Статистические критерии согласия на основе оценки скользящего экзамена

Неделько В. М.

Институт математики СО РАН, г. Новосибирск
nedelko@math.nsc.ru

«Интеллектуализация обработки информации»
(ИОИ-12), Италия, г. Гаэта, 8–12 октября 2018 г.

Задача восстановления зависимостей

Предположим, что единственная переменная X принимает целые значения $\{1, ..., M\}$.
Задача восстановления зависимостей:
целевая переменная Y — множество действительных чисел.
Решающей функцией называется соответствие $\lambda\colon X \to Y$.

В задаче оценивания регрессии используется квадратичная функция потерь $\mathcal{L}(y, y') = (y - y')^2$.
В этом случае критерий качества решения есть средний квадрат отклонения.

Метод построения решающих функций

Пусть X и Y — случайные величины с некоторым совместным распределением и пусть $V_N = ((x^i, y^i) \in D \,|\, i = 1, \ldots, N)$ — случайная независимая выборка.

Отображение $Q\colon (X \times Y)^N \to \Lambda$ называется методом (алгоритмом) построения решающих функций.

Здесь $(X \times Y)^N$ – пространство выборок объёма N, а Λ – некоторый класс решающих функций.

Задача дисперсионного анализа

Пусть, например, имеется 10 сортов некоторой культуры, и нужно проверить гипотезу, одинакова ли их урожайность.

В задачах анализа данных часто встречаются номинальные переменные, проверка гипотезы однородности даёт ответ на вопрос об их информативности.

На практике информативность переменных оценивается по скользящему экзамену.

Оказывается, что скользящий экзамен в некоторых случаях эквивалентен классическим критериям проверки гипотез.

Гипотеза однородности

Обозначим:
$N_x = \sum_{i=1}^{N} I(x_i = x)$ – количество точек с $x_i = x$,

$\overline{y}_x = \frac{1}{N_x} \sum_{i=1}^{N} y_i \cdot I(x_i = x)$ – среднее значение y в точке x,

$\overline{y} = \frac{1}{N} \sum_{i=1}^{N} y_i$ – среднее значение y по всей выборке,

$I(\cdot)$ — индикаторная функция.

Статистика Фишера для проверки гипотезы однородности в этих обозначениях запишется как

$$F(V) = \frac{(N-M) \sum\limits_{x=1}^{M} N_x (\overline{y}_x - \overline{y})^2}{(M-1) \sum\limits_{i=1}^{N} (y_i - \overline{y}_{x_i})^2}.$$

Скользящий экзамен

Функционал скользящего экзамена определяется как:

$$\breve{R}(V, Q) = \frac{1}{N} \sum_{i=1}^{N} \mathcal{L}(y^i, \lambda_{Q,V_i'}(x^i)),$$

где $V_i' = V \setminus (x^i, y^i)$ — выборка, получаемая из V удалением i-го наблюдения,

Скользящий экзамен — «суррогат» контрольной выборки.

Будем использовать скользящий экзамен для выбора одного из двух решений: оценивать среднее по объединённой выборке или по каждой подвыборке.

В предположении равенства средних

Пусть метод Q_0 в качестве решения возвращает $\overline{y}$. Тогда оценка качества этого метода посредством скользящего экзамена даст величину

$$\breve{R}(V, Q_0) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \frac{N\overline{y} - y_i}{N-1} \right)^2 = \frac{N}{(N-1)^2} \sum_{i=1}^{N} (y_i - \overline{y})^2.$$

В предположении неравенства средних

Пусть метод Q_1 в качестве решения возвращает $\lambda(x) = \overline{y}_x$. Тогда оценка качества этого метода посредством скользящего экзамена даст величину

$$\breve{R}(V, Q_1) = \frac{1}{N} \sum_{i=1}^{N} \left(y_i - \frac{N_{x_i} \overline{y}_{x_i} - y_i}{N_{x_i} - 1} \right)^2 =$$

$$\frac{1}{N} \sum_{i=1}^{N} \left(\frac{N_{x_i}}{N_{x_i} - 1} \right)^2 (y_i - \overline{y}_{x_i})^2.$$

Утверждение 1

Пусть $N_x = \frac{N}{M}$. Тогда условие $\breve{R}(V, Q_0) - \breve{R}(V, Q_1) > 0$ равносильно

$$F(V) > 1 + \frac{N-1}{N-M} \approx 2.$$

В случае, когда все N_x равны, метод скользящего экзамена эквивалентен критерию Фишера, с пороговым значением статистики $1 + \frac{N-1}{N-M} \approx 2$, что при $M = 2$ соответствует уровню значимости около 0.157.

Зависимое оценивание средних

Та же задача дисперсионного анализа. Но теперь мы знаем, что математические ожидания по выборкам не равны, более того, нет оснований считать, что эти величины вообще как-то связаны друг с другом.
Требуется наилучшим образом оценить эти средние.

Оказывается, нужно взять не просто средние по каждой из выборок, а скорректировать их (большие уменьшить, меньшие увеличить).
Звучит парадоксально: имеем M независимых случайных величин, но оценки их математических ожиданий зависят друг от друга.

Пример

Рассмотрим задачу Zillow Prize: Zillow's Home Value Prediction (Zestimate) (https://www.kaggle.com/c/zillow-prize-1).
Данные представляют собой значения 58 переменных, измеренные для 86795 объектов.

Целевая переменная Y' представляет собой некоторую характеристику, вычисляемую на основе цены объекта недвижимости.
В качестве переменной X будет выступать код региона, который принимает 317 различных значений.

К переменной Y' для удобства применим квантильное преобразование, т.е. введём переменную $Y = \widetilde{\mathsf{F}}(Y')$, $\widetilde{\mathsf{F}}(\cdot)$ – эмпирическая функция распределения.

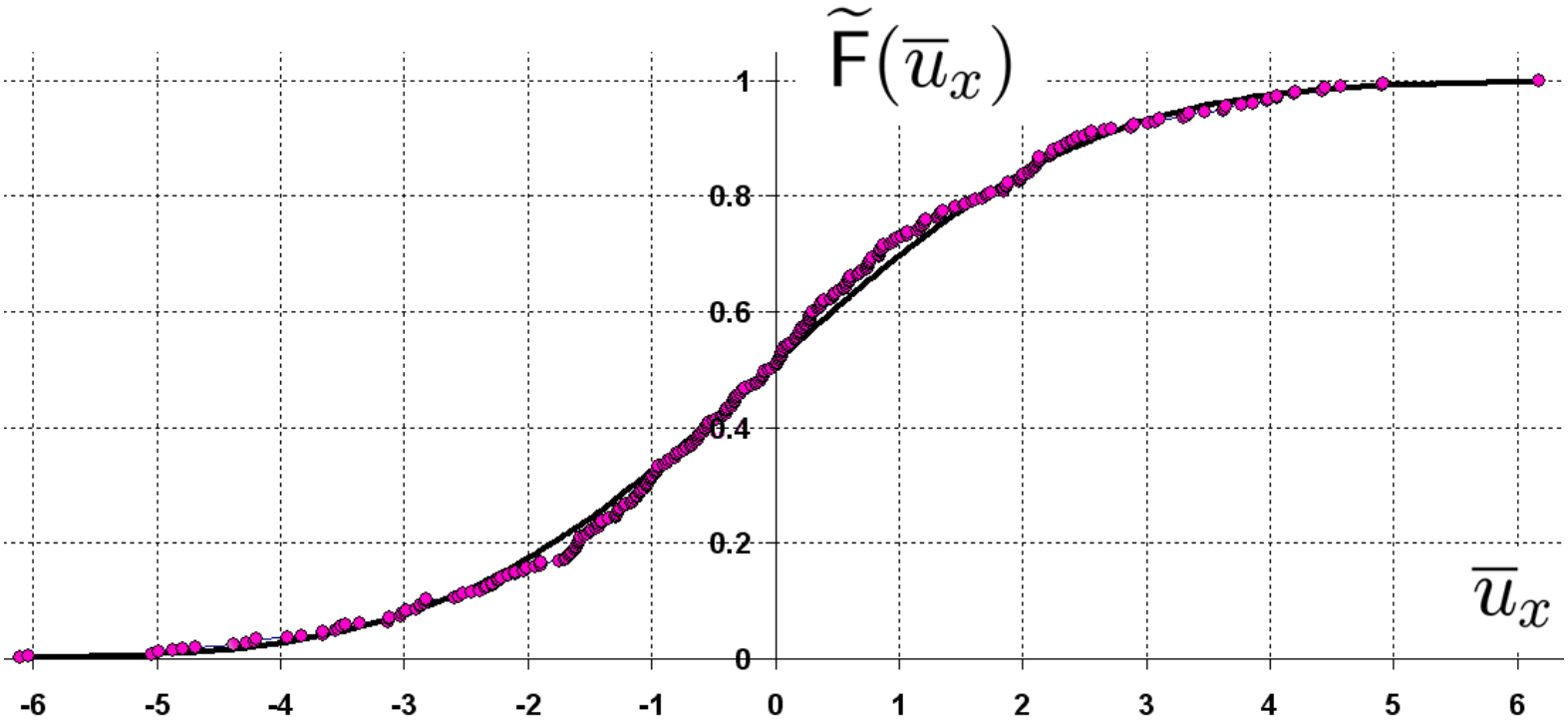
Гипотеза однородности

Введём величины $\overline{u}_x = \sqrt{12 \cdot N_x} \cdot (\overline{y}_x - 0.5)$.

Если справедлива гипотеза однородности (т.е. если Y не зависит от X), то величины $\overline{u}_x$ имеют распределение, близкое к нормальному с параметрами 0, 1.

На рисунке приведена (круглые маркеры) эмпирическая функция распределения $\widetilde{\mathsf{F}}(\overline{u}_x)$. Сплошная чёрная кривая изображает функцию нормального распределения с параметрами -0.06, 2.07, оцененными по обучающей выборке.

Эмпирическая функция распределения



Видим, что распределение очень похоже на нормальное, но дисперсия значимо больше 1, поэтому поведение $\overline{u}_x$ не может быть объяснено случайными флуктуациями частот.

Зависимое оценивание средних

Введём обозначения $\xi_x = \mathsf{E}[Y \mid x]$ – условное математическое ожидание целевой переменной в точке x и $\omega_x = \overline{u}_x - \xi_x$.

Предположим, что целевая переменная подчиняется следующей модели $Y = \xi_x + \omega$, $\mathsf{E}\omega = 0$, $\mathsf{D}\omega = \sigma^2$.
Тогда наблюдаемые значения $\overline{y}_x$ являются реализациями случайной величины $\zeta_x = \xi_x + \omega_x$, $\mathsf{E}\omega_x = 0$, $\mathsf{D}\omega_x = \frac{\sigma^2}{N_x}$.
В силу центральной предельной теоремы величина ω_x имеет приближённо нормально распределение.

Байесовская модель

В исходной постановке задачи значения ξ_x не являются случайными, но $\overline{y}_x$ эмпрически выглядит как величина с нормальным распределением.

Это даёт основания рассмотреть байесовскую постановку, в которой ξ_x будет случайной величиной с нормальным распределением.

Оптимальное решение

В байесовской модели оптимальное решение есть

$$\lambda(x) = \mathsf{E}[\xi_x \mid \zeta_x = \overline{y}_x] = \mathsf{E}Y + (\overline{y}_x - \mathsf{E}Y)\cdot\alpha_x = \mathsf{E}Y\cdot(1-\alpha_x) + \overline{y}_x\alpha_x,$$

где

$$\alpha_x = 1 - \frac{\mathsf{D}\omega_x}{\mathsf{D}\zeta_x} = 1 - \frac{\sigma^2}{N_x\mathsf{D}\zeta_x}.$$

В случае, когда все N_x одинаковы, при подстановке этих оценок в выражение для α_x получаем $\alpha_x = 1 - \frac{1}{F(V)}$.

Последнее выражение представляется достаточно интересным. В следующем разделе мы его получим ещё раз, из других соображений.

Усреднённое решение

Методы Q_0 и Q_1 строят решения, исходя из того, принимается или отвергается гипотеза однородности. Поскольку мы не можем быть полностью уверены в справедливости гипотез, представляется естественным рассмотреть решения, основанные на усреднении базовых решающих функций.

Пусть метод Q_α в качестве решения возвращает $\lambda(x) = \alpha\overline{y}_x + (1-\alpha)\overline{y}$. Тогда оценка качества этого метода посредством скользящего экзамена даст величину

$$\breve{R}(V, Q_\alpha) = \frac{1}{N}\sum_{i=1}^{N}\left(y_i - \frac{N_{x_i}\overline{y}_{x_i} - y_i}{N_{x_i} - 1}\cdot\alpha - \frac{N\overline{y} - y_i}{N-1}\cdot(1-\alpha)\right)^2.$$

Утверждение 2

Обозначим $\alpha^* = \arg\min_\alpha \breve{R}(V, Q_\alpha)$.

При $N_x = \frac{N}{M}$ имеет место

$$\alpha^* = \frac{F(V) - 1}{F(V) + \frac{M-1}{N-M}} \approx 1 - \frac{1}{F(V)}.$$

Выводы

- Метод скользящего экзамена в некоторых случаях эквивалентен критерию согласия Фишера.
- Оптимизация ансамбля по скользящему экзамену оказалась эквивалентна некоторой байесовской модели.
- Оценки средних для независимых переменных могут быть зависимы. Ансамблевые методы это учитывают.

Виды предположений

- Полная вероятностная модель: позволяет синтезировать данные.
- Неполная вероятностная модель: описывает целевую переменную. Пример: логистическая и классическая регрессии.
- Модель в виде решающей функции. В роли статистической гипотезы выступает предположение, что используемый метод «хорошо решает задачу». Такие «гипотезы» проверяются методом скользящего экзамена.

Параметрический и непараметрический подходы могут рассматриваться в рамках единого подхода.