Теория статистического обучения

Н. К. Животовский
nikita.zhivotovskiy@phystech.edu

2 марта 2016 г.

Материал находится в стадии разработки, может содержать ошибки и неточности. Автор будет благодарен за любые замечания и предложения, направленные по указанному адресу

Цель данного курса лекций заключается в доступном изложении основных результатов теории статистического обучения (Statistical learning theory – 'SLT'). Систематическое исследование теоретических основ вопросов машинного обучения привело к созданию теории статистического обучения и началось около 30 лет назад с работ Вапника и Червоненкиса. Одним из основных преимуществ разработанной ими теории была независимость основных результатов от того, по какому закону распределены данные. Таким образом, был осуществлен переход от подхода, ориентированного на модель данных (статистический подход), к подходу, заключающемуся в анализе в первую очередь методов обучения. Вторым важным шагом было получение необходимых и достаточных условий для равномерной по классу гипотез сходимости частот к вероятностям. Теперь процесс обучения можно контролировать вне зависимости от распределения данных и даже сложной процедуры выбора алгоритма из семейства. Общность подхода, конечно, имела очевидные недостатки, многие из которых были ликвидированы в последнее десятилетие. Катализатором исследований были два вероятностных раздела – теория эмпирических процессов и неравенства концентрации меры.

1 Вероятностная постановка задач распознавания

Предположим, что существует множество объектов $\mathcal{X}$ (объекты принято отождествлять с их признаковыми описаниями) и множество ответов $\mathcal{Y}$. Последнее, например, в случае задачи классификации на два класса может состоять всего из двух элементов (классы 1 и -1) или в случае задачи регрессии совпадать со множеством действительных чисел. Далее предполагается, что нам дана *обучающая* выборка из n пар (X, Y) из $\mathcal{X} \times \mathcal{Y}$.

Говоря неформально, цель статистического обучения заключается в том чтобы на основании имеющейся обучающей выборки построить некоторое правило, которое бы смогло предсказать ответ Y на основании нового объекта X. Тем не менее какое-то предположение о природе данных должно существовать.

В данной теории предполагается:

- На $\mathcal{X} \times \mathcal{Y}$ задана неизвестная вероятностная мера P.

- Все пары (X, Y) из обучающей выборки получены независимо согласно этой мере (вероятностному распределению).
- Любая новая пара (X, Y) получается согласно тому же самому распределению и независимо от остальных.

Предположим, что на основании обучающей выборки нам удалось построить функцию $\hat{f} : \mathcal{X} \to \mathcal{Y}$. В этом случае, говорят, что был использован некоторый *алгоритм* или *метод* обучения, а процесс его применения мы в дальнейшем будем называть *обучением*. Заметим, что наличие взаимосвязи между X и Y как-то характеризуется самой вероятностной мерой P. Для того чтобы делать какие-то предсказания логично предположить, что P не является произведением мер по X и Y, то есть объекты и, например, их классы вовсе не независимые случайные величины. Одновременно слишком сильное предположение заключается и в существовании строгой функциональной зависимости между X и Y. Поэтому P такова, что предполагается существование достаточно хорошей (в некотором смысле) связи между объектами и ответами. Для того чтобы формализовать эту идею нужно ввести *функцию ошибок*. Функция ошибок (функция потерь) — это некоторая неотрицательная функция $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}_+$, которая характеризует потери при относении объекта X к ответу $\hat{f}(X)$ в сравнении с его реальным ответом Y. Зафиксировав $\hat{f}$, удобно переопределить функцию ℓ на парах (X, Y) следующим образом:

$$\ell(\hat{f}, X, Y) := \ell(\hat{f}(X), Y)$$

Не вводя требований, которые обычно предъявляются к функциям потерь, перейдем сразу к типичным примерам:

- В случае задачи классификации бинарные потери $\ell(\hat{f}, X, Y) = \mathbf{I}\{\hat{f}(X) \neq Y\}$.
- В задачах регрессии $\ell(\hat{f}, X, Y) = (\hat{f}(X) - Y)^2$.
- или $\ell(\hat{f}, X, Y) = |\hat{f}(X) - Y|^2$.
- hinge loss $\ell(\hat{f}, X, Y) = \max\{0, 1 - Y\hat{f}(X)\}$.

Разумной характеристикой решающего правила была бы его ожидаемая ошибка по отношению к обучающей выборке, на основании которого оно построено:

$$L(\hat{f}) = \mathbb{E}\left[\ell(\hat{f}, X, Y) \mid (X^n, Y^n)\right]$$

Важно понимать, что математическое ожидание берется по новой паре (X, Y), в то время как решающее правило $\hat{f}$ само строится по случайной обучающей выборке (X^n, Y^n). Если считать $\hat{f}$ не случайным объектом, то $L(\hat{f}) = \mathbb{E}\left[\ell(\hat{f}, X, Y)\right]$ называют *риском* правила $\hat{f}$. Для того чтобы избавиться от зависимости от случайной реализации определим уже неслучайную величину, называемую в дальнейшем *средним риском*:

$$\mathbb{E}L(\hat{f}) := \mathbb{E}\left[\mathbb{E}\left[\ell(\hat{f}, X, Y) \mid (X^n, Y^n)\right]\right]$$

Данная терминология несколько отличается от принятой в математической статистике, однако, для удобства мы будем использовать введенные определения. Средний риск зависит только от меры P и способа выбора $\hat{f}$ и дает разумный критерий выбора способа построения $\hat{f}$: выбирается та оценка, которая доставляет минимальный средний риск. Напомним, что в общем случае $\hat{f}$ – это случайная конструкция, которая строится на основании обучающей выборки.

Если бы P была известна, то задача поиска оптимального $\hat{f}$ была бы лишь задачей оптимизации.

Пример 1.1. Пусть мы имеем дело с задачей классификации $\mathcal{Y} = \{1, 0\}$ с бинарной функцией потерь. В этом случае риск равен $P(\hat{f}(X) \neq Y)$. Среди всевозможных выборов $\hat{f}$ его минимизирует так называемое *байесовское решающее правило* $g(x) = \mathbf{I}\{\eta(x) \geq \frac{1}{2}\}$, где $\eta(x) = \mathbb{E}(Y|X = x)$. Отметим, что байесовское решающее правило зависит не от обучающей выборки, а от неизвестной меры P, поэтому одним из способов приближенного построения байесовского решающего правила являются так называемые plug-in правила, основанные на построении по наблюдаемой выборке эмпирического аналога $g(x)$.

В теории статистического обучения не принято задавать модель данных в явном виде или предполагать зависимость между X и Y. Наше априорное знание о задаче должно быть представлено в основном не ограничением на меру P, а априорно заданным семейством отображений $\mathcal{F}$, каждое из которых отображает X в Y. Алгоритмы в результате обучения выбирают решающее правило, принадлежащее $\mathcal{F}$. Такие алгоритмы обычно называют *proper learning* алгоритмами, в отличие от, например, построений композиций алгоритмов, которые мы будем обсуждать в дальнейшем. В литературе, однако, часто и семейство решающих правил $\mathcal{F}$ называется моделью, а выбор оптимального для задачи $\mathcal{F}$ — называется задачей выбора модели. В качестве семейства решающих правил могут выступать, например, гиперплоскости (афинные подпространства) в случае линейных классификаторов, семейства функций с определенными свойствами гладкости и так далее. Рассмотрим некоторые примеры:

Пример 1.2.

1. Бесшумный случай (в литературе часто называется *function learning*): существует некоторая функция $T : \mathcal{X} \to \mathcal{Y}$, такая, что $Y = T(X)$. При этом не делается предположений о том, что $T \in \mathcal{F}$. Очевидно, что, например, для квадратичной функции ошибки в этом случае функция T является еще и байесовским решающим правилом.

2. *Непараметрическая регрессия*: Зависимость $Y = f^*(X) + \varepsilon$, где ε — центрированная случайная величина с конечной дисперсией, не зависящая от X. Легко понять, что в этом случае обозначение f^* не случайно. Действительно, данная функция является байесковским решающим правилом относительного квадратичной функции потерь. Данный тип зависимостей обычно изучается в математической статистике, где практически всегда предполагается, что $f^* \in \mathcal{F}$.

3. В более общей задаче статистического обучения никаких функциональных связей между X и Y не преполагается, а рассматриваются отдельно лишь хорошо специфицированный случай, когда байесовское решающее правило $f^* \in \mathcal{F}$, и

агностический случай (*agnostic case*), когда не делается предположений о принадлежности байесовского решающего правила семейству $\mathcal{F}$. В обоих случаях анализ проводится на основании свойств семейства $\mathcal{F}$.

В дальнейшем мы проанализируем практически все рассматриваемые задачи. В частности, приведем в некотором смысле оптимальный алгоритм, который для широкого класса семейств $\mathcal{F}$ будет с точностью до константы иметь минимальный возможный риск.

Approximation–Estimation tradeoff

Ясно, что байесовское решающее правило, то есть, правило, минимизирующее риск, является в некотором смысле эталоном для любого метода обучения. Таким образом, возникает разумный вопрос о построении таких алгоритмов, обучаясь с помощью которых мы получили бы риск, сколь угодно близкий к байесовскому. Обозначая все измеримые отображения из X в Y как $\mathcal{Y}^{\mathcal{X}}$, рассмотрим разницу риска оценки $\hat{f}$ и байесовского решающего правила, то есть:

$$\mathbb{E}L(\hat{f}) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} L(f),$$

где $\inf\limits_{f \in \mathcal{Y}^{\mathcal{X}}} L(f) = L(f^*)$. Однако, в общих задачах статистического обучения $f^* \notin \mathcal{F}$, поэтому перепишем предыдущую разность в следующем виде:

$$\left(\mathbb{E}L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f)\right) + \left(\inf_{f \in \mathcal{F}} L(f) - \inf_{f \in \mathcal{Y}^{\mathcal{X}}} L(f)\right).$$

Левое слагаемое называют *ошибкой оценивания* (*estimation error*), а правое называется *ошибкой аппроксимации* (*approximation error*). Очевидно, что чем больше $\mathcal{F}$, тем меньше ошибка аппроксимации, но одновременно больше ошибка оценивания. Действительно, так как алгоритм выбирает правило из $\mathcal{F}$, то в лучшем случае его приближает именно $\inf\limits_{f \in \mathcal{F}} L(f)$. Действительно, если $\mathcal{F}$ состоит всего из одной функции, то ошибка оценивания равна нулю. Заметим также, что в реализуемом случае ошибка аппроксимации равна нулю. Компонента

$$L(\hat{f}) - \inf_{f \in \mathcal{F}} L(f)$$

называется *избыточным риском* и, таким образом, ошибка оценивания в наших обозначениях является математическим ожиданием избыточного риска.

Обучаемость

Введем классическое понятие *Probably Approximately Correct* – обучаемости. Ограничимся задачей классификации $\mathcal{Y} = \{0, 1\}$, бинарной функцией потерь и примем бесшумную модель, то есть для некоторой функции $T : \mathcal{X} \to \{0, 1\}$ имеет место $Y = T(X)$. Также предполагается *реализуемый* случай, то есть, что $T \in \mathcal{F}$.

Опр. 1.1 (PAC–learnability). *Семейство $\mathcal{F}$ называется PAC–обучаемым, если существует функция $n_{\mathcal{F}} : (0, 1)^2 \to \mathbb{N}$ и некоторый обучающий алгоритм, такие что*

для всех $\varepsilon, \delta \in (0,1)$ при любом вероятностном распределении на $\mathcal{X}$ и любой целевой функции T, если алгоритм на выборке из хотя бы $n \geq n_{\mathcal{F}}(\varepsilon, \delta)$ объектов выдает классификатор $\hat{f}$, такой что $\hat{f} \in \mathcal{F}$, то с вероятностью не меньшей чем $1-\delta$ (по отношению к обучающей выборке) имеет место неравенство

$$L(\hat{f}) \leq \varepsilon.$$

Обучаемость значит, что для достаточно большой выборки алгоритм выдает с большой вероятностью решение, обладающее маленькой вероятностью ошибки. Среди функций $n_{\mathcal{F}}$ ту, что принимает наименьшие значения, принято называть *выборочной сложностью* (sample complexity). Она показывает сколько нужно объектов, чтобы обучиться с заданной точностью. Обобщим введенное понятие на агностический случай:

Опр. 1.2 (Agnostic PAC–learnability). *Семейство $\mathcal{F}$ называется агностически PAC–обучаемым, если существует функция $n_{\mathcal{F}} : (0,1)^2 \to \mathbb{N}$ и некоторый обучающий алгоритм, такие что для всех $\varepsilon, \delta \in (0,1)$ при любом вероятностном распределении на $\mathcal{X} \times \mathcal{Y}$, если алгоритм на выборке из хотя бы $n \geq n_{\mathcal{F}}(\varepsilon, \delta)$ объектов выдает классификатор $\hat{f}$, такой что $\hat{f} \in F$, то с вероятностью не меньшей чем $1-\delta$ (по отношению к обучающей выборке) имеет место неравенство*

$$L(\hat{f}) \leq \inf_{f \in \mathcal{F}} L(f) + \varepsilon.$$

Заметим, что в агностическом смысле сравнение идет не с абсолютной ошибкой, то есть с $\inf\limits_{f \in \mathcal{Y}^{\mathcal{X}}} L(f)$, а с лучшим классификатором в классе. В терминах аппроксимации и оценивания нас интересует только часть, отвечающая за ошибку оценивания.

Минимаксные порядки

Предположим, что нам задано семейство решающих правил $\mathcal{F}$. Разумным вопросом является: на сколько хорошо может работать лучший обучающий алгоритм в случае самого плохого распределения? Пусть функция $\tilde{f}$ построена по обучающей выборке из n объектов с помощью некоторого обучающего алгоритма. Символ $\inf\limits_{\tilde{f}}$ будет означать минимизацию по всем обучающим алгоритмам, при этом $\tilde{f}$ будет обозначать функцию, полученную с помощью обучения этим алгоритмом на выборке из n элементов. Введем следующие минимаксные значения:

- В случае PAC–обучения $\mathcal{V}^{\text{PAC}}(\mathcal{F}, n) = \inf\limits_{\tilde{f}} \sup\limits_{\mathsf{P}_{\mathcal{X}}, T \in \mathcal{F}} \mathbb{E}\left(\mathsf{P}\left(\tilde{f}(X) \neq T(X)\right)\right)$.
- Непараметрическая регрессия $\mathcal{V}^{\text{NP}}(\mathcal{F}, n) = \inf\limits_{\tilde{f}} \sup\limits_{\mathsf{P}_{\mathcal{X}}, f^* \in \mathcal{F}} \mathbb{E}\left(\mathbb{E}\left(\tilde{f}(X) - f^*(X)\right)^2\right)$.
- В общем случае $\mathcal{V}^{\text{i.i.d.}}(\mathcal{F}, n) = \inf\limits_{\tilde{f}} \sup\limits_{\mathsf{P}} \mathbb{E}\left(L(\tilde{f}) - \inf\limits_{f \in \mathcal{F}} L(f)\right)$.

Упр. 1.1. Доказать, что для любого семейства $\mathcal{F}$, в котором функции принимают только значения $\{0,1\}$, выполнено следующее неравенство:

$\mathcal{V}^{\text{PAC}}(\mathcal{F}, n) \leq \mathcal{V}^{\text{NP}}(\mathcal{F}, n)$.

Упр. 1.2. Доказать, что для любого семейства $\mathcal{F}$ в случае квадратичных потерь выполнено:

$$\mathcal{V}^{\text{NP}}(\mathcal{F}, n) \leq \mathcal{V}^{\text{i.i.d.}}(\mathcal{F}, n).$$

Часто в литературе под *обучаемостью* имеется в виду сходимость к нулю соответствующих минимаксных значений.

Упр. 1.3. Как связаны PAC–обучаемость и сходимость к нулю соответствующих минимаксных порядков?

No Free Lunch Theorem

Теорема 1.1 (No Free Lunch Theorem). *Пусть $\mathcal{X}$ является конечным множеством, а $n \leq \frac{|\mathcal{X}|}{2}$. Тогда в задаче классификации для любого алгоритма найдется такое распределение на $\mathcal{X} \times \{0, 1\}$, что:*

- *Существует классификатор с нулевым риском.*
- *С вероятность не меньшей чем $\frac{1}{7}$ относительно обучающей выборки длины n риск правила, выбираемого алгоритмом больше чем $\frac{1}{8}$.*

Доказательство.

Доказательство в следующей лекции. ■

Очевидным следствием данной теоремы является следующая теорема

Теорема 1.2 (Необучаемость семейства всех функций). *Пусть $\mathcal{X}$ — бесконечное множество, а $\mathcal{F} = \mathcal{Y}^{\mathcal{X}}$. Тогда $\mathcal{F}$ не является PAC–обучаемым.*

Список литературы

[1] *Boucheron S., Bousquet O., Lugosi G.* Introduction to Statistical Learning Theory // 2004. — Pp. 169-207.

[2] *Rakhlin A.* Statistical Learning Theory and Sequential Prediction // Lecture notes, 2014, http://www-stat.wharton.upenn.edu/ rakhlin/

[3] *Shalev-Shwartz S., Ben-David S.* Understanding Machine Learning: From Theory to Algorithms // Cambridge University Press, 2014

[4] *Vapnik V.* Statistical Learning Theory. — John Wiley and Sons, New York, 1998.

[5] *L. G. Valiant* A theory of the Learnable. — Communications of the ACM, 27, 1984.