Теория статистического обучения

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

17 февраля 2014 г.

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

1 Введение

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 семейтсва. Общность подхода, конечно, имела очевидные недостатки, многие из которых были ликвидированы в последнее десятилетие. Катализатором исследлований были два вероятностных раздела – теория эмпирических процессов и неравенства концентрации меры. Перейдем теперь к постановке задачи.

§1.1 Постановка задачи

Начнем с так называемого обучения с учителем (supervised learning). 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 считается, что на $\mathcal{X} \times \mathcal{Y}$ существует некоторая неизвестная вероятностная мера P. И все пары (X, Y) из обучающей выборки получены независимо согласно этой мере (вероятностному распределению). Второе предположение заключается в том, что любая новая пара (X, Y) получается согласно тому же самому распределению.

Предположим, что на основании обучающей выборки нам удалось построить функцию $\hat{f} : \mathcal{X} \to \mathcal{Y}$. 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 независмые 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 ошибки*. Это некоторая функция $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, которая характеризует потери при относении объекта X к ответу $\hat{f}$ в сравнении с его реальным ответом Y. Удобно определить функцию ℓ на парах (X, Y) следующим образом:

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

Типичные примеры:

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

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

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

Важно понимать, что математическое ожидание берется по новому объекту (X, Y), в то время как само решающее правило $\hat{f}$ само строится по случайной выборке (X^n, Y^n). Для того чтобы избавиться от зависимости от случайной реализации определим уже неслучайную величину, называемую средним риском:

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

Средний риск зависит теперь только от меры P и способа выбора $\hat{f}$. Среди всех решающих правил ищется то, которое доставляет минимальный средний риск. Это соответствует так назвываемому Probably approximately correct learning [13], заключающемуся в выбоке решающего правила, которая хорошо приближает ненаблюдаемые данные. Опять же напомним, что в общем случае $\hat{f}$ – это случайная функция, 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 $\mathcal{Y} = \{1, -1\}$ с бинарной функцией потерь. В этом случае ожидаемый риск равен $P(\hat{f}(X) \neq Y)$. Среди всевозможных выборов $\hat{f}$ его минимизирует так называемое баесовское решающее правило $g(x) = \operatorname{sgn}(\mathbf{E}(Y|X = x))$. Отметим, что баесовское решающее правило зависит не от обучающей выборки, а от неизвестной меры P, поэтому одним из способов приближенного построения баесовского решающего правил являются так называемые plug-in rules, основанные на построении по наблюдаемой выборке эмпирического аналога $g(x)$.

Упр. 1.1. Для случая классификации докажите оптимальность баесовского решающего правила.

Одной из самых подробно изученных моделей в статистической теории является линейная. Попробуем кратко продемонстрировать разницу между постановками задач статистической теории и теории статистического обучения.

Пример 1.2 (Линейная регрессия (математическая статистика)). В качестве функции потерь принято считать квадратичное отклонение $\ell(f, x, y) = ||f(x) - y||^2$. Считая x — d мерным вектором ($n \times d$ матрицей), а y числом (n векотором), в случае линеной модели $f(x)$ можно рассматривать как значение матричного умножения некоторого вектора f на вектор (матрицу) x.

Пусть мы пронаблюдали векторы $x_1, \ldots, x_n$ (которые, в простой статистической постановке считаются неслучайными). Наблюдению x_t соответует ответ Y_t, причем для некоторой функции g^* имеет место равенство

$$Y_t = g^* x_t + \varepsilon_t,$$

где ε_t независимые в совокупности случайные величины с нулевым средним и дисперсией σ^2. Стандартное матричное представление записывается

$$Y = Xg^* + \varepsilon,$$

где строки матрицы X являются векторами x_t. Определим $\hat{\Sigma} = \frac{1}{n}\sum\limits_{t=1}^{n} x_t x_t^T$. Задача уменьшения среднего риска сводится теперь к поиску $\hat{g}$, минимизирующему

$$\mathbf{E}||\hat{g} - g^*||^2_{\hat{\Sigma}} = \mathbf{E}\left(\frac{1}{n}\sum_{t=1}^{n}(\hat{g}(x_t) - g^*(x_t))^2\right).$$

Считая, что $d \leq n$ и матрицы $\hat{\Sigma}$ обратима, построим по наблюдениям оценку наименьших квадратов

$$\hat{g} = \arg\min_{g \in \mathbf{R}^d} \frac{1}{n}\sum_{t=1}^{n}\left(Y_t - gx_t\right)^2.$$

Хорошо известная явная формула для оценки наименьших квадратов выглядит

$$\hat{g} = \hat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{t=1}^{n} Y_t x_t\right) = \frac{1}{n}\hat{\Sigma}^{-1}X^T Y.$$

Из $Y_t = g^* x_t + \varepsilon_t$ легко получить, что

$$g^* = \hat{\Sigma}^{-1}\left(\frac{1}{n}\sum_{t=1}^{n}\left(Y_t - \varepsilon_t\right)x_t\right)$$

Подставляя полученные выражения в $\mathbf{E}||\hat{g} - g^*||^2_{\hat{\Sigma}}$ получаем с учётом того, что матрица $\frac{1}{n}X\hat{\Sigma}^{-1}X^T$ является проектором на $Im(X)$

$$\mathbf{E}||\hat{g} - g^*||^2_{\hat{\Sigma}} = \frac{1}{n}\mathbf{E}\left(\varepsilon^T\left(\frac{1}{n}X\hat{\sigma}^{-1}X^T\right)\varepsilon\right) = \frac{\sigma^2}{n}\operatorname{tr}\left(\frac{1}{n}X\hat{\Sigma}^{-1}X^T\right) \leq \frac{\sigma^2 d}{n}.$$

Такую же скорость сходимости можно получить и в случае, если матрица X случайна (random design), но при некоторых ограничениях на распределение ее элементов [6].

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 не ограничением на меру P, а априорно заданным семейством отображений $\mathcal{F}$, каждое из которых отображает X в Y. 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 в случае линейных классификаторов.

Теперь для некоторого построенного решающего правила $\hat{f}$ разумной мерой качества будет так называемый ожидаемый избыточный риск (expected excess risk) [9]

$$\mathbf{E}\ell(\hat{f}, X, Y) - \inf_{f \in F}\mathbf{E}\ell(f, X, Y)$$

Он показывает насколько построенное правило хуже чем лучшее в среднем правило в семействе $\mathcal{F}$. Заметим важную особенность: в нашем определении усреднение берется также и по обучающей выборке, то есть, мы работаем с детерминированной величиной и можем давать верхние оценки. Некоторые авторы [7] предпочитают работать с более общей величиной избыточного риска (excess risk), в которой усреднение в $\mathbf{E}\ell(\hat{f}, X, Y)$ берется только по паре X, Y и, таким образом, избыточный риск является случайной велчиной и все утверждения про него носят вероятностный характер.

Пример 1.3 (Линейная регрессия(статистическое обучение)). Пусть $\mathcal{F}$ представляет собой векторы d-мерного шара единичного радиуса. Функция потерь также остаётся квадратичной. Теперь пары из обучающей выборки (x_t, Y_t) получены независимо согласно неизвестной мере P. В отличие от статистической постановки мы не предполагаем никаких соотношений между X и Y. Более того нам не нужно даже чтобы баесовское решающее правило $(f(x) = \mathbf{E}(Y|X = x))$ принадлежало семейству $\mathcal{F}$.

Мы сможем показать, что в случае, если $\hat{f} = \arg\min\limits_{f \in \mathcal{F}} \frac{1}{n}\sum\limits_{t=1}^{n}(Y_t - fx_t)^2$, то для некоторой абсолютной константы C

$$\mathbf{E}\ell(\hat{f}, X, Y) - \inf_{f \in F}\mathbf{E}\ell(f, X, Y) \leq C\left(\frac{d}{n}\right).$$

Таким образом, без ограничений на распределние, пользуясь лишь структурой $\mathcal{F}$, можно получить приближение оценки наименьших квадратов к лучшему линейному приблежению с порядком $O(\frac{d}{n})$.

2 Оценки обобщающей способности

§2.1 Минимизация эмпирического риска

Для того чтобы получить хотя бы один содержательный результат нам нужно получить простое неравенство концентрации меры [3].

Теорема 2.1 (неравенство Хеффдинга). *Пусть $Z_1, \dots, Z_n$ - независимые случайные величины, такие что почти наверное $Z_i \in [a_i, b_i]$. Обозначим $S_n = \sum\limits_{i=1}^{n} Z_i$, тогда для любого $t > 0$ имеют место неравенства*

$$\mathsf{P}\left\{S_n - \mathbf{E}S_n \geq t\right\} \leq \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right)$$

и

$$\mathsf{P}\left\{S_n - \mathbf{E}S_n \leq -t\right\} \leq \exp\left(\frac{-2t^2}{\sum\limits_{i=1}^{n}(b_i - a_i)^2}\right)$$

Доказательство.
На лекции. ∎

В данном разделе для удобства обозначим $L(f) = \mathbf{E}\ell(f, X, Y)$. Для классификатора f введем понятие эмпирического риска $L_n(f) = \frac{1}{n}\sum\limits_{i=1}^{n} \ell(f, X_i, Y_i)$. Разумно выбирать такой классификатор, который минимизирует эмпирический риск, то есть $\hat{f} = \arg\min_{f \in \mathcal{F}} L_n(f)$.. Методы обучения, основанные на этой идее будем называть методами *минимизации эмпирического риска.*

Отметим, что $L(f) - L_n(f)$ является случайной величной даже если f не зависит от обучающей выборки. Оценкой обобщающей способности называется всякая верхняя оценка на вероятность уклонений среднего риска от эмпирического, то есть для $t > 0$ оценка на $\mathsf{P}\{L(f) - L_n(f) \geq t\}$. Далее будем для простоты считать, что функция ошибок ℓ равномерно ограничена единицей. Отсюда можно вывести простое соотношение

Теорема 2.2. *Пусть $\mathcal{F} = \{f\}$, то есть $\hat{f} = f$. Тогда для любого $\delta > 0$ с вероятностью не меньшей $1 - \delta$ выполнено*

$$L(\hat{f}) \leq L_n(\hat{f}) + \sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Упр. 2.1. Доказать теорему.

Хотелось бы обобщить приведенный результат на случай, когда $\mathcal{F}$ содержит более одного решающего правила. Проблема заключается в том, что в оценках сложно

учесть специфику, связанную с тем, что мы выбираем именно лучшее на обучающей выборке решающее правило. Поэтому начнем с классического подхода, который заключается в получении равномерных оценок, то есть исследовании

$$\sup_{f\in\mathcal{F}}|L(f)-L_n(f)|,$$

или

$$\sup_{f\in\mathcal{F}}(L(f)-L_n(f)).$$

Теорема 2.3 (Конечный класс функций). *Пусть* $\mathcal{F}=\{f_1,\dots,f_n\}$, *то есть* $\hat{f}=f$. *Тогда для любого* $\delta>0$ *одновременно с вероятностью не меньше* $1-\delta$ *выполнено*

$$\forall f\in F\mathcal{F}L(\hat{f})\le L_n(\hat{f})+\sqrt{\frac{\log\frac{1}{\delta}}{2n}}.$$

Доказательство.

С помощью неравенств Хеффдинга и Буля имеем

$$\mathsf{P}\{\exists f\in\mathcal{F}: L(f)-L_n(f)>\varepsilon\}\le\sum_{i=1}^{n}\mathsf{P}\{L(f)-L_n(f)>\varepsilon\}\le N\exp(-2n\epsilon^2).$$

Отсюда

$$\mathsf{P}\{\forall f\in\mathcal{F}: L(f)-L_n(f)\le\varepsilon\}\ge 1-N\exp(-2n\epsilon^2).$$

Обращая оценку(что и нужно было научиться делать в упражнении), получаем утверждение теоремы. ■

Список литературы

[1] *Boucheron S., Bousquet O., Lugosi G.* Theory of classification: A survey of some recent advances // ESAIM: Probability and Statistics. — 2005. — No. 9. — Pp. 323–375.

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

[3] *Boucheron S., Lugosi G., Massart P.* Concentration Inequalities: A Nonasymptotic Theory of Independence // 2013. —

[4] *Devroye L., Lugosi G.* Combinatorial Methods in Density Estimation // Springer Series in Statistics. Springer-Verlag, 2001.

[5] *Haussler D.* Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension // Journal of Combinatorial Theory. — 1995. — Pp. 217–232.

[6] *Hsu D., Kakade S. M, Zhang T.* An Analysis of Random Design Linear Regression // . — 2011 http://arxiv.org/pdf/1106.2363v1.pdf

[7] *Koltchinskii V.* Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems // Ecole d'Etŕe de Probabilitŕes de Saint-Flour XXXVIII-2008. Lecture Notes in Mathematics. Springer-Verlag, 2011.

[8] *Ledoux M.* The Concentration of Measure Phenomenon // American Mathematical Society, 2005.

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

[10] *Talagrand M.* The Generic Chaining. Upper and Lower Bounds of Stochastic Processes // Springer Monographs in Mathematics, 2005.

[11] *Talagrand M.* New concentration inequalities in product spaces // Inventiones mathematicae 1996. — Pp. 505-563.

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

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

[14] *Vorontsov K. V.* Exact combinatorial bounds on the probability of overfitting for empirical risk minimization // *Pattern Recognition and Image Analysis.* — 2010. — Vol. 20, no. 3. — Pp. 269–285.