Модель алгоритмов классификации: информационный подход

С.И. Гуров

МГУ имени М.В. Ломоносова, Москва, Россия
e-mail: sgur@cs.msu.ru

XXII Международная конференция
«Математика. Экономика. Образование»

27 мая – 3 июня 2014 г. / Краснодарский край, пос. Дюрсо

Содержание

Классифицирующий алгоритм

- Задано множество *объектов* $\mathcal{X}$.
- Известно, множество $\mathcal{X}$ разбито на классы $\mathcal{X}_0$ и $\mathcal{X}_1$, однако информация о таком разбиении содержится только в указании о принадлежности к данным классам элементов конечной *обучающей последовательности* или выборки *прецедентов* из $\mathcal{X}$.
- Рассматриваются множество *классифицирующих алгоритмов* $\{A\}$, реализующих отображения $A : \mathcal{X} \to \{0, 1\}$:

$$A(x) = i \;\Leftrightarrow\; x \in \mathcal{X}_i,\; i \in \{0, 1\}.$$

Построение алгоритма классификации и его надёжность

- *Надёжность классификации* $P(A)$ — вероятность правильной классификации данным алгоритмом A произвольного объекта из $\mathcal{X}$.
- Считается, что построение алгоритма классификации A проводилось
  - (возможно, исключительно) на основе информации о принадлежности элементов обучающей последовательности (*обучение по прецедентам*);
  - так, чтобы максимизировать надёжность $P(A)$.

Оценка надёжности классификации

Задача:

Оценить вероятность $P(A)$.

— важнейшая задача в теории обучения по прецедентам

Проблема подхода:

На основе каких подходов строить оценки $P(A)$?

VC-теория

Большинство исследователей решает эту задачу в рамках того или иного развития *теории Вапника-Червоненкиса* (*VC-теория*) базирующейся на следующих допущениях:

Гипотеза 1. На множестве $\mathcal{X}$ определено вероятностное пространство с (возможно неизвестной) вероятностной мерой $P(X)$, $X \subseteq \mathcal{X}$, и любой рассматриваемый набор образов $\{x_1, x_2, \ldots, x_n\}$ из $\mathcal{X}$ считается (если явно не указано иначе), реализацией независимой выборки n случайных величин из генеральной совокупности с распределением $P(X)$.

Гипотеза 2. Алгоритм классификации выбирается из **фиксированного заранее** семейства F, $F \subset \{X \to Y\}$.

Математическая и информационная модели

Физическая модель

...

Математическая модель

описывает объект с помощью уравнений, соотношений, закономерностей и т.д., моделирующих реальные физические процессы.

Информационная модель

представляет объект в терминах формальных зависимостей «вход-выход», обычно никак не связанных с указанными процессами и обеспечивающим только правильное нахождение реакции объекта на данный входной сигнал («чёрный ящик» в кибернетике).

Критика VC-теории

- VC-теория даёт чрезвычайно заниженные оценки надёжности, что является причиной постоянных попыток её модификации.
- **Гипотезу 1** не удаётся полностью эллиминировать (иногда заменяется на более слабые утверждения).
- Результатом принятия **Гипотезы 2** становится необходимость решать непростые, а подчас и неразрешимые вопросы определения различных характеристик класса F: ёмкость, функция роста, коэффициент разнообразия и др. (их нахождение «*сводится к громоздким комбинаторным вычислениям, которые не всегда можно провести*»).

Критика VC-теории (продолжение)

- **Гипотеза 2** часто не соответствует реальному процессу построения алгоритмов (тип классификатора определяется непосредственно в процессе решения задачи и семейство F заранее не фиксируется).
- Одна и та же дискриминантная функция на $\mathcal{X}$ может быть получена реализацией различных алгоритмов, принадлежащих разным классам, что приведёт, вообще говоря, к разным оценкам надёжности осуществляемой ей классификации.
- Можно сначала определить оптимальный алгоритм A^*, а затем заново формально решить задачу классификации, полагая $F = \{A^*\}$ и $|F| = 1$.
- ...

Место VC-теории: лирическое отступление

VC-теория — это ...

теория равномерной сходимости частот к вероятностям и её применение к задаче оценки надёжности классификаторов в общей постановке неоправдано.

VC-теория может успешно применяться

- при адекватности **Гипотезы 2** процессу построения и/или оценки надёжности классификаторов, например:
  - классификатор действительно выбирается из заранее фиксированного семейства F,
  - оценка определяется по обучающей выборке (скользящий контроль), ...
- при теоретико-статистических исследованиях;
- при теоретико-функциональных исследованиях и т.д.

Место VC-теории: лирическое отступление

VC-теория — это ...

теория равномерной сходимости частот к вероятностям и её применение к задаче оценки надёжности классификаторов в общей постановке неоправдано.

VC-теория может успешно применяться

- при адекватности **Гипотезы 2** процессу построения и/или оценки надёжности классификаторов, например:
  - классификатор действительно выбирается из заранее фиксированного семейства F,
  - оценка определяется по обучающей выборке (скользящий контроль), ...
- при теоретико-статистических исследованиях;
- при теоретико-функциональных исследованиях и т.д.

Место VC-теории: лирическое отступление

VC-теория — это ...

теория равномерной сходимости частот к вероятностям и её применение к задаче оценки надёжности классификаторов в общей постановке неоправдано.

VC-теория может успешно применяться

- при адекватности **Гипотезы 2** процессу построения и/или оценки надёжности классификаторов, например:
  - классификатор действительно выбирается из заранее фиксированного семейства F,
  - оценка определяется по обучающей выборке (скользящий контроль), ...
- при теоретико-статистических исследованиях;
- при теоретико-функциональных исследованиях и т.д.

Место VC-теории: лирическое отступление

VC-теория — это ...

теория равномерной сходимости частот к вероятностям и её применение к задаче оценки надёжности классификаторов в общей постановке неоправдано.

VC-теория может успешно применяться

- при адекватности **Гипотезы 2** процессу построения и/или оценки надёжности классификаторов, например:
  - классификатор действительно выбирается из заранее фиксированного семейства F,
  - оценка определяется по обучающей выборке (скользящий контроль), ...
- при теоретико-статистических исследованиях;
- при теоретико-функциональных исследованиях и т.д.

Место VC-теории: лирическое отступление

VC-теория — это ...

теория равномерной сходимости частот к вероятностям и её применение к задаче оценки надёжности классификаторов в общей постановке неоправдано.

VC-теория может успешно применяться

- при адекватности **Гипотезы 2** процессу построения и/или оценки надёжности классификаторов, например:
  - классификатор действительно выбирается из заранее фиксированного семейства F,
  - оценка определяется по обучающей выборке (скользящий контроль), ...
- при теоретико-статистических исследованиях;
- при теоретико-функциональных исследованиях и т.д.

Постановка задачи: отказ от VC-теории

Оценка надёжности: данные

Единственными данными для оценки надёжности классифицирующего алгоритма является информация о том, что на *контрольной выборке* длины n алгоритм совершил m ошибок.

Задача в такой постановке имеет решение методами математической статистики, опирающимися на классическую информационную модель алгоритма классификации.

Постановка задачи: отказ от VC-теории

Оценка надёжности: данные

Единственными данными для оценки надёжности классифицирующего алгоритма является информация о том, что на *контрольной выборке* длины n алгоритм совершил m ошибок.

Задача в такой постановке имеет решение методами математической статистики, опирающимися на классическую информационную модель алгоритма классификации.

Постановка задачи: отказ от VC-теории

Оценка надёжности: данные

Единственными данными для оценки надёжности классифицирующего алгоритма является информация о том, что на *контрольной выборке* длины n алгоритм совершил m ошибок.

Задача в такой постановке имеет решение методами математической статистики, опирающимися на классическую информационную модель алгоритма классификации.

Классическая информационная модель алгоритма классификации

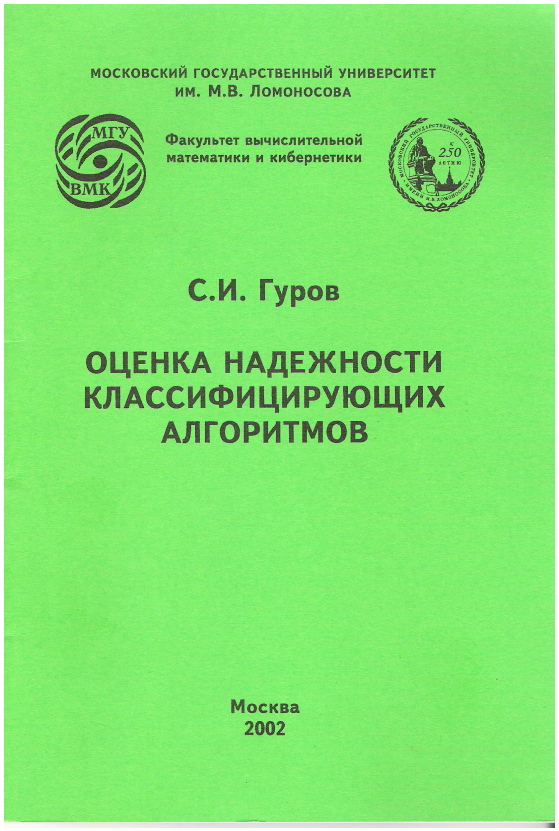
$\underline{P(A) = r}$

1) получение описания объекта $x \in \mathcal{X}$;
2) получение значения r случайной величины, равномерно распределённой на отрезке $[0, 1]$;
3) сравнение r с внутренним параметром $P \in [0, 1]$ алгоритма;
4) выдача правильного ответа о принадлежности x, если $r \leqslant P$, и неправильного — в противном случае.

— классическая урновая схема Бернулли.

Что даёт классический подход?

Оценки вероятности $P(A)$, полученные в рамках классических статистических моделей без привлечения дополнительных предположений, приведены в



Даны результаты для частотного и бейесовского подходов в случаях двух и более классов

Критика классического подхода

- В силу своей универсальности схема Бернулли является слишком грубой формализацией процесса классификации, осуществляемой алгоритмом A, не учитывающей его специфику.
- В частности, неучёт факта построения (настройки) A по обучающейся последовательности, в силу чего больши́е (но не слишком близкие к 1) значения искомого параметра P более вероятны, чем малые, и более того, $P \geqslant \frac{1}{2}$.
- Часто принимают ряд дополнительных предположений (например, о типах распределений), в результате чего для оценки P требуются трудноопределимые или даже ненаблюдаемые параметры.

Предлагаемая информационная модель алгоритма классификации базируется на следующих предположениях:

(А) информация о значениях n и m является адекватной для оценки вероятности P;

(Б) каждый классифицирующий алгоритм $A(t)$ характеризуется значением некоторого скрытого неслучайного, но неизвестного параметра $t \in [0, 1]$, определяющим его степень «необученности».

Условие (А) будет выполняться, например, при справедливости **Гипотезы 1** (независимость обучающей и экзаменационной выборок).

Условие (Б) — конкретизация информационной модели *IM*

$\underline{P(A(t)) = r}$

1) получение описания объекта $x \in \mathcal{X}$;
2) получение значения r случайной величины, равномерно распределённой на отрезке $[0, 1]$
(шаги 1) и 2) совпадают с классической моделью);
3) сравнение r с внутренним параметром $t \in [0, 1]$ алгоритма;
4) в случае $t \leqslant r$ — выдача правильного ответа о принадлежности x, иначе — выдача правильного ответа с вероятностью p.

— **двухуровневое** испытание по урновой схеме Бернулли.

Параметр p

Вследствие обучения алгоритма A, на последнем шаге работы модели IM значение вероятности p правильного ответа будет не менее $\frac{1}{2}$ (этот минимум реализуется при условии равновероятного появления объектов из каждого класса).

Точное значение p, очевидно, неопределимо, однако всегда можно положить $p = \frac{1}{2}$, корректируя соответствующим образом параметр t, и получая, возможно, заниженные оценки $P(A(t))$.

Оценка $\widehat{P}(A(t))$ через определение t

t — мера (пере)обученности алгоритма

$$P(A(t)) = 1 - t + tp \geqslant 1 - \frac{t}{2}, \qquad 0 \leqslant \frac{m}{n} \leqslant t \leqslant 1$$

Теперь задача — оценить значение t.

Сделаем это, построив на основе приведённой информационной модели алгоритма две его статистические модели: дискретную и непрерывную.

Статистические модели для t

Дискретная (DIM): $[(1-t)n]$ объектов алгоритм отклассифицировал правильно по построению, а классификация остальных — результат случайного угадывания с вероятностью успеха p.

Непрерывная (CIM): классификация каждого объекта происходит строго по описанной выше модели *IM*.

CIM и DIM соответствуют осреднению по времени и по совокупности соответственно в рамках эргодической теории случайных процессов

Крайние значения t означают, что:

$t = 0$ — удалось построить алгоритм, всегда правильно осуществляющий классификацию любого предъявляемого объекта;

$t = 1$ — построенный алгоритм выдаёт случайный, с вероятностью $P = p$ — правильный, ответ о принадлежности объекта.

Реальный алгоритм описывается некоторым значением t, лежащим между этими крайними идеальными случаями.

Методы определения параметра t

Частотный подход: максимум правдоподобия;

Бейесовский подход: по апостериорному распределению —

- максимум апостериорной вероятности;
- математическое ожидание;
- медиана;
- интервальные оценки;
- ...

Функция правдоподобия $L(k)$

Полагаем $t = \frac{k}{n}$, $0 \leqslant m \leqslant k \leqslant n$ $\quad k = ?$

Функция правдоподобия —

$$L(k;\, n, m) \;=\; C_k^m p^{n-m}(1-p)^{m-n+k} \;=\; \frac{C_k^m}{2^k}.$$

$$\arg\max_k \{L(k;\, n, m)\} \;=\; \{\, 2m-1,\; 2m \,\}$$

Оценка по максимуму правдоподобия:

$$\widehat{P}_{ML}(A) \geqslant 1 - \frac{m}{n} \;=\; \widehat{P}_{Cl}(A)\,.$$

Отличие $\widehat{P}_{ML}$ от классической оценки $\widehat{P}_{Cl}$ в знаке: здесь стоит $\geqslant$, а не $=$, показывая заниженность классической точечной оценки $\widehat{P}_{Cl}$.

Функция правдоподобия $L(k)$

Полагаем $t = \frac{k}{n}$, $0 \leqslant m \leqslant k \leqslant n$ $\quad k = ?$

Функция правдоподобия —

$$L(k;\, n, m) \;=\; C_k^m p^{n-m}(1-p)^{m-n+k} = \frac{C_k^m}{2^k}.$$

$$\arg\max_k \{L(k;\, n, m)\} \;=\; \{\, 2m-1,\; 2m \,\}$$

Оценка по максимуму правдоподобия:

$$\widehat{P}_{ML}(A) \geqslant 1 - \frac{m}{n} \;=\; \widehat{P}_{Cl}(A)\,.$$

Отличие $\widehat{P}_{ML}$ от классической оценки $\widehat{P}_{Cl}$ в знаке: здесь стоит $\geqslant$, а не $=$, показывая заниженность классической точечной оценки $\widehat{P}_{Cl}$.

Достоверность оценок

— определяется рамках бейесовского подхода для равномерного априорного распределения значения k:

$$\eta = \eta(M; n, m) = \frac{\sum_{k=m}^{M} \frac{C_k^m}{2^k}}{\sum_{k=m}^{n} \frac{C_k^m}{2^k}}.$$

По этой формуле можно определить некоторые значения достоверности η оценки

$$\widehat{P}(A) = 1 - \frac{M}{2n}$$

Достоверность оценок (продолжение)

Имеем $\eta(2;n,1)=\frac{1}{2-(n+2)/2^n}$, т.е. при $m=1$ оценка $\widehat{P}_{ML}(A)$ для не слишком малых n является медианной.

Оценки $P(A)$ при $n=10,\ m=1$

Модель *IM*	**Классическая модель**
$0,90 \leqslant P(A)$ $\eta = 0,503$	$0,85 \leqslant P(A)$ $\eta = 0,500$
$0,75 \leqslant P(A)$ $\eta = 0,895$	$0,66 \leqslant P(A) \leqslant 0,99$ $\eta = 0,900$
$0,7 \leqslant P(A)$ $\eta = 0,943$	$0,61 \leqslant P(A) \leqslant 0,99$ $\eta = 0,950$

Случай 0-события ($m = 0$)

Формально имеем $L(k;\, n, 0) = 2^{-k}$ и

$$\eta \;=\; \eta(M; n, 0) \;=\; \frac{\sum_{k=0}^{M} 2^{-k}}{\sum_{k=0}^{n} 2^{-k}}\,.$$

Отсюда, например, для $n = 10$ и $m = 0$ получим

$$0,80 \leqslant P(A) \;\text{ с достоверностью }\; \eta \geqslant 0,937 \;\; (M = 4),$$
$$0,75 \leqslant P(A) \;\text{ с достоверностью }\; \eta \geqslant 0,968 \;\; (M = 5),$$
$$0,70 \leqslant P(A) \;\text{ с достоверностью }\; \eta \geqslant 0,984 \;\; (M = 6).$$

Функция правдоподобия $L(t)$

$$L(t;\, n, m) \;=\; C_n^m \left(1 - \frac{t}{2}\right)^{n-m} \left(\frac{t}{2}\right)^m, \quad 1 < \frac{m}{n} \leqslant t \leqslant 1$$

$$\arg\max_t \{L(t;\, n, m)\} \;=\; \frac{2m}{n}$$

Совпадение оценки с аналогичной в DIM говорит об «эргодичности» рассматриваемой информационной модели.
Несложный анализ показывает, что CIM — предельный случай DIM при $n \to \infty$.
Выбор той или иной модели определяется лишь их адекватностью реальному процессу распознавания в конкретной задаче.

Бейесовский подход: интервальные оценки

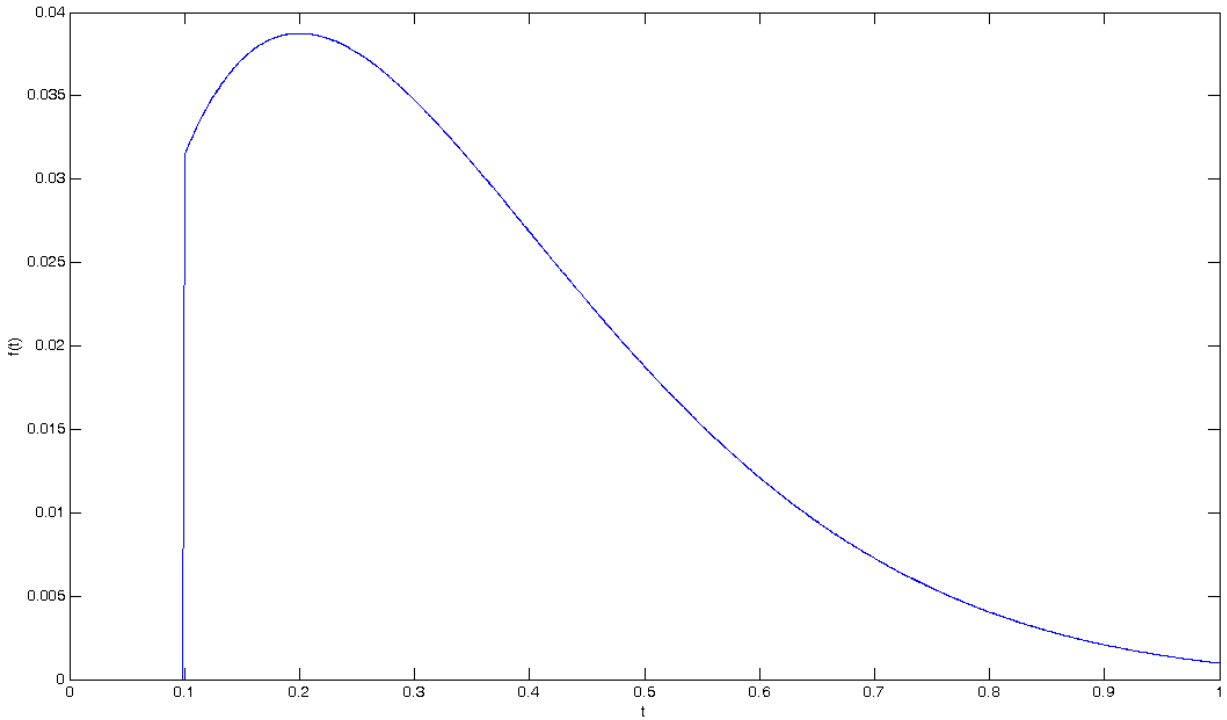
Для равномерного f_{a_prior} имеем

$$f_{a\_post}(t;\, n, m) \;=\; \frac{\left(1-\frac{t}{2}\right)^{n-m}\left(\frac{t}{2}\right)^{m}}{M(n,\, m)}\,,$$

$$\frac{m}{n} \leqslant t \leqslant 1\,,$$

где $M(n,\, m) \;=\; \int\limits_{\frac{m}{n}}^{1} \left(1-\frac{t}{2}\right)^{n-m}\left(\frac{t}{2}\right)^{m} dt$ — нормирующий множитель.

График функции $f(t) = (1 - \frac{t}{2})^9 \frac{t}{2} \propto f_{a\_post}(t;\, 10, 1)$



По формуле для $f_{a\_post}(t;\, n, m)$ возможно получение оценок $\widehat{P}$ для произвольных значений η, более точных чем классические.

Сравнение оценок

Некоторые интервальные бейесовские оценки (P^-, P^+), полученные по CIM с равномерным априорное распределением t в сравнении с классическими:

$\eta = 0,90$ $\quad$ $\mathbf{n = 10,\ m =}$		**0**	**1**	**2**
классические	P^-	0,7410	0,6060	0,4930
	P^+	1,0000	0,9950	0,9630
по CID (бейесовские)	P^-	0,8114	0,6856	0,6012
	P^+	1,0000	0,9500	0,9000

$\eta = 0,90$ $\quad$ $\mathbf{n = 20,\ m =}$		**0**	**1**	**2**
классические	P^-	0,8610	0,7840	0,7170
	P^+	1,0000	0,9970	0,9820
по CID (бейесовские)	P^-	0,8961	0,8220	0,7611
	P^+	1,0000	0,9750	0,9500

Сравнение оценок... Развитие модели

$\eta = 0,90$ $\mathbf{n = 100,\ m =}$		**0**	**1**	**10**
классические	P^-	0,9700	0,9617	0,8368
	P^+	1,0000	0,9974	0,9447
по CID (бейесовские)	P^-	0,9775	0,9609	0,8513
	P^+	1,0000	0,9950	0,9500

Рассматривая выборку прецедентов только из одного данного классов, можно определить ошибки первого и второго родов.

Очевидным способом подход переносится на случай многих классов.

Выводы

- Предложенная информационная модель более адекватно описывает работу классифицирующего алгоритма, чем классическая.
- Модель не использует **Гипотезу 2**, являющуюся наиболее слабым местом VC-теории.
- На основе введённой модели возможно получение более точных, чем классические, оценок надёжности классификации.
- Интересным направлением дальнейших исследований является сравнение новых оценок с известными, получаемыми на основе различных подходов, не использующих **Гипотезу 2** и адаптация указанных подходов к введённой модели.

Выводы

- Предложенная информационная модель более адекватно описывает работу классифицирующего алгоритма, чем классическая.
- Модель не использует **Гипотезу 2**, являющуюся наиболее слабым местом VC-теории.
- На основе введённой модели возможно получение более точных, чем классические, оценок надёжности классификации.
- Интересным направлением дальнейших исследований является сравнение новых оценок с известными, получаемыми на основе различных подходов, не использующих **Гипотезу 2** и адаптация указанных подходов к введённой модели.

Выводы

- Предложенная информационная модель более адекватно описывает работу классифицирующего алгоритма, чем классическая.
- Модель не использует **Гипотезу 2**, являющуюся наиболее слабым местом VC-теории.
- На основе введённой модели возможно получение более точных, чем классические, оценок надёжности классификации.
- Интересным направлением дальнейших исследований является сравнение новых оценок с известными, получаемыми на основе различных подходов, не использующих **Гипотезу 2** и адаптация указанных подходов к введённой модели.

Выводы

- Предложенная информационная модель более адекватно описывает работу классифицирующего алгоритма, чем классическая.
- Модель не использует **Гипотезу 2**, являющуюся наиболее слабым местом VC-теории.
- На основе введённой модели возможно получение более точных, чем классические, оценок надёжности классификации.
- Интересным направлением дальнейших исследований является сравнение новых оценок с известными, получаемыми на основе различных подходов, не использующих **Гипотезу 2** и адаптация указанных подходов к введённой модели.

Выводы

- Предложенная информационная модель более адекватно описывает работу классифицирующего алгоритма, чем классическая.
- Модель не использует **Гипотезу 2**, являющуюся наиболее слабым местом VC-теории.
- На основе введённой модели возможно получение более точных, чем классические, оценок надёжности классификации.
- Интересным направлением дальнейших исследований является сравнение новых оценок с известными, получаемыми на основе различных подходов, не использующих **Гипотезу 2** и адаптация указанных подходов к введённой модели.

Спасибо за внимание

Вопросы?