Санкт-Петербургский государственный университет
Прикладная математика и информатика
Статистическое моделирование

Понизова Вероника, Федяев Игорь

Активное обучение

ML–семинар, конспект

Санкт-Петербург
2019

1. Введение

Активное обучение («query learning» или «optimal experimental design») — раздел машинного обучения, где ключевая гипотеза заключается в том, что если алгоритму позволено самостоятельно выбирать объекты, на которых он будет обучаться, то он будет работать лучше, чем алгоритм, обучающая выборка аналогичного размера для которого выбиралась случайно. Такое свойство важно в следующей ситуации: представим, что размеченная выборка у нас очень мала, но мы имеем возможность узнавать ответы на объектах по запросу к «учителю». Получение правильного ответа от оракула очень сложно, трудоемко или дорого, поэтому хотелось бы понимать, на каком объекте лучше спросить ответ учителя. Задачи такого рода возникают в следующих областях:

- информационный поиск сложно–структурированных объектов, где для обучения поискового алгоритма требуются дорогостоящие асессорские оценки;
- планирование экспериментов в естественных науках, например в комбинаторной химии требуется подбирать условия проведения реакции;
- управление ценами и ассортиментами в торговых сетях: результаты эксперимента могут дорого обойтись, если ставить его «случайным образом».

2. Постановка задачи

Пусть $\mathcal{X}$ — множество объектов, $\mathcal{Y}$ — множество ответов, $y : \mathcal{X} \to \mathcal{Y}$ — неизвестная зависимость. Пусть также имеется начальная размеченная выборка $X^l = (x_i, y_i)_{i=1}^l \subset \mathcal{X} \times \mathcal{Y}$ и некоторый набор неразмеченных объектов $X^K \subset \mathcal{X}$ мощности K, то есть некоторый пул, который мы можем просматривать и выбирать оттуда кандидатов для запроса к оракулу.

Требуется найти функцию $a : \mathcal{X} \to \mathcal{Y}$, приближающую y на всем множестве $\mathcal{X}$ (обучить предсказательную модель), при условии, что получение ответов y_i на неразмеченной части X^K стоит дорого. Цель у нас при этом следующая: за фиксированное число запрошенных у оракула ответов k достичь как можно лучшего качества модели ($k \leqslant K$).

Алгоритм активного обучения примерно таков:

Algorithm 1: Активное обучение

Вход : начальная размеченная выборка $X^l = (x_i, y_i)_{i=1}^l$ и неразмеченный набор объектов X^K;

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

Обучить модель a по начальной выборке $(x_i, y_i)_{i=1}^l$;

while *остаются неразмеченные объекты* $x_{l+1}, \ldots, x_{l+k}$*:* **do**

- выбрать неразмеченный объект $x_i \in X^K$;
- узнать для него ответ y_i (спросить у «оракула»);
- дообучить модель a ещё на одном примере (x_i, y_i).

Качество модели активного обучения определяется следующим образом:

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

где $\mathscr{L}$ — функция потерь, C_i — стоимость информации y_i. Для имеющейся изначально размеченной выборки X^l стоимость информации равна единице: $C_i = 1 \ \forall i \in 1, \ldots, l$.

Таким образом, мы описали общий подход активного обучения. Далее пойдет речь о том, каким образом мы непосредственно выбираем кандидата x_i^* из неразмеченного множества объектов, на котором необходимо спросить ответ у оракула. Все подходы мы будем иллюстрировать на примере байесовского классификатора:

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

где $\mathcal{Y}$ — множество меток классов конечной мощности.

3. Стратегии активного обучения: сэмплирование по неуверенности (uncertainty sampling)

Основная идея этого подхода — на каждом шаге выбирать объект x_i^* такой, что на нем на данный момент достигается наибольшая неопределенность классификатора. Для всех $x \in X^K$ обозначим $p_j(x)$, $j = 1, \ldots, |\mathcal{Y}|$ — ранжированные по убыванию вероятности $P(y|x)$ классов $y \in \mathcal{Y}$ на объекте x.

3.1. Принцип наименьшей достоверности (least confidence)

Для всех $x \in X^K$ рассмотрим наибольшую апостериорную вероятность $p_1(x)$. Понятно, что $p_1(x) \geqslant 1/|\mathcal{Y}|$, и чем меньше эта величина, тем неуверенее классификатор $a(x)$ классифицирует данный объект (апостериорное распределение становится похожим на равномерное). Таким образом, на i–том шаге

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

Например, если $\mathcal{Y} = \{0, 1\}$ и классификатор $a(x)$ является бинарным, то выбираться будут объекты, для которых вероятность быть отнесенными к положительному классу близка к $0,5$ (т.е. объекты на границе).

3.2. Принцип наименьшей разности отступов (margin sampling)

Предыдущий критерий учитывает вероятность только о наиболее вероятной метке, отбрасывая информацию об остальном распределении меток. Чтобы исправить это, вводится следующий принцип выбора x_i^* :

$$x_i^* = \arg\min_{u \in X^K} (p_1(u) - p_2(u)),$$

то есть в рассмотрение включается второй наиболее вероятный класс.

Интуитивно: легко классифицировать объекты, имеющие большое значение отступа, поскольку классификатор не сомневается в выборе между двумя наиболее вероятными классами и легко относит такие объекты или к одному, или к другому. По сравнению с ними, на объектах, где значение отступа мало, достигается наименьшая определенность. На таких x_i и предлагается узнавать ответы: это поможет классификатору более эффективно отличать объекты с маленьких отступом (т.е. объекты на границе). Однако этот метод не будет панацеей в случае, если мощность множества $\mathcal{Y}$ велика.

3.3. Принцип максимума энтропии (maximum entropy)

Третий принцип использует энтропию как меру неопределенности:

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

Здесь идет минимизация, поскольку в выражении для самой энтропии еще есть минус перед суммой. Если $\mathcal{Y} = \{0, 1\}$ и классификатор $a(x)$ является бинарным, то выбираться будут объекты опять таки с апостериорными вероятностями близкими к 0.5, поскольку на таких объектах энтропия максимальна.

4. Стратегии активного обучения: сэмплирование по несогласию в комитете (query by committee)

Этот подход учитывает мнение нескольких моделей $a_t(x_i) = \arg\max\limits_{y \in Y} P_t(y|x)$, $t = 1, \ldots, T$, обученных на изначально размеченных данных X^l. А именно — предлагает выбирать те объекты x_i^*, на которых достигается наибольшая несогласованность комитета моделей.

4.1. Принцип максимума энтропии (maximum entropy)

Рассмотрим долю моделей, которые классифицируют кандидата $u \in X^K$ как объект из класса y: $\hat{p}(y|u) = \frac{1}{T}\sum_{t=1}^{T}[a_t(u) = y]$. Рассматривая все $y \in \mathcal{Y}$ для фиксированного $u \in X^K$ получим некоторое эмпирическое распределение. Выбираться будет тот объект, для которого энтропия такого распределения максимальна:

$$x_i^* = \arg\min_{u \in X^k} \sum_{y \in Y} \hat{p}(y|u) \ln \hat{p}(y|u). \tag{1}$$

4.2. Принцип максимума средней KL–дивергенции (average Kullback-Leibler divergence)

Напомним, что KL–дивергенция вероятностного распределения $q(x)$ относительно $p(x)$ выглядит следующим образом:

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

Основная мысль здесь заключается в том, что стоит рассматривать не ответы комитета, как в принципе максимума энтропии, а вероятности, а именно вероятности, так как чтобы получить ответ, классификатор округляет соответствующую вероятность в большую или меньшую сторону. Наряду с энтропией KL–дивергенция также может служить мерой несогласия распределений. Согласно этому принципу, будет выбираться объект,

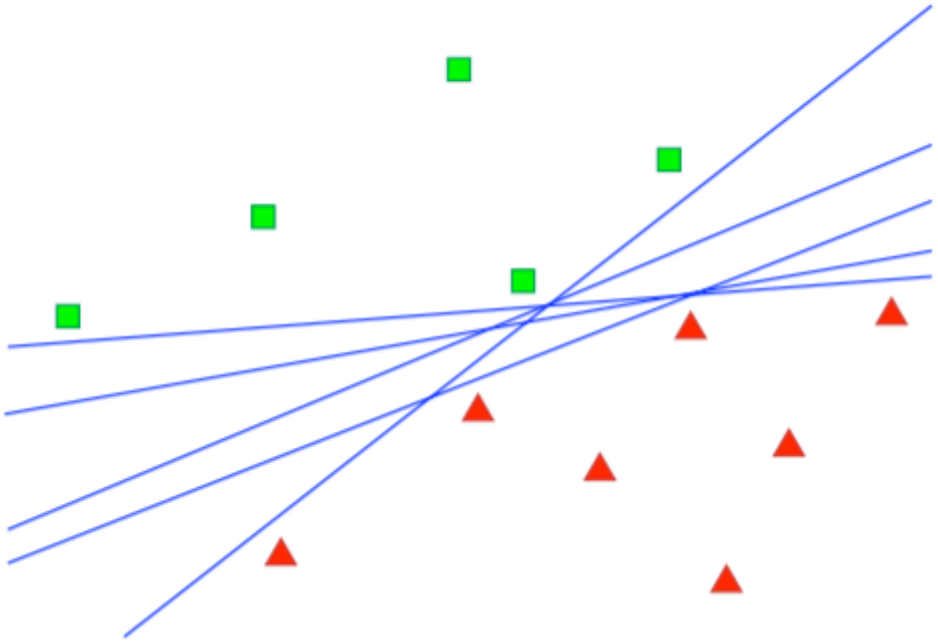
выбираем x_i^*, на котором показания моделей $P_t(y|x_i)$ максимально различны:

$$x_i^* = \arg\max_{u \in X^k} \sum_{t=1}^{T} \mathrm{KL}(P_t(y|u)||\bar{P}(y|u)) = \arg\max_{u \in X^k} \sum_{t=1}^{T} \sum_{i=1}^{|\mathcal{Y}|} P_t(y_i|u) \log \frac{P_t(y_i|u)}{\bar{P}(y_i|u)}, \qquad (2)$$

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

4.3. Иллюстрация

Снова рассмотрим случай бинарного линейного классификатора: Обратим на три



наиболее близких к «точке пересечения» решающих прямых объекта: два красных треугольника и один зеленый квадрат. Понятно, что каждая из моделей в комитете для обоих классов будет давать вероятности, примерно равные 0.5, тогда логарифм под двойной суммой в (2) будет примерно равен нулю (или энтропия в (1) минимальна), и такие объекты не будут максимизировать рассматриваемое выражение (энтропию, соответственно). С объектами, находящимися близко к верху и низу изображения, такая же ситуация: логарифм будет близок к нулю, поскольку все модели выдают большую и примерно одинаковую вероятность принадлежать зеленому или красному классу. Объекты, близкие к правому и левому краям изображения, как раз таки являются кандидатами на выбор: какая–то модель классифицирует их более уверенно, а какие–то менее уверенно, из–за чего логарифм уже не будет около нуля. Таким образом, в целом этот подход «прощупывает» далекие области.

5. Ожидаемое изменение модели (expected model change)

Как известно, большое число методов ML основано на методе стохастического градиента: на каждом объекте из обучающей выборки производится шаг по градиенту. Рассмотрим параметрический байесовский классификатор:

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

Основная идея заключается в следующем: 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 $\bigtriangledown_\theta \mathscr{L}(\theta; u, y)$ — вектор градиента функции потерь. Вводится принцип максимума ожидаемой длины градиента:

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

где $||\cdot||$ — евклидова норма результирующего вектора градиента. Интуитивно, данный способ также не обращает внимания на объекты, находящиеся близко к границам — поскольку теоретически они не слишком сильно изменят модель, то есть если и сдвинут решающую гиперплоскость, то ненамного. Соответственно, он выбирает дальние объекты, получение ответа на которых может сильно изменить форму гиперплоскости. Стоит отметить, что здесь существенна стандартизация данных, иначе некоторые объекты в пуле могут быть переоценены в своей значимости из–за необычайно больших значений какого–то из признаков. По этой же причине этот метод чувствителен к аутлаерам.