Работает ли технический анализ?

Фадеева Анастасия,Каглинская Мария

ВШЭ, ФКН, БПМИ-141

28 ноября 2016 г.

Data snooping

Data snooping - это применение анализа данных для поиска статистически значимых закономерностей, без предварительного построения гипотез о причинно-следственных связях.

- проверяем много гипотез в поисках корреляций на заданном уровне значимости
- какие-то из них оказываются случайными и ничего не значат
- делаем неверные выводы

Постановка задачи

Обозначения:

- Стратегии (модели) $\delta_{k,t}$ $k = 1, \ldots, m$
- $\delta_{0,t}$ базовая модель
- Моменты времени с данными $t = 1, \ldots, n$
- Целевая переменная ξ_t
- Функция потерь $L_{k,t} = L(\xi_t, \delta_{k,t})$
- Сравнение k-ой модели с базовой $d_{k,t} = L_{0,t} - L_{k,t}$

Задача:
Сравнить m гипотез с ассимптотическим контролем FWE (вероятность отвергнуть хотя бы одну верную нулевую гипотезу).

$$\limsup_{n} FWE_P \leqslant \alpha$$

Имеющиеся методы:

Существует ли модель лучше базовой?

- Reality Check
- Superior Predictive Ability

Какая модель лучше базовой?

- stepwise Reality Check
- stepwise Superior Predictive Ability

Центральная предельная теорема

$$n^{1/2}(\bar{d} - \mu) \to N(0, \Omega)$$

- $\bar{d}$ – это вектор средних (по времени) значений для моделей
- $\mu = (E(d_{1,t}) \ldots E(d_{m,t}))$
- $\Omega = avar(n^{1/2}\ \bar{d})$

Reality Check, White

Зададим вектор μ так, что $\mu_k = E(d_{k,t})$.

- H_0: $\mu \leq 0$
- $T^{RC} = max(n^{1/2}\bar{d}_1, \dots, n^{1/2}\bar{d}_m)$ - статистика
- предполагаем все $\mu_i = 0$, хотя все $\mu_i < 0$ также соответствуют H_0
- по теореме нулевое распределение: $n^{1/2}\bar{d} \sim N(0, \Omega)$
- По сути с помощью статистики проверяем, что все модели хуже базовой, если хотя бы одна нет- отвергаем гипотезу.

Superior Predictive Ability, Hansen

- H_0: $\mu \leq 0$ аналогично White: все модели хуже базовой
- $T^{SPA} = \max(\max_{k=1 \dots m} \frac{n^{1/2}\bar{d}_k}{\widehat{\omega_k}}, 0)$ - статистика
- Теперь не предполагается, что $\mu_k = 0$. Введем оценку на $\widehat{\mu_k} = \bar{d}_k I\{\frac{n^{1/2}\bar{d}_k}{\widehat{\omega_k}} \geq -\sqrt{2 \log \log n}\}$
- ЦПТ $n^{1/2}(\bar{d} - \mu) \to N(0, \Omega)$

Reality Check vs Superior Predictive Ability

Основным отличием и преимуществом SPA является то, что в нулевом распределении используется оценка для μ, учитывающая ее отрицательные значения, что позволяет повысить мощность.

Реализация Superior Predictive Ability

- $d^*_{b,t}$ $b = 1 \dots B$ вектор длины m полученный с помощью бутстрепа
- вычисляем $\widehat{\omega^*_{k,B}}$
- $Z^*_{k,b,t} = d^*_{k,b,t} - \bar{d}_k I\{\bar{d}_k \geq -\sqrt{\frac{\widehat{\omega^2_k}}{n} 2 \log \log n}\}$
- $\bar{Z}^*_{k,b} = n^{-1} \sum_{t=1}^{n} Z^*_{k,b,t}$ $k = 1 \dots m$
- $T^{SPA_*}_{b,n} = \max(0, \max_{k=1 \dots m} \frac{n^{1/2} \bar{Z}^*{}_{k,b}}{\widehat{\omega}_k})$ $b = 1 \dots B$
- $\widehat{p}_{SPA} = \sum_{b=1}^{B} \frac{I\{T^{SPA_*}_{b,n} > T^{SPA}_{b,n}\}}{B}$
- получили p-value и можем проверить гипотезу

Stepwise Reality Check

Хотим найти модели, которые хуже базовой, а не просто знать, что они есть.
Будем для каждой модели проверять:

- H_0^i: $\mu_i \leq 0$, где $i = 1, \ldots, m$
- Необходима поправка на множественную проверку гипотез
- Хотим, с одной стороны, сохранить асимптотический контроль, с другой- хотим как можно большую мощность (вероятность отвергнуть неверную гипотезу)
- По сути с помощью статистики проверяем, что конкретная модель хуже базовой

Stepwise Reality Check Algorithm

1. Сортируем статистики всех методов от самой большой к самой маленькой, пронумеруем их от r_1 до r_m
2. Зададим $j = 1,\ R_0 = 0$
3. Строим доверительный интервал для μ вида: $[d_{r_1} - c_i; \infty) \times \cdots \times [d_{r_m} - c_i; \infty)$ уровня доверия α (где i- номер интерации)
4. Отвергаем все нулевые гипотезы, доверительные интервалы для которых не содержат 0 и удаляем их статистики (длz интервалов с $R_j + 1$ до m)
5.
   - Если на этом шаге ни одну гипотезу не отвергли, заканчиваем процедуру
   - Иначе, $j = j + 1, R_j =$ число гипотез отвергнутых к этому моменту, возращаемся на шаг 3

Таким образом все модели, гипотезы для которых мы отвергли, дают качество лучше, чем базовая.

Stepwise Reality Check

Как же ищем c_i?

- Будем искать $c_i = inf\{x : Prob\{\max_{R_j+1 \leq i \leq m}(d_i - \mu_i) \leq x\} \geq 1 - \alpha\}$
- По сути c_i — это $1 - \alpha$-квантиль распределения $\max(d_i - \mu_i)$ (просто при применении определения)

Stepwise Superior Predictive Ability

Это stepwise метод, основанный на алгоритме Хансена.

- H_0^i: $\mu_i \leq 0$, где $i = 1, \dots, m$
- ЦПТ $n^{1/2}(\bar{d} - \mu) \to N(0, \Omega)$

Алгоритм:

- Сравниваем $\sqrt{n}\bar{d}_1 \dots \sqrt{n}\bar{d}_m$ с $\widehat{q}^*_{\alpha_0}$ (аналог c_i) методом stepwise
- $\widehat{q}_{\alpha_0} = inf\{q : Prob\sqrt{n}max_{k=1\dots m}(\bar{d}^*_k - \bar{d}_k + \widehat{\mu}_k) \leq q\} \geq 1 - \alpha_0\}$
- $\widehat{q}^*_{\alpha_0} = \max(\widehat{q}_{\alpha_0}, 0)$
- $\widehat{\mu}_k = \bar{d}_k I\{\sqrt{n}\bar{d}_k \leq -\widehat{\omega}_k\sqrt{2\log\log n}\}$

Studentisation

Все вышеперечисленные модели можно улучшить (и они улучшались таким образом своими авторами) с помощью стьюдентизации.
Зачем?

- Дает большую мощность (можно показать, что после стьюдентизации будем отвергать не меньше гипотез)
- Для решения прикладных задач будем использовать бутстреп, а для него доверительные интервалы строятся на основании стьюдентизированных статистик
- Благодаря стьюдентизации получаются более сбалансированные в полане индивидуального покрытия доверительные интервалы (за счет того, что домножаем на дисперсии).

Чем step-алгоритмы лучше

Step-алгоритмы (Stepwise Reality Check Algorithm и Stepwise Superior Predictive Ability) находят множество гипотез, которые статистически значимо лучше базовой модели, а Reality Check Algorithm и Superior Predictive Ability проверяют наличие модели лучшей чем базовая.

И еще раз методы:

Существует ли модель лучше базовой?

Reality Check

Superior Predictive Ability

Какая модель лучше базовой?

stepwise Reality Check

stepwise Superior Predictive Ability

Примеры задач

- Трейдер хочет сравнить модель, которую он использует, с новыми. У него есть данные за некоторый прошедший период.
- Менеджер выбирает в какой из хэдж фондов вложить деньги. Базовая модель в этом случае – безрисковое вложение.

Спасибо за внимание!