Математические основы машинного обучения и прогнозирования

В.В.Вьюгин

ПреМоЛаб

logo

Задачи машинного обучения

- Статистическая теория машинного обучения
- Предсказания с использованием прогнозов экспертов
- Универсальные предсказания

Статистическая теория машинного обучения

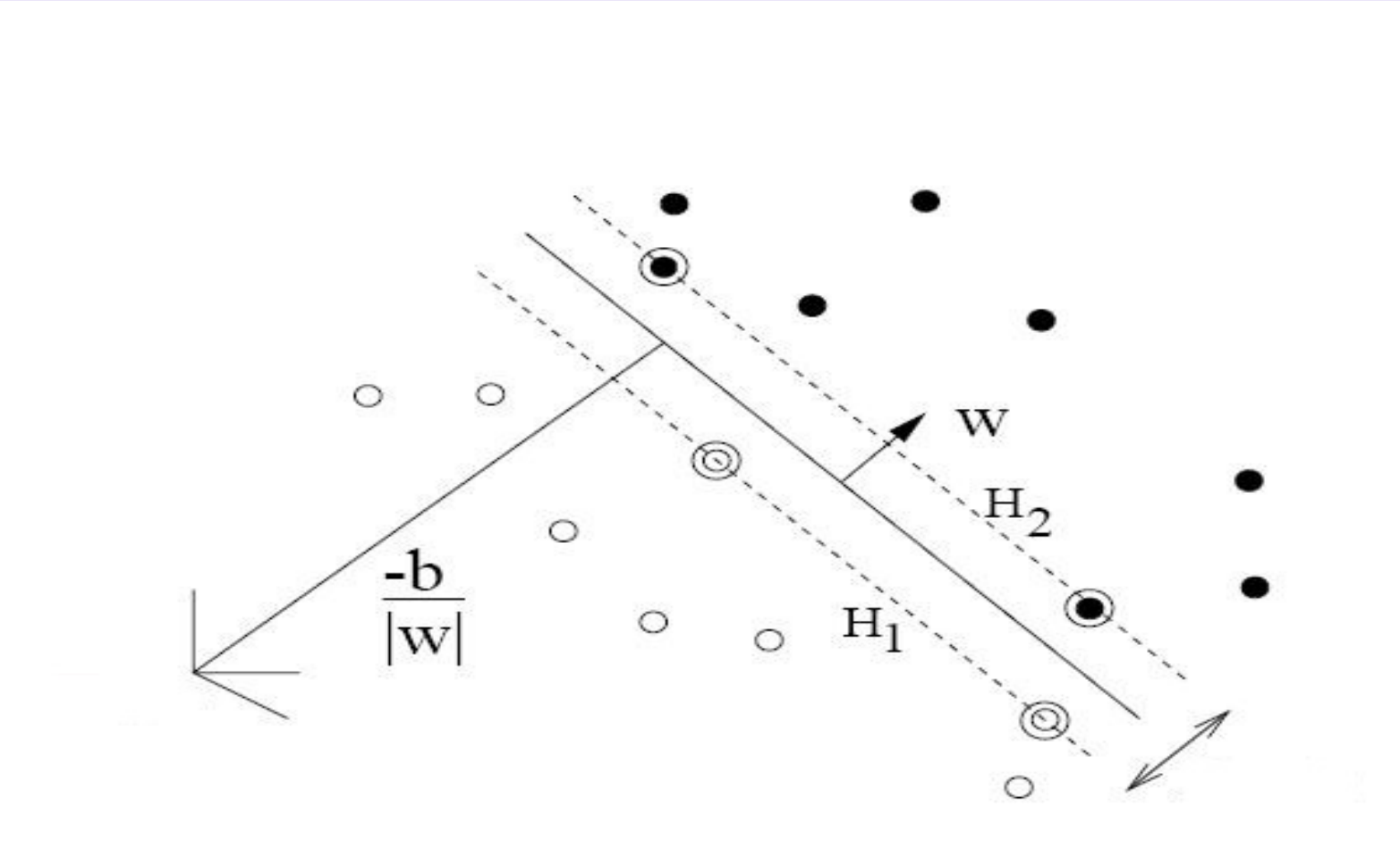
$S = ((x_1, y_1), \ldots, (x_l, y_l))$ – случайная выборка, где $x_i \in \mathscr{R}^n$, $y_i \in \{0, 1\}$

Задача 1: построить классификатор $h : \mathscr{R}^n \to \{0, 1\}$ по выборке S

Задача 2: оценить предсказательную способность классификатора $h(x)$

Ошибка обобщения $\mathrm{err}_P(h) = P\{h(x) \neq y\}$

Эмпирическая ошибка $\mathrm{err}_S(h) = \frac{1}{l}|\{i : h(x_i) \neq y_i, 1 \leq i \leq l\}|$


w
H2
H1
-b
|w|

Theorem

Задан класс H функций классификации. При $l > 2/\varepsilon$ имеет место оценка

$$P^l\{S : (\exists\, h \in H)(\mathrm{err}_S(h) = 0 \,\&\, \mathrm{err}_P(h) > \varepsilon)\} \leq 2B_H(2l)e^{-\varepsilon l/4}.$$

Сложность класса H оценивается с помощью функции роста

$$B_H(l) = \max_{(x_1, x_2, \ \ldots, \ x_l)} |\{(h(x_1), h(x_2), \ \ldots, \ h(x_l)) : h \in H\}|$$

$$B_H(l) \leq 2^l \text{ при } l \leq d \text{ и } B_H(l) = O(l^d) \text{ при } l > d$$

d – VC-размерность (может быть $d = \infty$)

Линейные классификаторы

$$h(x) = \begin{cases} 1, \text{ если } f(x) \geq 0, \\ -1 \text{ в противном случае,} \end{cases}$$

где $f(x) = (w \cdot x) + b$ – гиперплоскость, $w, x \in \mathscr{R}^n$.

VC-размерность класса линейных классификаторов равна $n+1$

Способы измерения емкости класса функций:

- VC-размерность
- пороговая размерность (fat-размерность)
- средние Радемахера

С вероятностью $1-\delta$ для всех $f \in \mathscr{F}$

$$E_P(f) \leq \tilde{E}_{z^l}(f) + 2\mathscr{R}_l(\mathscr{F}) + \sqrt{\frac{\ln \frac{2}{\delta}}{2l}}$$

Prediction with expert advice

Предсказания с использованием прогнозов экспертов

Общая схема:

FOR $t = 1, 2, \ldots, T$
Эксперт i выдает прогноз p_t^i, $i = 1, \ldots, N$
Предсказатель выдает прогноз p_t
Природа выдает исход $\omega_t \in \{0, 1\}$
ENDFOR

Цель предсказателя – предсказывать не хуже наилучшего эксперта (с некоторой точностью).

Алгоритм взвешенного большинства

FOR $t = 1, 2, \ldots, T$
Эксперт i выдает прогноз p_t^i, $i = 1, \ldots, N$
Алгоритм $WMA(\varepsilon)$ выдает прогноз p_t:
IF $\sum_{i:p_t^i=0} w_t^i > \sum_{i:p_t^i=1} w_t^i$ THEN $p_t = 0$ ELSE $p_t = 1$
Природа выдает исход $\omega_t \in \{0, 1\}$
Производим пересчет весов экспертов:

$$w_{t+1}^i = \begin{cases} (1-\varepsilon) w_t^i, \text{ если } p_t^i \neq \omega_t, \\ w_t^i, \text{ в противном случае.} \end{cases}$$

ENDFOR

Theorem

Для любого T выполнено $L_T \leq (2+\varepsilon) \min_i L_T^i + \left(\frac{2}{\varepsilon}\right) \ln N$.

Вероятностный алгоритм

Выбираем эксперта i с вероятностью

$$P_t^i = \frac{w_t^i}{\sum_{j=1}^{N} w_t^j}$$

и полагаем прогноз нашего алгоритма $p_t = p_t^i$.

Theorem

Для любого T выполнено $E(L_T) \leq (1+\varepsilon)\min_i L_T^i + O(\ln N)$.

Предсказания с использованием прогнозов экспертов

Прогноз $p \in [0,1]$, исход ω – произвольный,

Примеры: $\lambda(\omega,p) = (\omega - p)^2$, $\lambda(\omega,p) = |\omega - p|$,

$$\lambda(\omega,\gamma) = \begin{cases} -\ln\gamma, \text{ если } \omega = 1, \\ -\ln(1-\gamma), \text{ если } \omega = 0. \end{cases}$$

FOR $t = 1, 2, \ldots, T$

Эксперт i выдает прогноз p_t^i, $i = 1, \ldots, N$

Алгоритм выдает прогноз p_t

Природа выдает исход $\omega_t \in \{0,1\}$

$L_t^i = L_{t-1}^i + \lambda(\omega_t, p_t^i)$

$L_t = L_{t-1} + \lambda(\omega_t, p_t)$

ENDFOR

Алгоритм экспоненциального взвешивания

Прогноз Алгоритма вычисляется по формуле

$$p_t = \sum_{i=1}^{N} w^*_{i,t-1} p^i_t, \tag{1}$$

где $w^*_{i,t-1} = \frac{e^{-\eta L^i_{t-1}}}{\sum_{j=1}^{N} e^{-\eta L^j_{t-1}}}$, $\eta > 0$.

Theorem

Пусть $\lambda(\omega, p)$ выпуклая по p.
Для любого i выполнено $L_T \leq \min_i L^i_T + O(\sqrt{T \ln N})$.

Агрегирующий алгоритм Вовка

Для специальных функций потерь $\lambda(\omega, p) = (\omega - p)^2$,

$$\lambda(\omega, \gamma) = \begin{cases} -\ln \gamma, \text{ если } \omega = 1, \\ -\ln(1-\gamma), \text{ если } \omega = 0. \end{cases}$$

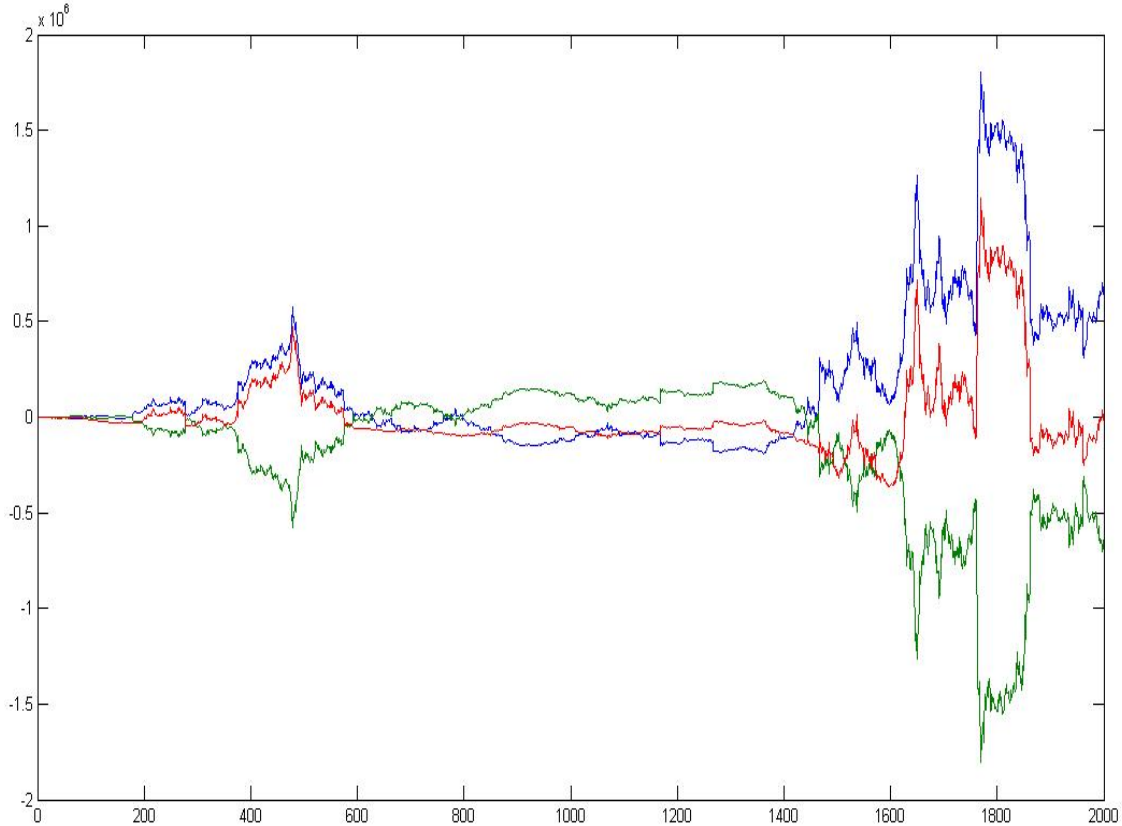
регрет смешивающего алгоритма не зависит от T:

Theorem

Для любого i выполнено $L_T \leq \min\limits_i L_T^i + O(\ln N)$.

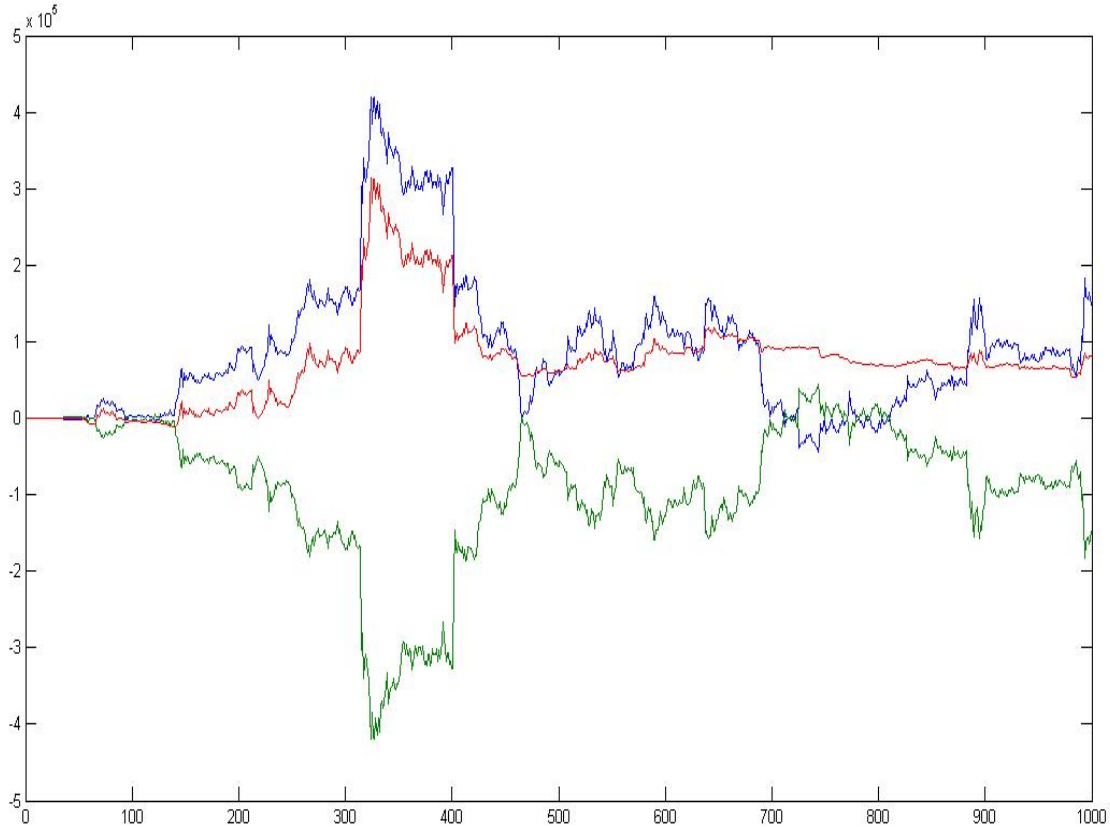
Для остальных функций потерь – регрет как обычно $O(\sqrt{T \ln N})$.

Два эксперта с нулевой суммой



logo

Два эксперта с нулевой суммой



logo

Универсальные предсказания

Универсальные предсказания

Источник исходов $\omega_1, \omega_2 \ldots$ – “черный ящик”

FOR $n = 1, 2, \ldots$
Предсказатель анонсирует прогноз p_n.
Природа анонсирует исход ω_n.
ENDFOR

Цель предсказателя: $p_n - \omega_n \to 0$ при $n \to 0$ для любого источника последовательности $\omega_1, \omega_2, \ldots$.

Калибруемость предсказаний

- $\omega_n = 1$ – дождь в n-й день
- p_n – вероятность дождя в n-й день
- Предсказатель – калибруемый, если дождь случается так же часто, как он прогнозируется предсказателем:
- если дождь случается в 80% всех дней, для которых предсказатель давал прогноз $p_n = 0.8$, и т.д.
- Величина среднего отклонения частоты исходов ω_n от прогнозов p_n для тех n, где $p_n \approx p^*$, для различных значений p^* может использоваться как тест для выявления “плохих” предсказателей.

Универсальные предсказания

Рассмотрим произвольные индикаторные функции $I(p)$ на $[0,1]$: $I(p) = 0$ или $I(p) = 1$.

Последовательность прогнозов $p_1, p_2, \ldots$ калибруется по Дэвиду на бесконечной последовательности $\omega_1, \omega_2, \ldots$, если для любой индикаторной функции $I(p)$ калибровочная ошибка стремится к нулю при $n \to \infty$:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(p_i)(\omega_i - p_i) \to 0.$$

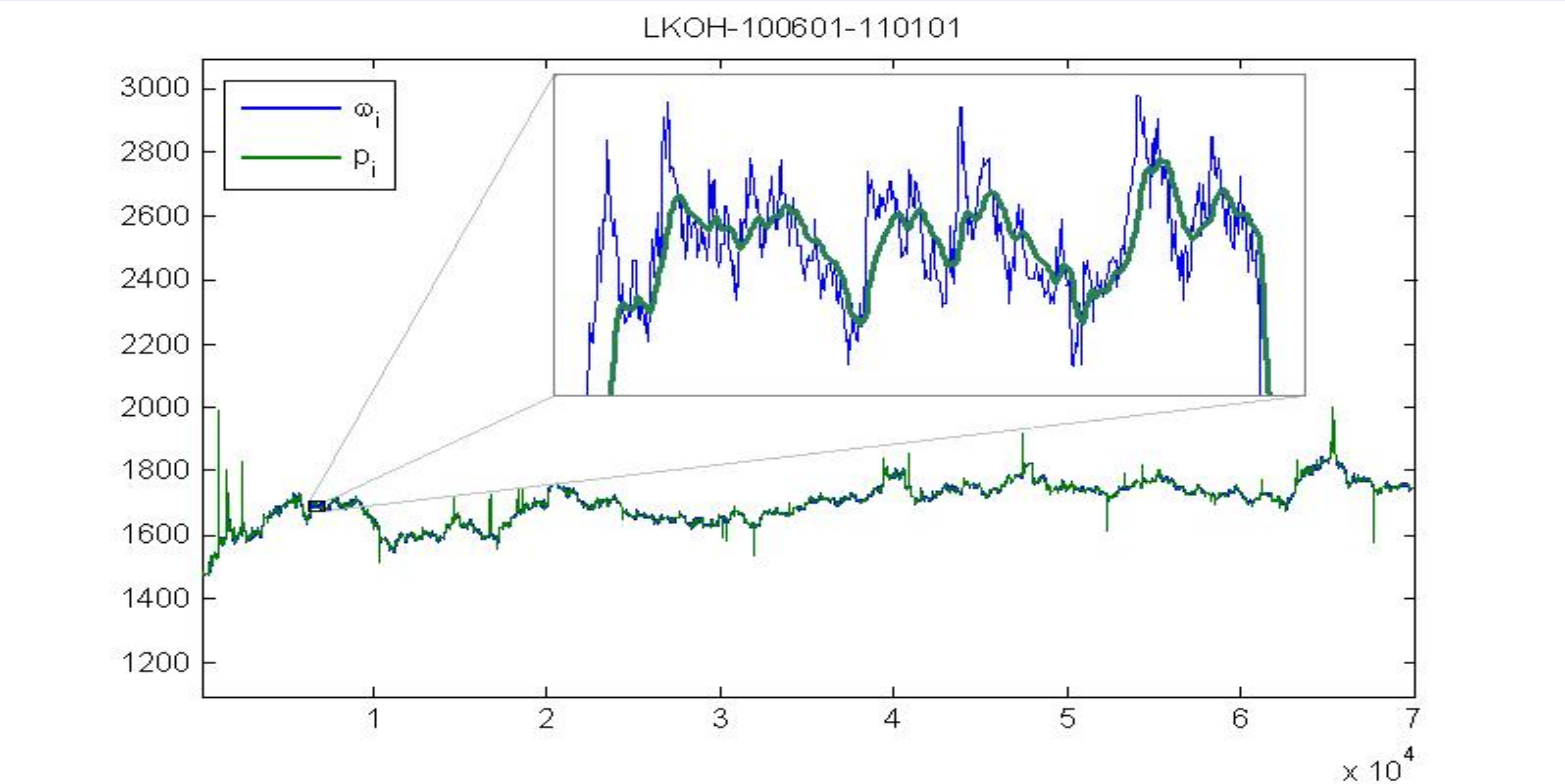
Универсальные предсказания

- Не существует универсального детерминированного предсказателя, который выдерживает все калибровочные тесты
- Существует универсальный рандомизированный предсказатель, выдерживающий каждый калибровочный тест с вероятностью единица (Foster, Vohra)

Существует рандомизированный алгоритм, выдающий случайные числа $\tilde{p}_i$ так, что для любой индикаторной функции $I(p)$ с вероятностью единица

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i)(\omega_i - \tilde{p}_i) \to 0.$$

Универсальные предсказания



Пример последовательности данных $\omega_1, \omega_2, \ldots$ и последовательности калибруемых прогнозов $p_1, p_2, \ldots$

Игра на финансовом рынке

FOR $n = 1, 2 \ldots$
Рынок сообщает информацию x_n
Трейдер покупает C_n акций по цене S_{n-1}
Рынок сообщает новую цену S_n акции.
Трейдер продает все акции по S_n:
$\mathscr{K}_n = \mathscr{K}_{n-1} + C_n(S_n - S_{n-1})$, где $\mathscr{K}_0 = 0$.
ENDFOR
Строим рандомизированные предсказания цены $\tilde{p}_i$ так, что для любой индикаторной функции $I(p, x)$ с вероятностью единица

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} I(\tilde{p}_i, \tilde{x}_i)(S_i - \tilde{p}_i) \to 0.$$

Сравниваем стратегии двух видов:

- $C_n = D(x_n)$ – стационарная стратегия (непрерывная функция)
- $C_n = \tilde{M}_n$ – выдается рандомизированным алгоритмом, который использует калибруемые предсказания цены S_n акции.

Theorem

Для любой непрерывной функции $D(x)$

$$\liminf_{n\to\infty} \frac{1}{n}\left(\mathcal{K}_n^M - \mathcal{K}_n^D\right) \geq 0$$

почти наверное.

Russian Crisis 2008: 080501-081101

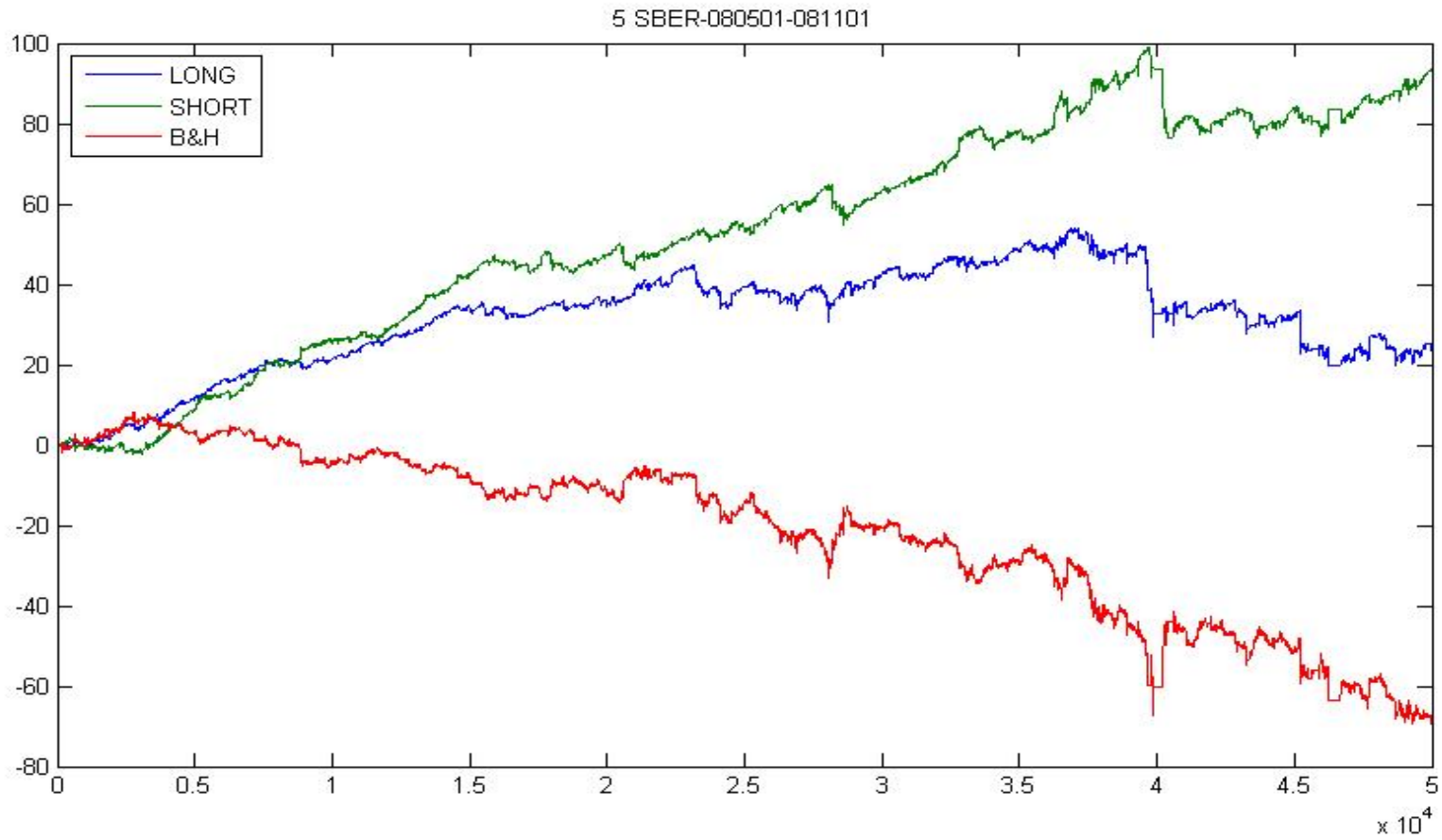


Рис.: SBER mean income of 100 runs

Russian Crisis 2008: 080501-081101

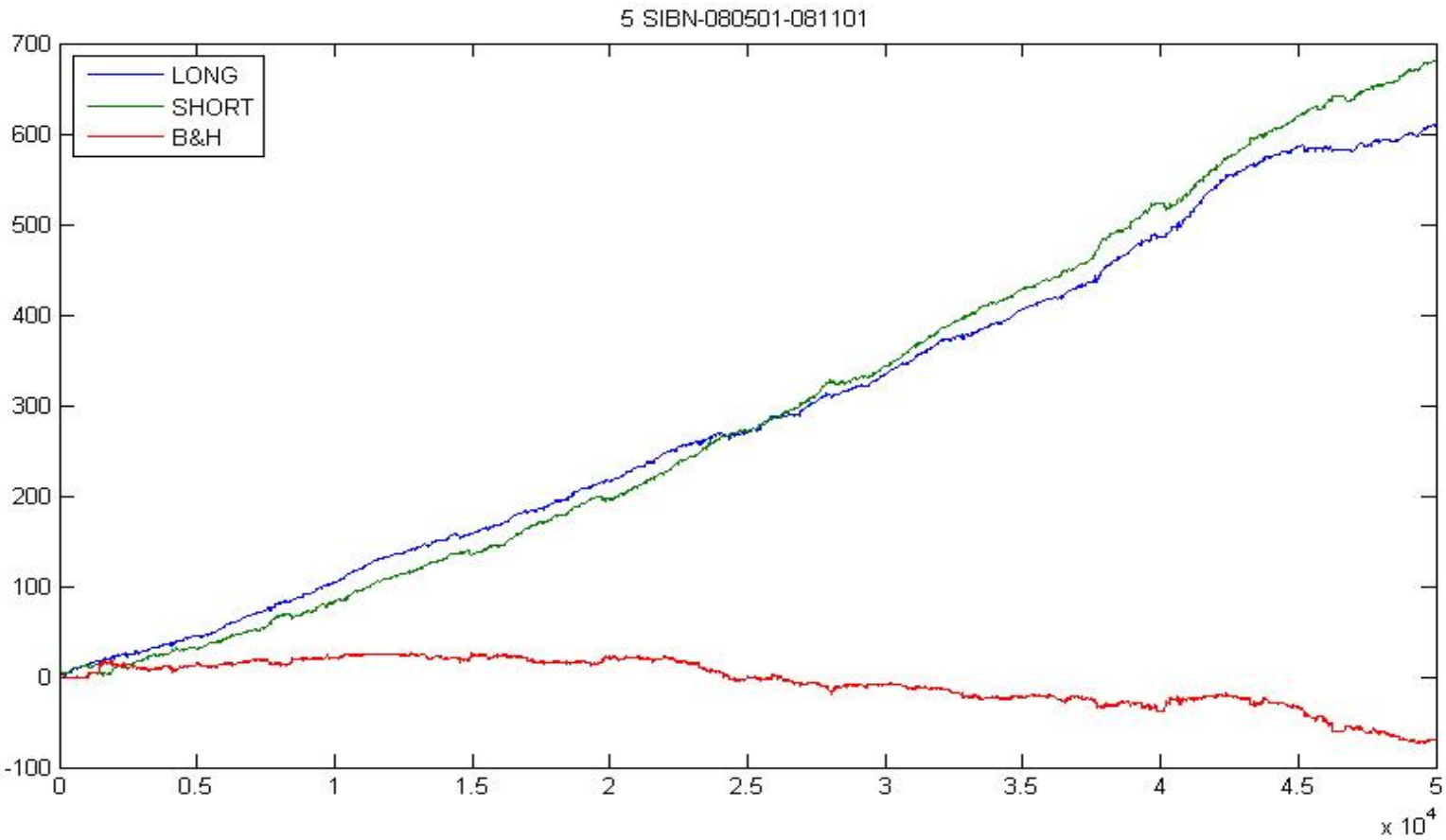
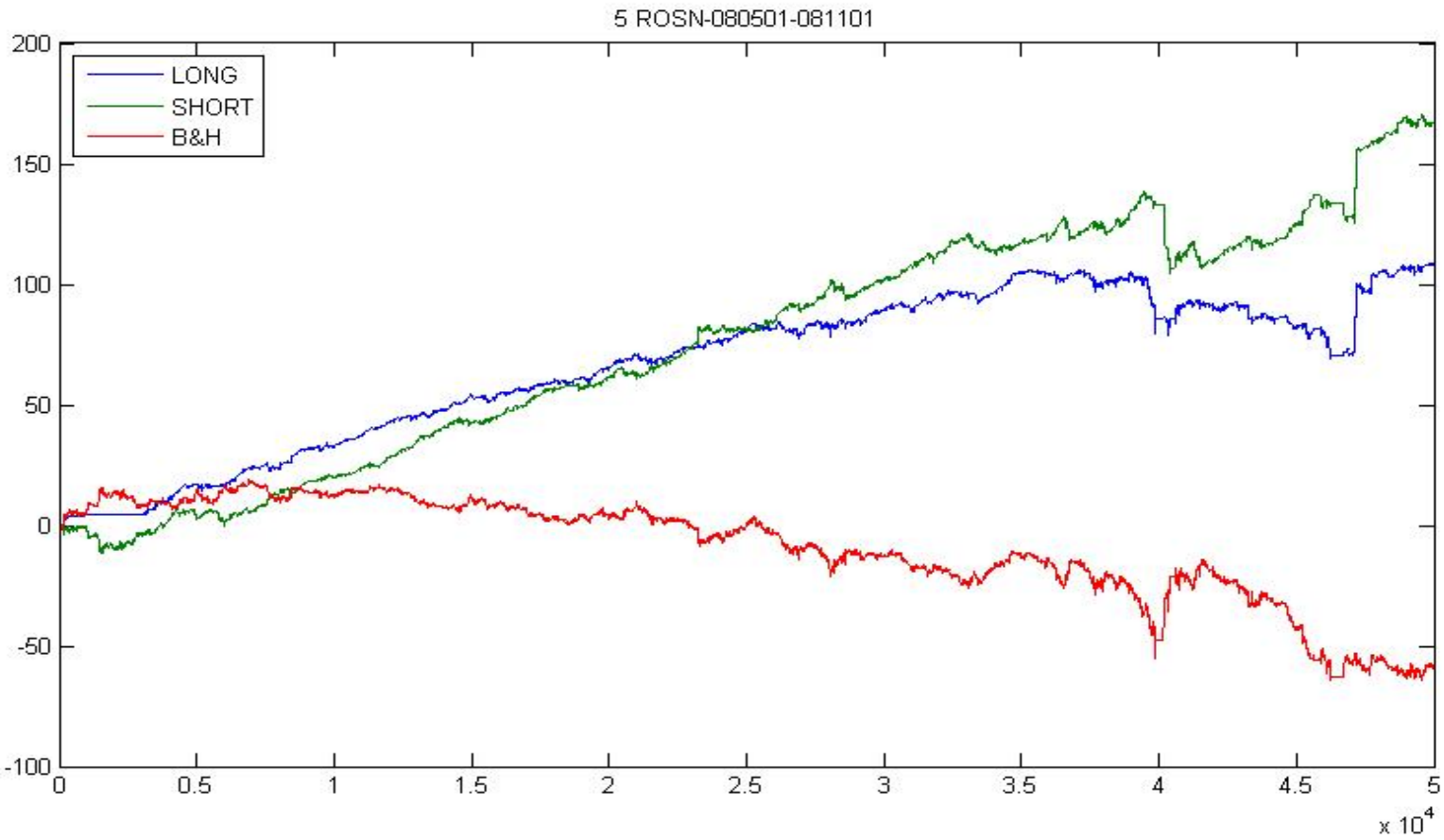


Рис.: SIBN mean income of 100 runs

Russian Crisis 2008: 080501-081101



logo

Рис.: ROSN mean income of 100 runs

Russian Crisis 2008: 080501-081101

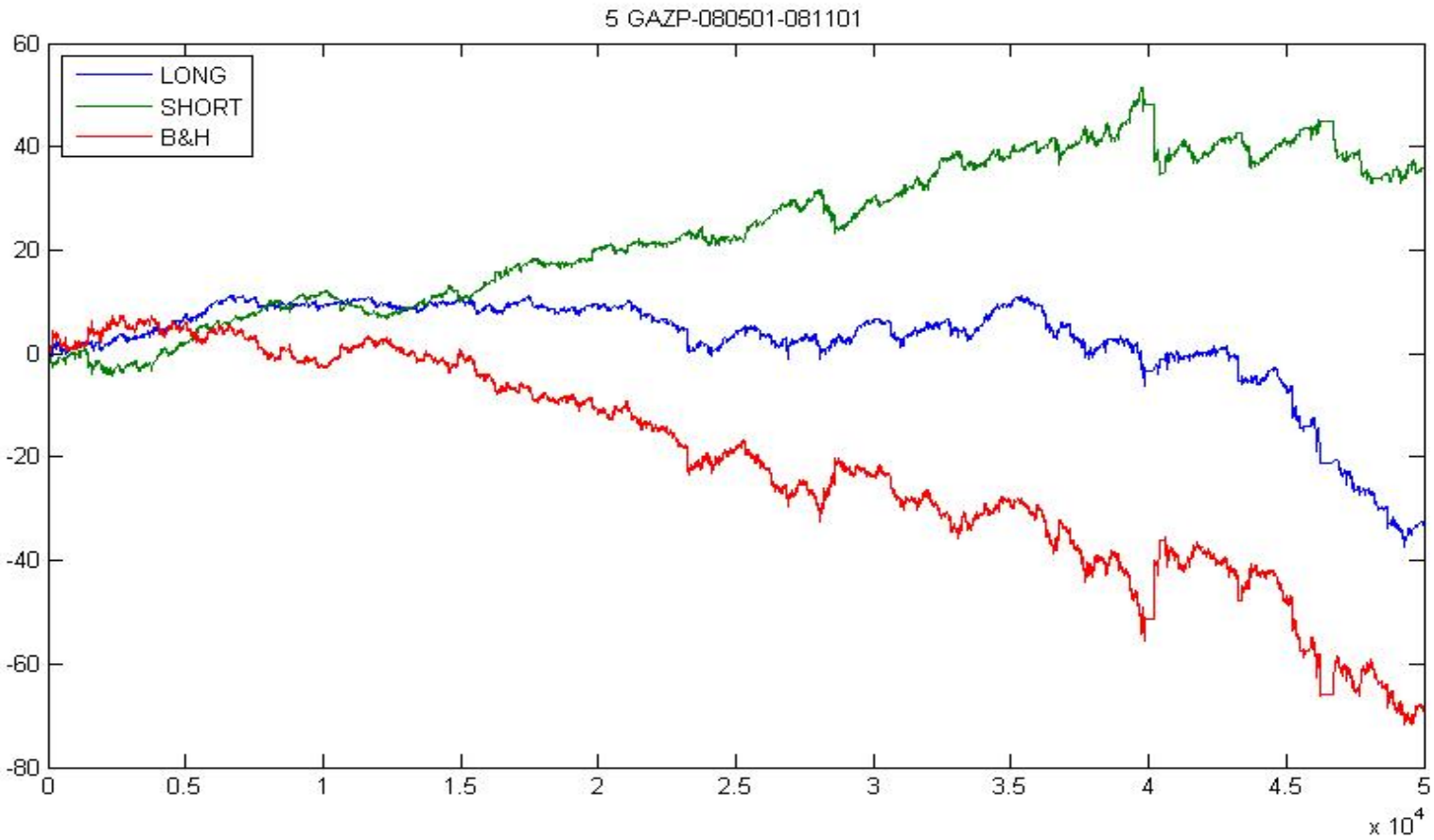


Рис.: GAZP mean income of 100 runs