}\epsilon_i = 0$, $\mathbb{D}\epsilon_i = \sigma^2$, ϵ_i — независимы.

Знаем: $\mathbb{D}\hat{Y}_j = \mathbb{D}(X_j^{\mathrm{T}}\hat{B} + \mathcal{E}) = \sigma^2 + \frac{\sigma^2}{n} + \frac{\sigma^2}{n}(Z - \bar{X})^{\mathrm{T}}\mathrm{S}_{xx}^{-1}(Z_j - \bar{X})$, где $X_j = (1, Z_j)$, — вектор j–го индивида, S_{xx} — «выборочная ковариационная матрица» центрированных данных, $\bar{X}$ — вектор средних.
Тогда:

$$x_i = \arg\min_{u \in X^k} \mathbb{D}\hat{Y}_j.$$