ЛЕКЦИЯ 12

Кроме классической статистики

Что ещё есть интересного за рамками классической статистики? Как можно ценой больших вычислений отказаться от ограничивающих предположений при оценке параметров, проверке гипотез и построении доверительных интервалов? Об этом в обзоре последней лекции.

12.1 Экзотичный доверительный интервал

Наверняка мы всё уже сталкивались с ситуацией, когда мы вычислили по данным какую-то метрику, оценили параметры, но сделали это настолько сложно, что классическая статистика не пробросилась через весь пайплайн. Кроме этого, данных конечное число, больше нам никто не даст, хотя у нас в очень крайнем случае и могут быть ещё силы разметить десяток картинок для классификации. Что делать, если мы хотим узнать больше про полученные оценки: получить доверительный интервал (чтобы понять, насколько устойчива оценка), проверить гипотезу (если мы работаем с нетривиальными метриками в АБ-тестировании), оценить смещение и дисперсию оценки?

Пример 12.1. *Вам, команде аналитиков, задают вопрос: как вы можете быть уверены, что ваш метод случайного леса устойчив к данным? Возможно, если выкинуть или добавить 2-3 наблюдения результат сильно изменится? Есть ли доверительный интервал для ваших оценок? (есть целая область Uncertainty Quantification, которая интересуется подобными задачами)*

Пример 12.2. *С другой стороны, вы получили AUC, точность и полноту. Это у вас средний показатель хороший, вам с данными повезло или всё же если имеющиеся данные поменять, результат останется неплохим? Насколько уверенный результат? Тут тоже поможет доверительный интервал.*

Пример 12.3. *Известная мера риска в финансовой математике – отношение Шарпа (Sharpe ratio)*

$$SR = \frac{\mathbb{E}[X]}{\sigma_x} \approx \frac{\overline{X}}{\hat{\sigma}_x} = \widehat{SR},$$

которое может оцениваться, к примеру, в контексте портфелей, где X – доходность. Если X – гауссовские (что неправда в жизни), то оценка $\widehat{SR}$ имеет распределение Стьюдента. А в других случаях?..

Во всех этих случаях мы сталкиваемся с тем, что ничего конкретного не можем сказать про распределение статистик или про распределение данных. Данные, как мы уже замечали, дополнительно получать может быть дорого, поэтому надежды на ЦПТ нужно строить с внимательной оглядкой на количество наблюдений. Кстати, какой типичный размер тестовой выборки вы у себя задаёте на Кагле?..

Но оказывается, что есть способы жить даже в такой ситуации.

12.2 Бутстреп

Бутстреп (bootstrap) можно рассматривать как набор техник, позволяющих искусственно генерериовать больше данных из уже существующих, при этом получается распределение, похожее на исходное распределение выборки. С этими данными можно манипулировать, добиваясь нужных для ЦПТ и других великих теорем предположений.

Пример 12.4. *Положим, у нас есть выборка из* 10 *наблюдений* $X_1, .., X_{10}$ *из некоторого распределения* F *с плотностью. Мы хотели бы оценить среднее, выборочное среднее – наш кандидат:*

$$\mathbb{E}[X] \approx \frac{1}{10} \sum_{i=1}^{10} X_i.$$

Но вот как построить доверительный интервал для $\mathbb{E}[X]$*? Даже ещё проще: как оценить дисперсию* $\frac{1}{10} \sum_{i=1}^{10} X_i$ *при том, что дисперсия* X_i *неизвестна?*

Давайте детальнее всмотримся в

$$\frac{1}{10} \sum_{i=1}^{10} X_i.$$

Эту сумму можно рассматривать как выборочное среднее, взятое от величин $X_1, .., X_{10} \sim F$, как мы привыкли. Но можно посмотреть с другой стороны. Представим себе, что у нас есть дискретная случайная величина, распределённая равновероятно с десятью (положим, повторений в данных нет) возможными значениями $x_1, .., x_{10}$ (реализация выборки). Тогда наверху написано точное среднее; более того, если рассмотреть $f(X_i)$ вместо X_i это всё ещё будет точным средним. Теорема Гливенко-Кантелли даже формальнее утверждает, что эмпирическая функция распределения $\hat{F}_n$ равномерно сходится к настоящей F при росте n:

$$\sup_{t \in \mathbb{R}} |F(t) - \hat{F}_n(t)| \to 0, \quad n \to \infty.$$

Неформально говоря (формально говорить сложно), эмпирическое распределение, о котором мы только что говорили, *похоже* на распределение, из которого была взята выборка, а значит, мы можем попробовать использовать это знание, чтобы сгенерировать как можно больше данных и использовать привычную нам ЦПТ. В этом и состоит ключевая идея бутстрепа.

Важно понять один тонкий момент, который часто упускают из вида: бутстреп НЕ позволяет сделать оценку точнее или доверительный интервал уже; причина простая – для этого нужно больше информации, а значит больше данных об истинном распределении F, а мы умеем семплировать только из $\hat{F}$. Методы бутстрепа (если грубо) всего лишь из данных из непонятного распределения получают данные из более понятного распределения. Главное назначение бутстрепа – это вообще дать возможность построить доверительный интервал в условиях, когда совершенно не ясно, как его строить.

12.2.1 Перцентильный бутстреп

Простейший и самый универсальный метод бутстрепа – перцентильный или наивный бутстреп. Пусть дана выборка $X_1, .., X_n$, и мы умеем по данным вычислять некоторую статистику $\hat{\theta}$. Для того, чтобы построить доверительный интервал для θ, предлагается такой алгоритм:

1. Сгенерировать n_{boot} бутстреп-выборок: в каждую равновероятной выборкой с возвращением из $X_1, .., X_n$ набирается n наблюдений;
2. Для каждой бутстреп-выборки $X_1^{(k)}, .., X_n^{(k)}$ вычислить оценку $\hat{\theta}^{(k)}$, отстортировать оценки по возрастанию;
3. В качестве доверительного интервала уровня $1-\alpha$ задать интервал $(q_{\alpha/2}, q_{1-\alpha/2})$, где q – посчитанные по отсортированному набору оценок перцентили.

Удивительно, но если мы попробуем перегенерировать выборку $X_1, .., X_n$ и повторить эту процедуру много раз, то мы увидим, что вероятность накрытия интервалом истинного значения θ очень близка к заданной $1 - \alpha$ даже при относительно небольших объёмах выборки. То есть, если выборка более-менее репрезентативна (нет такого, что мы забыли в выборке представить хвост распределения с ощутимой вероятностью), то результат неплохой. Но при этом одним только бутстрепом оценку не уточнить.

Если в общем, то теория бутстрепа говорит, что при стремлении к бесконечности n_{boot} и n доверительный интервал сходится к настоящему. При фиксированном n всё равно остаётся ошибка, порождённая разницей между распределением F и распределением $\hat{F}$.

12.2.2 Бутстреп t-статистик

Перцентильный бутстреп неявно использует идею симметричности распределения статистик при построении доверительного интервала. Однако это предположение может быть неверно: некоторые метрики в АБ-тестировании, например, скошены в положительную сторону. Для того, чтобы побороть эту проблему был предложен t-бутстреп.

Предположим, что $X_1, .., X_n$ приходит из скошенного распределения, многие метрики в АБ-тестировании оказываются скошенными вправо в силу их специфики. Переход к t-статистикам позволяет скомпенсировать эффект скошенности и лучше работать с несимметричными распределениями, теперь квантили строятся по распределению без эффекта скоса.

1. Вычислить $\overline{X}$ по исходной выборке;

2. Сгенерировать B бутстреп-выборок, по каждой вычислить

$$t_k^* = \frac{\overline{X} - \overline{X^{(k)}}}{\hat{\sigma}(X^{(k)})/\sqrt{n}}.$$

3. Предложить доверительный интервал уровня $1-\alpha$ как $(\overline{X} + \hat{\sigma} q_{\alpha/2}, \overline{X} + \hat{\sigma} q_{1-\alpha/2})$, где q – перцентили из набора t-статистик.

Но почему только выборочное среднее? Можно построить такую процедуру для общей оценки $\hat{\theta}$, нужно поменять определение t-статистики

$$t_k^* = \frac{\hat{\theta} - \hat{\theta}_k^*}{\hat{\sigma}(\hat{\theta}_k^*)},$$

и доверительный интервал задать как

$$(\hat{\theta} + \sigma(\hat{\theta}) q_{\alpha/2}, \hat{\theta} + \sigma(\hat{\theta}) q_{1-\alpha/2})$$

но тут понадобится как-то вычислить точно или аппроксимировать дисперсию бутстреп-оценки и оценки по исходным данным. Это можно сделать в том числе ещё одним вложенным бутстрепом.

12.3 Проверка гипотез с помощью бутстрепа

До сих пор мы строили доверительные интервалы, но, как мы помним, они имеют прямое отношение к проверке гипотез. Как можно использовать бутстреп для этой задачи? Давайте для конкретики обсудим критерий для проверки гипотезы о равенстве средних в двух выборках, которая часто возникает в контексте АБ-тестирования.

Определимся, что нам нужно от статистического критерия: нужно уметь считать p-value. Нам даны выборки $X_1, .., X_n$ и $Y_1, .., Y_m$, если сложим в одну выборку, то получим $Z_1, .., Z_{n+m}$. Мы проверяем гипотезу

$$H_0: \mathbb{E}[X] = \mathbb{E}[Y], \quad H_A: \mathbb{E}[X] > \mathbb{E}[Y].$$

Положим, гипотеза верна, нам нужно вычислить наименьший уровень значимости, при котором гипотеза отвергается (это и есть p-value) при данном значении статистики. Нам нужно, таким образом, предложить статистику, и посчитать пороговую вероятность ошибки. Один из классических подходов, предложенный Эфроном, состоит в следующем.

1. Тестовая статистика
$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{\hat{\sigma}_x^2/n + \hat{\sigma}_y^2/m}}.$$

2. Модифицируем данные, чтобы сопоставить смещения: $X' = X - \overline{X} + \overline{Z}$, $Y' = Y - \overline{Y} + \overline{Z}$.

3. Сгенерировать n_{boot} бутстреп-выборок $X^{(k)}$ и $Y^{(k)}$ и вычислить по каждой паре выборок $X^{(k)}, Y^{(k)}$ статистику
$$T_k^* = \frac{\overline{X^*}^{(k)} - \overline{Y^*}^{(k)}}{\sqrt{\hat{\sigma}_{x^{(k)}}^2/n + \hat{\sigma}_{y^{(k)}}^2/m}}.$$

4. Оценить p-value из определения как
$$p = \frac{\sum_{i=1}^{n_{boot}} I(T_i^* \geq T)}{n_{boot}}.$$

Для двустороннего теста в последнем пункте нужно оценивать
$$p = 2\min\left(\mathbb{P}\left(T^* \geq T\right),\ \mathbb{P}\left(T^* \leq T\right)\right),$$
но в остальном процедура та же.

Если мы откажемся от p-value, то в случае двусторонней альтернативы здесь транслируется та же идея с доверительными интервалами. Мы оцениваем доверительный интервал уровня $1-\alpha$ с помощью бутстрепа, а дальше, если статистика не попадает в интервал, то гипотеза отвергается на уровне значимости α.

12.4 Jackknife

Есть ещё один интересный способ, как можно оценить смещение и дисперсию вообще почти любой оценки. В сущности, мы тоже, как и в бутстрепе пытаемся вытащить как можно больше из имеющейся выборки. Одна из проблем бутстрепа состоит в том, что он требует много времени и много бутстреп-выборок (часто больше n), если вычисление оценки $\hat{\theta}$ достаточно долгое. В некоторых случаях оказывается полезным более экономичный приём.

Идея Jackknife (швейцарский нож) нацелена на исследование смещения оценки $\hat{\theta}$ и её дисперсии. Оба числа характеризуют изменчивость оценки при изменяющейся выборке.

Здесь предлагается рассмотреть в качестве возмущения n jackknife-выборок, каждая из которых $X(-i)$ есть $X_1, .., X_{i-1}, X_{i+1}, .., X_n$, то есть, исходная выборка без i-го наблюдения. Для иллюстрации давайте рассмотрим выборочное среднее, а позже выпишем более общие оценки.

В основе идеи jackknife есть предположение, что матожидание оценки $\hat{\theta}_n$ (по выборке размера n) раскладывается в степенной ряд по $(1/n)$

$$\mathbb{E}\left[\hat{\theta}_n\right] = \theta + \frac{a_1}{n} + \frac{a_2}{n^2} + \ldots ,$$

где коэффициенты a_i определяются распределением выборки. Если так, то аналогичное верно и для уменьшенной jackknife-выборки:

$$\mathbb{E}\left[\hat{\theta}^*_{(j)}\right] = \mathbb{E}\left[\hat{\theta}_{n-1}\right] = \theta + \frac{a_1}{n-1} + \frac{a_2}{(n-1)^2} + \ldots ,$$

где мы обозначили за $\hat{\theta}^*_{(j)}$ выборочное среднее по выборке с выкинутым наблюдением. Если правильно вычтем из одного другое, то заметим, что

$$\mathbb{E}\left[\hat{\theta}^*_{(j)} - \hat{\theta}_n\right] = 0 + a_1\frac{1}{n(n-1)} + \ldots,$$

смещение теперь порядка $O(1/n^2)$! То есть, мы можем пытаться скорректировать смещённую оценку как

$$\hat{\theta}_{cor} = \hat{\theta} - \widehat{bias}(\hat{\theta}).$$

через оценку смещения

$$\widehat{bias}(\hat{\theta}) = (n-1)(\hat{\theta}^*_{(\cdot)} - \hat{\theta}_n),$$

где $\hat{\theta}^*_{(\cdot)}$ – выборочное среднее jacknife-оценок, а справа стоит выборочное среднее отклонений от оценки по полной выборке. Выборочное среднее – конечно, несмещённая оценка, но попробуйте посмотреть, что будет, если вы в качестве оценки возьмёте смещённую выборочную дисперсию. Но и это не всё, мы ещё можем оценить дисперсию оценки, если применим ещё больше техник:

$$\frac{1}{(n-1)^2}\sum_{i=1}^{n}(\hat{\theta}^*_{(i)} - \hat{\theta}_{(\cdot)})^2.$$

Для более общих оценок, чем выборочное среднее, jackknife записывается похожим образом:

$$\hat{\theta}^*_{\cdot} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)},\ \widehat{bias}(\hat{\theta}) = (n-1)(\hat{\theta}^*_{(\cdot)} - \hat{\theta}_n),,\ \widehat{Var}(\hat{\theta}) = \frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}^*_{(i)} - \hat{\theta}^*_{(\cdot)})^2.$$

Если оценки $\hat{\theta}$ несмещённые, то оценка смещения просто равна нулю. С другой стороны, для распределений, допускающих разложение, оценка смещения будет несмещённой. Jackknife в этом смысле не очень хорошо работает с порядковыми статистиками, квантилями и медианой. Сейчас бутстреп в целом является более гибким и широко применимым. Больше деталей можно найти в [5].

Литература

[1] T. W. Anderson and D. A. Darling. Asymptotic Theory of Certain "Goodness of Fit"Criteria Based on Stochastic Processes. *The Annals of Mathematical Statistics*, 23(2):193 – 212, 1952.

[2] Peter C Austin, Muhammad M Mamdani, David N Juurlink, and Janet E Hux. Testing multiple statistical hypotheses resulted in spurious associations: a study of astrological signs and health. *J Clin Epidemiol*, 59(9):964–969, July 2006.

[3] Erika Cule, Paolo Vineis, and Maria De Iorio. Significance testing in ridge regression for genetic data. *BMC Bioinformatics*, 12(1):372, Sep 2011.

[4] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society. Series B (Methodological)*, 39(1):1–38, 1977.

[5] B. Efron. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.

[6] R. E. Kalman. A New Approach to Linear Filtering and Prediction Problems. *Journal of Basic Engineering*, 82(1):35–45, 03 1960.

[7] S.J. Levitt, S.D. Dubner. *Freakonomics*. NY: Harper Trophy, 2006.

[8] Richard Lockhart, Jonathan Taylor, Ryan J. Tibshirani, and Robert Tibshirani. A significance test for the lasso. *The Annals of Statistics*, 42(2):413–468, 2014.

[9] J. Scott Long and Laurie H. Ervin. Using heteroscedasticity consistent standard errors in the linear regression model. *The American Statistician*, 54(3):217–224, 2000.

[10] Halbert White. A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48(4):817–838, 1980.

[11] Платонов Е.Н. Горяинова Е.Р., Панков А.Р. *Прикладные методы анализа статистических данных*. Изд. дом Высшей школы экономики, 2012.

[12] Ю.М. Кельберт, М.Я. Сухов. *Вероятность и статистика в примерах и задачах. Т.3: теория информации и кодирования*. М.: МЦНМО, 2013.

[13] А.Н. Ширяев. *Основы стохастической финансовой математики*. МЦНМО, 2016.