

Прикладные байесовские методы

Максим Кочуров

МГУ им. М.В. Ломоносова

Лекция 1



Содержание

- ① Обо мне
- ② Формальности
- ③ Мотивация
- ④ Теорема Байеса
 - Априорное распределение
 - Правдоподобие
 - Апостериорное распределение
- ⑤ Модель
- ⑥ Обсуждение
- ⑦ Дополнительные материалы
 - Программирование



Обо мне

- Выпускник
 - бакалавриат – ЭФ МГУ (2018)
 - магистратура - Skoltech DS (2020)
- Один из разработчиков PyMC
- Ведущий data сайентист в PyMC Labs
- Выпускник BayesGroup
- Имею опыт работы с:
 - машинным обучением и компьютерным зрением;
 - дифференциальной геометрией для графовых нейросетей;
 - байесовскими методами для глубокого обучения;
 - прикладной байесовской статистикой в индустрии (A/B тестирование, биоинформатика).



Skoltech



p(B|A)yesgroup.ru





В этом курсе

Вы узнаете...

- как критически думать о вашей модели;
- какие инструменты использовать для проверки достоверности результатов;
- как презентовать ваши результаты;
- как строить непараметрические модели временных рядов.



Оценки

Оценка состоит из:

- 60% домашние задания
- 40% групповой проект

Конвертация баллов в оценку

- 5 - 85%+
- 4 - 65%+
- 3 - 40%+
- 2 - < 40%



- Использование в продвинутых исследованиях
 - ЦБ РФ - байесовская модель DSGE ([ссылка](#))
 - В Google Scholar **очень много** ([ссылка](#)) статей, использующих PyMC
- Использование в индустрии
 - Маркетинг в Indigo ([ссылка](#))
 - Разработка медикаментов в Roche ([ссылка](#))
 - Портфельная теория в Quantopian ([ссылка](#))
 - Финансовые консультации в EverySk ([ссылка](#))
 - Проведение опросов в Civiqs ([ссылка](#))
- Возможности карьерного роста
 - Связывают многие дисциплины и карьеры
 - Интересные предложения работы в индустрии
 - Открывают новые возможности для исследований

Теорема Байеса



$$p(\Theta|D) = \frac{\overbrace{p(D|\Theta)}^{\text{FACT}} \overbrace{p(\Theta)}^{\text{The Thinker}}}{p(D)}$$



FACT



D = Данные

Θ = Состояние мира



Априорное распределение



$$p(\Theta|\mathcal{D}) = \frac{p(\mathcal{D}|\Theta) \overbrace{p(\Theta)}^{\text{prior}}}{p(\mathcal{D})}$$



Авторы: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Пример из байесовской эконометрики.

Исследовательский вопрос

Нужно ли усиливать контроль за потреблением подростками марихуаны?

- Долгосрочное влияние наркотиков на мозговую активность в случае употребления в раннем возрасте



Авторы: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Пример из байесовской эконометрики.

Исследовательский вопрос

Нужно ли усиливать контроль за потреблением подростками марихуаны?

- Долгосрочное влияние наркотиков на мозговую активность в случае употребления в раннем возрасте
- Почти нет существующих исследований по теме
 - нет исследований именно о связи употребления марихуаны и мозговой активности
 - развитие сопутствующих заболеваний



Авторы: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Пример из байесовской эконометрики.

Исследовательский вопрос

Нужно ли усиливать контроль за потреблением подростками марихуаны?

- Долгосрочное влияние наркотиков на мозговую активность в случае употребления в раннем возрасте
- Почти нет существующих исследований по теме
 - нет исследований именно о связи употребления марихуаны и мозговой активности
 - развитие сопутствующих заболеваний
- Нехватка данных



Авторы: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Пример из байесовской эконометрики.

Исследовательский вопрос

Нужно ли усиливать контроль за потреблением подростками марихуаны?

- Долгосрочное влияние наркотиков на мозговую активность в случае употребления в раннем возрасте
- Почти нет существующих исследований по теме
 - нет исследований именно о связи употребления марихуаны и мозговой активности
 - развитие сопутствующих заболеваний
- Нехватка данных
- Классическая эконометрика не работает (16 наблюдений в группе)



Авторы: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Пример из байесовской эконометрики.

Исследовательский вопрос

Нужно ли усиливать контроль за потреблением подростками марихуаны?

- Долгосрочное влияние наркотиков на мозговую активность в случае употребления в раннем возрасте
- Почти нет существующих исследований по теме
 - нет исследований именно о связи употребления марихуаны и мозговой активности
 - развитие сопутствующих заболеваний
- Нехватка данных
- Классическая эконометрика не работает (16 наблюдений в группе)
- Хорошо бы учесть мнение экспертов



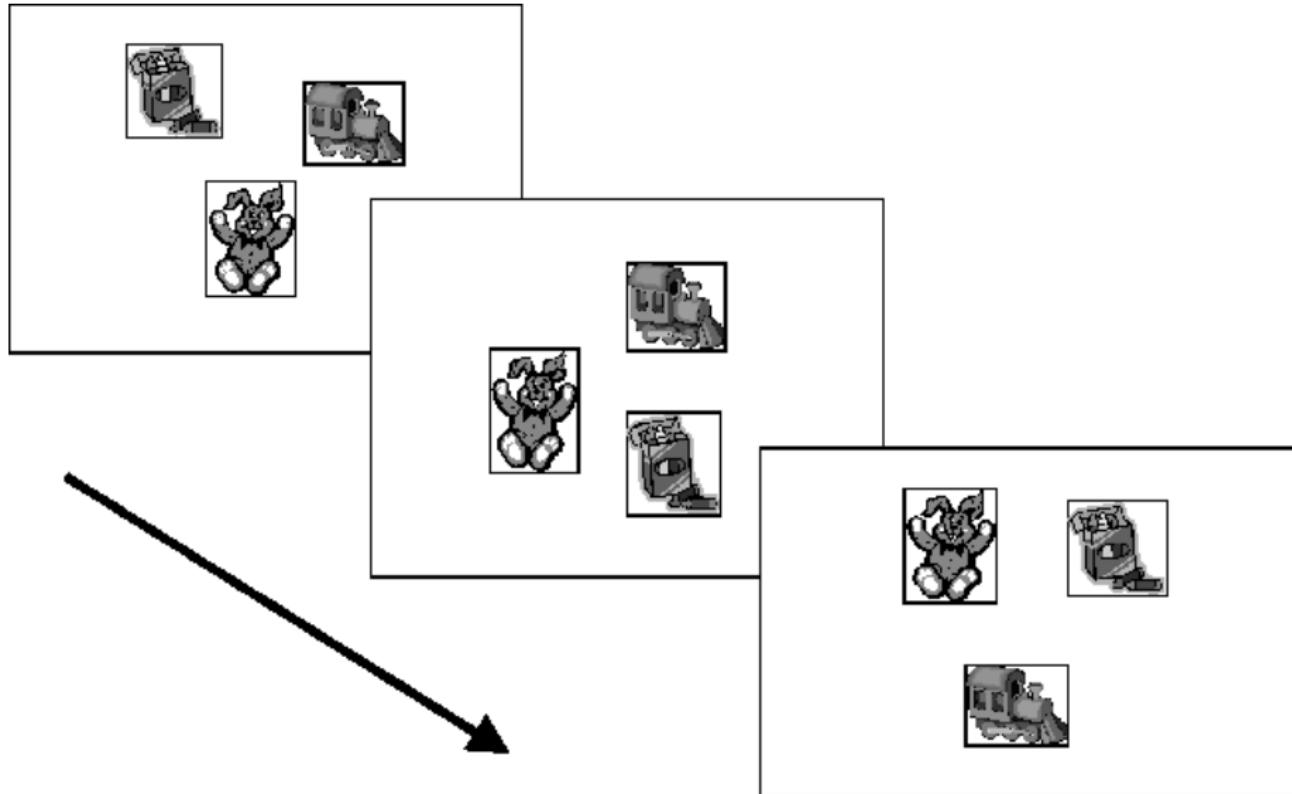
Авторы: Marielle Zondervan-Zwijnenburg, Margot Peeters, Sarah Depaoli, Rens van de Schoot [3]. Пример из байесовской эконометрики.

Исследовательский вопрос

Нужно ли усиливать контроль за потреблением подростками марихуаны?

- Долгосрочное влияние наркотиков на мозговую активность в случае употребления в раннем возрасте
- Почти нет существующих исследований по теме
 - нет исследований именно о связи употребления марихуаны и мозговой активности
 - развитие сопутствующих заболеваний
- Нехватка данных
- Классическая эконометрика не работает (16 наблюдений в группе)
- Хорошо бы учесть мнение экспертов
- Для принятия информированного решения необходимы

Игра





Кратко о статье

Измерение наличия и степени употребления марихуаны

- Развитие мозга можно измерить с помощью игры (Self Ordered Pointing Test [1])
- Участники проверяются на потребление марихуаны
- Подростки проходят игру 2 раза в год
- Результаты сравниваются между людьми, употребляющими каннабис **часто и редко**

Оставшиеся вопросы

- Как объём употреблённой марихуаны влияет на развитие мозга?
- Люди какого возраста менее подвержены отрицательным эффектам?
- Какой политики следует придерживаться для минимизации отрицательных эффектов?



Кейс: априорное распределение

Для разработки априорного распределения исследователи совместили много источников информации

① Знания имевшиеся до сбора данных



② Результаты прошлых исследований:



③ Экспертное мнение



④ Ограничения

Максим Кочуров



Кейс: априорное распределение

Для разработки априорного распределения исследователи совместили много источников информации

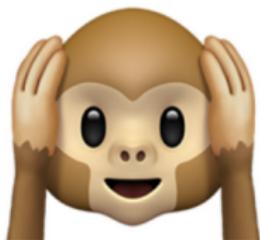
① Знания имевшиеся до сбора данных

- Практический диапазон для метрики SOPT и темпов роста

② Результаты прошлых исследований:



③ Экспертное мнение



④ Ограничения

Максим Кочуров





Кейс: априорное распределение

Для разработки априорного распределения исследователи совместили много источников информации

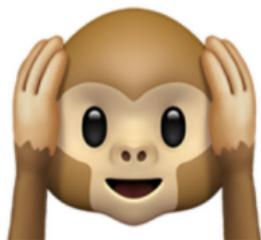
① Знания имевшиеся до сбора данных

- Практический диапазон для метрики SOPT и темпов роста

② Результаты прошлых исследований:

- размер эффекта от употребления марихуаны был смешан с другими заболеваниями, или о нём не было данных
- "в 13 из 693 статей была полезная информация"

③ Экспертное мнение



④ Ограничения

Максим Кочуров



Кейс: априорное распределение

Для разработки априорного распределения исследователи совместили много источников информации

① Знания имевшиеся до сбора данных

- Практический диапазон для метрики SOPT и темпов роста

② Результаты прошлых исследований:

- размер эффекта от употребления марихуаны был смешан с другими заболеваниями, или о нём не было данных
- "в 13 из 693 статей была полезная информация"

③ Экспертное мнение

- Прошлые исследования были изучены исследователями
- Связь заболеваний и поведения с употреблением марихуаны

④ Ограничения

Максим Кочуров





Кейс: априорное распределение

Для разработки априорного распределения исследователи совместили много источников информации

① Знания имевшиеся до сбора данных

- Практический диапазон для метрики SOPT и темпов роста

② Результаты прошлых исследований:

- размер эффекта от употребления марихуаны был смешан с другими заболеваниями, или о нём не было данных
- "в 13 из 693 статей была полезная информация"

③ Экспертное мнение

- Прошлые исследования были изучены исследователями
- Связь заболеваний и поведения с употреблением марихуаны

④ Ограничения

Максим Кочуров





Кейс: априорное распределение

Для разработки априорного распределения исследователи совместили много источников информации

① Знания имевшиеся до сбора данных

- Практический диапазон для метрики SOPT и темпов роста

② Результаты прошлых исследований:

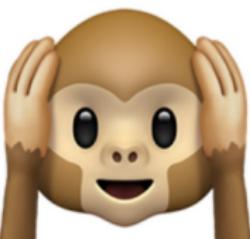
- размер эффекта от употребления марихуаны был смешан с другими заболеваниями, или о нём не было данных
- "в 13 из 693 статей была полезная информация"

③ Экспертное мнение

- Прошлые исследования были изучены исследователями
- Связь заболеваний и поведения с употреблением марихуаны

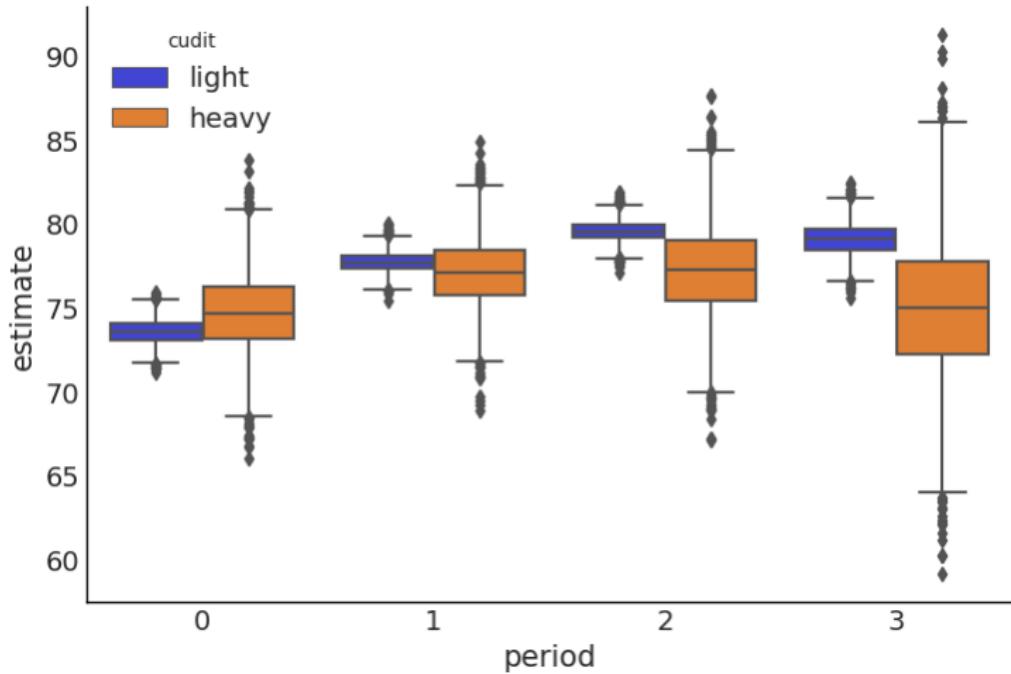
④ Ограничения

Максим Кочуров





Результаты





Частые проблемы

Авторы исследования обосновали выбор априорного распределения

- ① Априорное распределение субъективно
- ② Спецификация априорного распределения непонятна
- ③ Априорное распределение неверно специфицировано



Частые проблемы

Авторы исследования обосновали выбор априорного распределения

- ① Априорное распределение субъективно
 - Сравнение информированного априорного распределения с неинформированным
- ② Спецификация априорного распределения непонятна
- ③ Априорное распределение неверно специфицировано



Частые проблемы

Авторы исследования обосновали выбор априорного распределения

- ① Априорное распределение субъективно
 - Сравнение информированного априорного распределения с неинформированным
- ② Спецификация априорного распределения непонятна
 - Принципы выбора априорного распределения описаны в разделе Logbook: <https://osf.io/aw8fy/>
- ③ Априорное распределение неверно специфицировано



Авторы исследования обосновали выбор априорного распределения

- 1 Априорное распределение субъективно**
 - Сравнение информированного априорного распределения с неинформированным
- 2 Спецификация априорного распределения непонятна**
 - Принципы выбора априорного распределения описаны в разделе Logbook: <https://osf.io/aw8fy/>
- 3 Априорное распределение неверно специфицировано**
 - Анализ небезупречен, но это выходит за рамки лекции



FACT

$$p(\Theta | \mathcal{D}) = \frac{\overbrace{p(\mathcal{D} | \Theta)}^{\text{FACT}} p(\Theta)}{p(\mathcal{D})}$$



Кейс: А/Б тест

Вы продаёте орехи. Вы хотите продать больше орехов! Как увеличить объёмы продаж?

- увеличить вероятность покупки



- увеличить размер заказа

ТОВ "АТБ-Маркет"	
МАГАЗИН "ПРОДУКТИ-556"	
м.Чернігівська, 56а років СРСР, 5	
АТБ - ЦЕНР РАСПРОДІЛ!	
емл: atbmarket.com	
нн 394872184175	
00005	кося Я. І.
Касса 5	
чек 861621.	з 941
Пакет п/е	40х60 з1
Святковий сим	1,45 А
СУММА	1.45
ПДВ А	28,00%
СУММА ПОДАТКУ	0,24
ГОДІВКА	2,00
РЕБЕТА	0,55
ДЯКУЮ ЗА ПОКУПКУ!	
ГАРЯЧА ЛІНІЯ 8 000 500 415	
1 АРТИКУЛ	
0242196 0509204 17-01-2018 14:19:28	
ЗН.ИП00001005 04 2801001754	
ФІСКАЛЬНИЙ ЧЕК	
© Екселло	





Кейс: А/Б тест

Вы продаёте орехи. Вы хотите продать больше орехов! Как увеличить объёмы продаж?

- увеличить вероятность покупки
 - Сделать баннер о здоровой еде
 - Добавить баннер с рецептами
 - Улучшить макет сайта
- увеличить размер заказа





Кейс: А/Б тест

Вы продаёте орехи. Вы хотите продать больше орехов! Как увеличить объёмы продаж?

- увеличить вероятность покупки
 - Сделать баннер о здоровой еде
 - Добавить баннер с рецептами
 - Улучшить макет сайта
- увеличить размер заказа
 - Снизить цену
 - Повысить качество
 - Улучшить упаковку





Кейс: А/Б тест

Вы продаёте орехи. Вы хотите продать больше орехов! Как увеличить объёмы продаж?

- увеличить вероятность покупки
 - Сделать баннер о здоровой еде
 - Добавить баннер с рецептами
 - Улучшить макет сайта
- увеличить размер заказа
 - Снизить цену
 - Повысить качество
 - Улучшить упаковку

Что лучше?

А/Б тестирование может ответить

Максим Кочуров





- Значительная часть данных – просто нули
- Классический t-тест для 2-х выборок предполагает нормальное распределение, это не наш случай
- Исследователи признают недостатки t-теста в таких случаях[2]



Не все потребители покупают орехи



- Значительная часть данных – просто нули
- Классический t-тест для 2-х выборок предполагает нормальное распределение, это не наш случай
- Исследователи признают недостатки t-теста в таких случаях[2]

Решение

Придумаем правдоподобие, не основанное на нормальном распределении



Zero Inflation (избыток нулей)

Значительная доля наблюдений в точности равна нулю

1	0	1	0	0	0	0	1	1	1
0	1	0	0	1	0	0	1	1	0
0	1	0	1	1	1	0	0	0	1
0	0	1	0	0	0	1	1	1	1
0	1	1	0	1	1	0	0	1	0
1	0	1	0	0	1	0	1	0	1
1	0	1	1	1	1	0	1	1	1
0	0	0	0	0	1	0	0	1	1
0	0	0	1	0	0	1	1	0	0
0	1	0	0	1	1	1	0	1	0

Примеры:

- Время ожидания в очереди (нет очереди – ноль)
- Дефекты на производстве (нет дефектов – ноль)
- Уровень осадков (нет осадков – ноль)
- Покупки (нет орехов – ноль)



Мудрость

В любом распределении можно повысить вероятность возникновения нулей

Пример: Zero Inflated Gamma. Параметры α, β гамма распределения и p - вероятность ненулевого наблюдения

Сэмплирование:

$$z \sim \text{Bernoulli}(p)$$

$$\text{sample} \sim \begin{cases} \text{Gamma}(\alpha, \beta) & , z = 1 \\ 0 & , z = 0 \end{cases}$$



Мудрость

В любом распределении можно повысить вероятность возникновения нулей

Пример: Zero Inflated Gamma. Параметры α, β гамма распределения и p - вероятность ненулевого наблюдения

Сэмплирование:

$$z \sim \text{Bernoulli}(p)$$

$$\text{sample} \sim \begin{cases} \text{Gamma}(\alpha, \beta) & , z = 1 \\ 0 & , z = 0 \end{cases}$$

Логарифм плотности

$$\log p(x | p, \alpha, \beta) = \begin{cases} \log(1 - p) & , x = 0 \\ \log(p) + \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)} & , x > 0 \end{cases}$$



Мудрость

В любом распределении можно повысить вероятность возникновения нулей

Пример: Zero Inflated Gamma. Параметры α, β гамма распределения и p - вероятность ненулевого наблюдения

Сэмплирование:

$$z \sim \text{Bernoulli}(p)$$

$$\text{sample} \sim \begin{cases} \text{Gamma}(\alpha, \beta) & , z = 1 \\ 0 & , z = 0 \end{cases}$$

Логарифм плотности

$$\log p(x | p, \alpha, \beta) = \begin{cases} \log(1 - p) & , x = 0 \\ \log(p) + \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)} & , x > 0 \end{cases}$$



Zero inflation как смесь

Мудрость

Zero inflated распределения – разновидность смешанных распределений

Компоненты смеси:

- ① Constant(0)
- ② Gamma(α, β)

Параметр "смешанности" p (здесь – доля гамма распределения)

$$\text{ZI-Gamma}(p, \alpha, \beta) \equiv \text{Mixture}([1 - p, p], [\text{Constant}(0), \text{Gamma}(\alpha, \beta)])$$



Вернёмся к примеру

заказ сделан $\sim \text{Bernoulli}(p)$

размер заказа $\sim \begin{cases} \text{Gamma}(\alpha, \beta) & , \text{заказ сделан} = 1 \\ 0 & , \text{заказ сделан} = 0 \end{cases}$





Выводы

- Хорошее правдоподобие помогает лучше понять проблему
 - разделили вероятность покупки и размер заказа
 - больше возможностей по сравнению с классическим t-тестом
- Понимание проблемы – первый шаг к хорошему правдоподобию



$$\underbrace{p(\Theta | \mathcal{D})}_{\text{BREAKING NEWS}} = \frac{p(\mathcal{D} | \Theta)p(\Theta)}{p(\mathcal{D})}$$

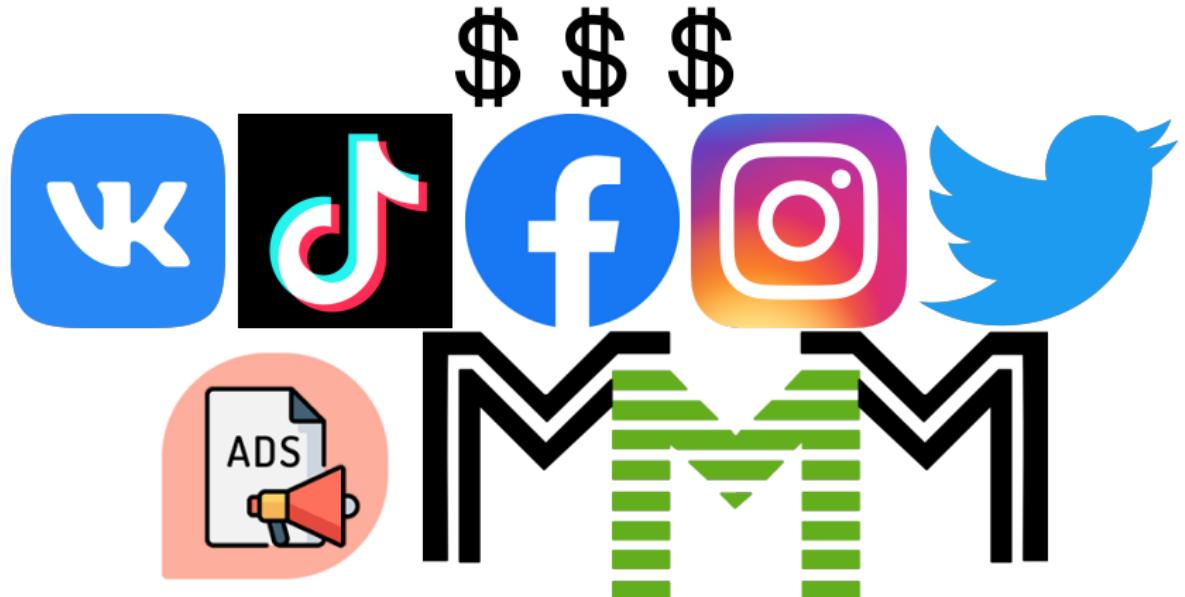




$p(\text{что вы думаете} | \text{данные})$

$\propto p(\text{данные} | \text{что вы думаете})p(\text{что вы думаете})$





Media Mix Model

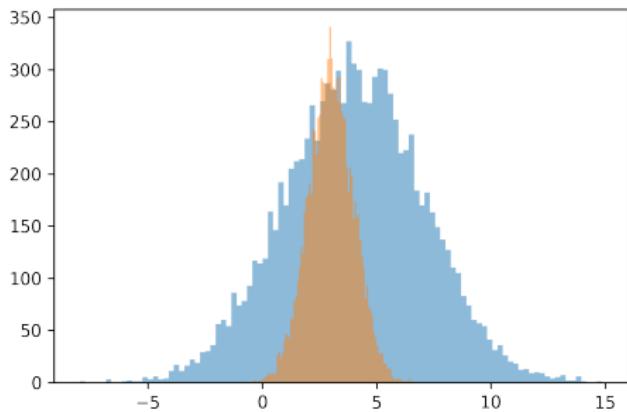
МММ – помогает оценивать маркетинговые каналы по историческим данным



Неопределённость

Media Mix Models ([подробнее](#)
[здесь](#))

- Насколько ценные дополнительные инвестиции в \$1000 для ВК? или Яндекса?
- Насколько велика неопределённость оценок?
- Большая ценность, но большая неопределённость, или небольшая ценность небольшая неопределённость?
- На что выделить деньги?

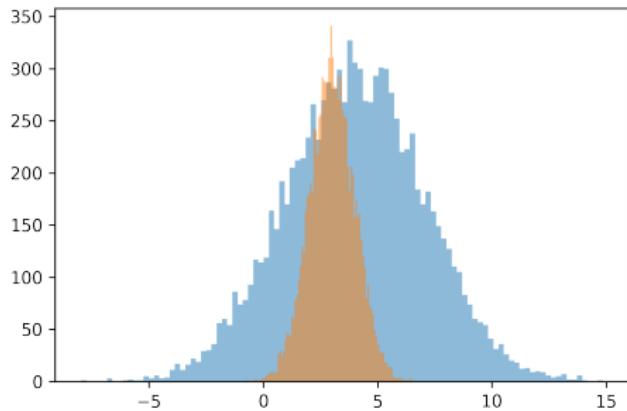




Неопределённость

Media Mix Models ([подробнее](#)
[здесь](#))

- Насколько ценные дополнительные инвестиции в \$1000 для ВК? или Яндекса?
- Насколько велика неопределённость оценок?
- Большая ценность, но большая неопределённость, или небольшая ценность небольшая неопределённость?
- На что выделить деньги?



Вывод

Оценка неопределённости помогает принять более информированные решения



Резюме

Байесовский фреймворк состоит из:

- априорного распределения (что мы изначально думаем о параметрах)
- правдоподобия (имеющиеся факты)
- апостериорного распределения (представление о параметрах, обновлённое на основе априорного распределения и данных)



Модель

$$p_{\mathcal{M}}(\Theta | \mathcal{D}) = \frac{p_{\mathcal{M}}(\mathcal{D} | \Theta) p_{\mathcal{M}}(\Theta)}{p_{\mathcal{M}}(\mathcal{D})}$$

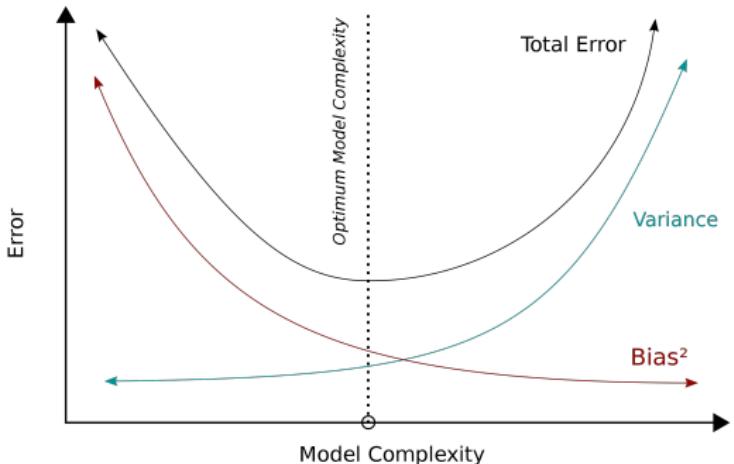
Модель – "одно из многих" описаний задачи.



Дilemma смещения-дисперсии

Как получить хорошую модель

- ① Начните с заведомо простой модели
- ② Сделайте так, чтобы она хорошо сэмплировалась
- ③ Усложните модель
- ④ ...
- ⑤ Выберите наилучшую модель, используя кросс-валидацию

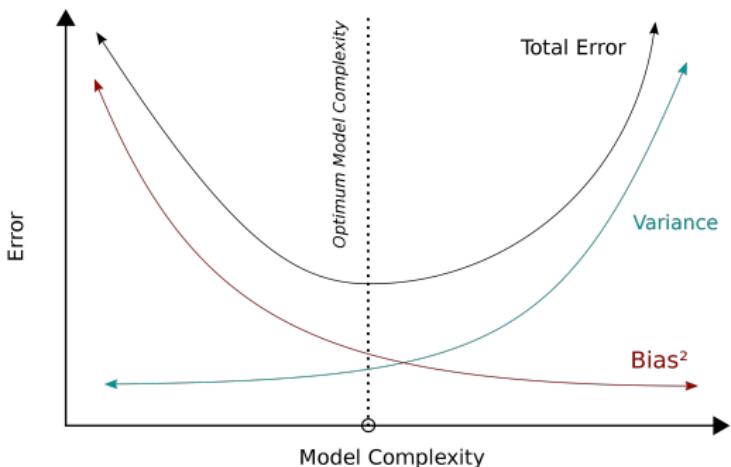




Дilemma смещения-дисперсии

Как получить хорошую модель

- ① Начните с заведомо простой модели
- ② Сделайте так, чтобы она хорошо сэмплировалась
- ③ Усложните модель
- ④ ...
- ⑤ Выберите наилучшую модель, используя кросс-валидацию



Частая ошибка

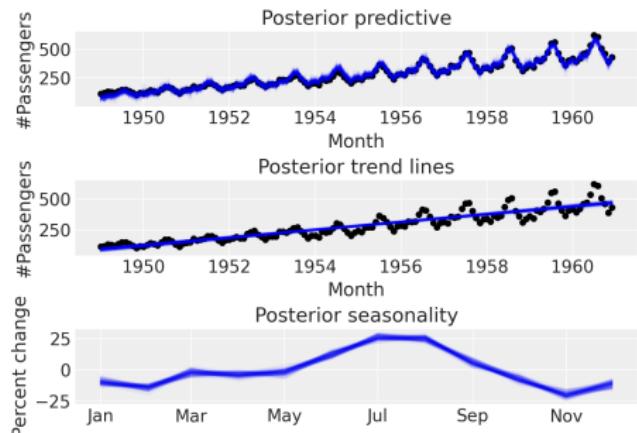
Если начать со сложной модели, её будет сложно отлаживать



$$y(t) = \underbrace{g(t)}_{\text{тренд}} + \underbrace{s(t)}_{\text{сезонность}} + \underbrace{r(X_t)}_{\text{регрессоры}} + \varepsilon_t$$

Усложнение модели

- ① начнём с простой модели тренда
- ② добавим сезонность
- ③ детализируем сезонность, добавим выходные дни или другие признаки



Когда использовать байесовские методы?



Байесовские методы не подходят для всех случаев одновременно

- Они требуют дополнительных навыков по сравнению с классическим машинным обучением
- Вам могут не понадобиться оценки неопределённости



Байесовские методы не подходят для всех случаев одновременно

- Они требуют дополнительных навыков по сравнению с классическим машинным обучением
- Вам могут не понадобиться оценки неопределённости

Утверждение

Байесовские модели начинаются там, где заканчивается парадигма fit-predict



Байесовские методы не подходят для всех случаев одновременно

- Они требуют дополнительных навыков по сравнению с классическим машинным обучением
- Вам могут не понадобиться оценки неопределённости

Утверждение

Байесовские модели начинаются там, где заканчивается парадигма fit-predict

- Байес даёт интерпретируемые "доверительные интервалы"
- Байес даёт гибкость и контроль над моделью
- Байес надёжен при недостатке данных

Но у всего есть своя цена...

Вы ДОЛЖНЫ понимать свою модель



- ① Чистый Python!
- ② Автоматический статистический вывод
- ③ Без сложных формул для МСМС!
- ④ Визуализации с ArviZ
- ⑤ Воспроизводимые исследования
- ⑥ Используется в индустрии
- ⑦ Огромное сообщество
- ⑧ Активно разрабатывается

<https://github.com/pymc-devs/pymc>





Библиография I

-  L. Cragg and K. Nation.
Self-ordered pointing as a test of working memory in typically developing children.
Memory (Hove, England), 15:526–35, 08 2007.
-  F. McElduff, M. Cortina-Borja, S.-K. Chan, and A. Wade.
When t-tests or wilcoxon-mann-whitney tests won't do.
Advances in Physiology Education, 34(3):128–133, 2010.
PMID: 20826766.
-  M. Zondervan-Zwijnenburg, M. Peeters, S. Depaoli, and R. V. de Schoot.
Where do priors come from? applying guidelines to construct informative priors in small sample research.
Research in Human Development, 14(4):305–320, 2017.