Прикладная статистика в машинном обучении
Лекция 9
Байесовский взгляд на подбор моделей

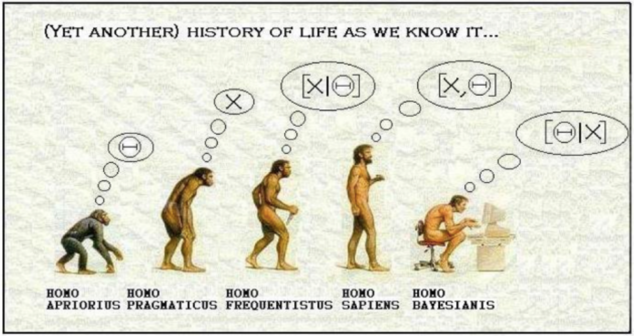
И. К. Козлов
(*Мехмат МГУ*)

2022

Эволюция

В прошлый раз мы совершили “эволюцию сознания”, познав формулу Байеса

$$p(\theta \mid x^n) = \frac{p(x^n \mid \theta)\, p(\theta)}{\int p(x^n \mid \theta)\, p(\theta) d\theta}.$$



Краткое содержание предыдущих серий

Вспомним — какая на прошлом занятии была главная проблема с байесовским выводом?

Сопряжённые распределения

- Проблема. Может не считаться знаменатель в теореме Байеса

$$\int p(x^n \mid \theta)\, p(\theta) d\theta.$$

Сопряжённые распределения

- Проблема. Может не считаться знаменатель в теореме Байеса

$$\int p(x^n \mid \theta)\, p(\theta) d\theta.$$

- Решение. Рассматривать сопряжённые семейства распределений:
    - Априорное $p(\theta)$ — из $\mathcal{A}$,
    - Функция правдоподобия $p(x^n \mid \theta)$— из $\mathcal{B}$.
    - Апостериорное $p(\theta \mid x^n)$ — тоже из $\mathcal{A}$.

Сопряжённые распределения

- Проблема. Может не считаться знаменатель в теореме Байеса

$$\int p(x^n \mid \theta)\, p(\theta) d\theta.$$

- Решение. Рассматривать сопряжённые семейства распределений:
  - Априорное $p(\theta)$ — из $\mathcal{A}$,
  - Функция правдоподобия $p(x^n \mid \theta)$— из $\mathcal{B}$.
  - Апостериорное $p(\theta \mid x^n)$ — тоже из $\mathcal{A}$.

- **Вопрос:** Есть ли большое семейство распределений, для которого известно сопряжённое семейство, и легко производить байесовский вывод?

Экспоненциальное семейство распределений

Экспоненциальный класс

Экспоненциальный класс

Семейство распределений, чья плотность может быть представлена в виде:

$$f(x|\theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right),$$

Экспоненциальный класс

Семейство распределений, чья плотность может быть представлена в виде:

$$f(x|\theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right),$$

- $h(x) \geq 0$,
- $\theta^T u(x)$ означает $\theta_1 u_1(x) + \dots + \theta_k u_k(x)$,
- Чтобы интеграл от плотности равнялся единице:

$$g(\theta) = \int h(x) \exp\left(\theta^T u(x)\right) dx.$$

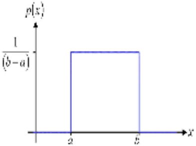
Экспоненциальный класс

YES

- normal
- exponential
- gamma
- Bernoulli
- Poisson
- categorical
- chi-squared
- beta
- Dirichlet
- geometric
- binomial (fixed number of trials)
- multinomial (fixed number of trials)
- negative binomial (fixed number of failures)

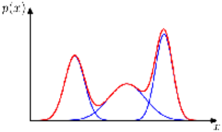
Большинство стандартных распределений

NO



Uniform Distributions

Student



Mixtures

Сопряжённые семейства

Для экспоненциального класса легко строятся сопряжённые семейства.

Рассмотрим экспонециальный класс:

$$f(x|\theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right),$$

Сопряжённые семейства

Для экспоненциального класса легко строятся сопряжённые семейства.

Рассмотрим экспонецициальный класс:

$$f(x|\theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right),$$

Посмотрим на него, как на функцию от θ:

$$\frac{C}{g(\theta)^\nu} \exp(\text{Линейно по } \theta). \tag{1}$$

Замечание. Семейство вида (1) сохраняется при умножении на $f(x|\theta)$. Это и есть сопряжённое семейство.

Сопряжённые семейства

Для экспоненциального класса

$$f(x \mid \theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right)$$

сопряжённым является семейство

$$g(\theta \mid \chi, \nu) = \frac{q(\chi, \nu)}{g(\theta)^\nu} \exp(\theta^T \chi).$$

Сопряжённые семейства

Для экспоненциального класса

$$f(x \mid \theta) = \frac{h(x)}{g(\theta)} \exp\left(\theta^T u(x)\right)$$

сопряжённым является семейство

$$g(\theta \mid \chi, \nu) = \frac{q(\chi, \nu)}{g(\theta)^\nu} \exp(\theta^T \chi).$$

Находим **апостериорное распределение** — произведение априорного на правдоподобия (т.е. на произведение вероятности выборки):

$$f(X \mid \theta) g(\theta \mid \chi, \nu) = \prod_{i=1}^{n} f(x_i \mid \theta) g(\theta \mid \chi, \nu) \propto$$

$$\propto \frac{1}{g(\theta)^{\nu+n}} \exp(\theta^T \left[\chi + \sum_{i=1}^{n} u(x_i)\right]).$$

Сопряжённые семейства

Элементарно находится апостериорное распределение.

Параметры апостериорного семейства:

$$\nu' = \nu + n, \qquad \chi' = \chi + \sum_{i=1}^{n} u(x_i)$$

Да, весь байесовский вывод - в 1 строчку, пару функций сложить.

Сопряжённые семейства

“Физический смысл” параметров априорного распределения:

- ν — число наблюдений.
- χ — сумма достаточных статистик $u(x_i)$.

Далее обсудим — что такое достаточные статистики, и какие у них основные свойства.

Достаточные статистики

1. Экспоненциальное семейство распределений
2. Достаточные статистики
   - Теорема Рао–Блэквелла–Колмогорова
3. Байесовский подход к ML
   - Этап обучения
   - Этап тестирования
4. Принцип максимума обоснованности

Ужать информацию

Иногда хочется оставить только самое нужное,

и ужать информацию *покомпактней*.

Достаточные статистики

Пусть X — i.i.d. выборка из F_θ.

Достаточная статистика — статистика $T = \mathrm{T}(X)$, которая содержит в себе всю информацию о параметре θ, содержащуюся в выборке X.

Формально, $T = \mathrm{T}(X)$ — достаточная статистика, если

$$\mathbb{P}(X \mid \mathrm{T}(X) = t,\ \theta) = \mathbb{P}(X \mid \mathrm{T}(X) = t).$$

Достаточные статистики

Простейший пример достаточной статистики — вся выборка X.

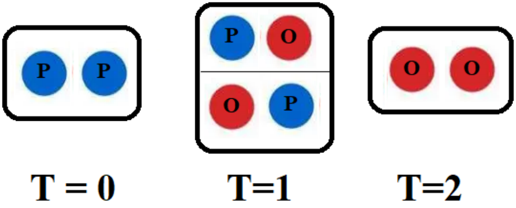
Хочется хранить информацию покомпактней.

Достаточные статистики

Пример-2. Рассмотрим схему Бернулли (“монетка”).

Достаточная статистика T — число выпавших орлов. Если фиксировать $T = t$, то все варианты $X = x$ равновероятны.

Вероятности $P(X = x \mid T = t)$ не зависят от вероятности выпадения орла θ.



Достаточные статистики

Пусть $f(x|\theta)$ — плотность распределения.

Простой критерий для достаточных статистик.

Теорема факторизации Фишера

$T(x)$ — достаточная статистика $\Leftrightarrow$ существует разложение

$$f(x \mid \theta) = h(x)\, q(\theta, T(x)).$$

Грубо говоря, *всё, что в плотности зависит от θ, встречается вместе с $T(x)$*.

Докажем теорему факторизации на семинаре.

Достаточные статистики

Напрашивающийся пример. Плотность для экспоненциального класса

$$f(x \mid \theta) = h(x) \cdot \frac{1}{g(\theta)} \exp\left(\theta^T u(x)\right)$$

имеет требуемый вид

$$f(x \mid \theta) = h(x) \cdot q(\theta, T(x)).$$

Достаточная статистика $T(x) = u(x)$.

Достаточные статистики

У экспоненциального класса **конечное число** k достаточных статистик

$$u_1(X), \quad \ldots, \quad u_k(X).$$

Оказывается, "другого такого класса не найти":

Достаточные статистики

У экспоненциального класса **конечное число** k достаточных статистик

$$u_1(X), \quad \ldots, \quad u_k(X).$$

Оказывается, “другого такого класса не найти”:

Среди гладких семейств распределений $f(x;\theta)$ с фиксированным носителем

$$\{x : f(x;\theta) > 0\}$$

только у экспоненциальных семейств существуют непрерывные k-мерные достаточные статистики.

Подробнее — см. Глава 10, §3, Теорема 2

М. Б. Лагутин,
Наглядная математическая статистика.

Теорема Рао–Блэквелла–Колмогорова

Теорема Рао–Блэквелла–Колмогорова

Достаточные статистики делают несмещённые оценки лучше.

- Пусть T — достаточная статистика;
- $\hat{\theta}$ — несмещённая оценка на θ и $\mathbb{V}\hat{\theta} < \infty$.

Теорема Рао–Блэквелла–Колмогорова

Достаточные статистики делают несмещённые оценки лучше.

- Пусть T — достаточная статистика;
- $\hat{\theta}$ — несмещённая оценка на θ и $\mathbb{V}\hat{\theta} < \infty$.

Теорема Рао–Блэквелла–Колмогорова

Оценка
$$\hat{\theta}_T = \mathbb{E}\left(\hat{\theta} \mid T\right)$$

- будет несмещённой оценкой
$$\mathbb{E}\hat{\theta}_T = \theta,$$
- с небо́льшей дисперсией
$$\mathbb{V}\hat{\theta}_T \le \mathbb{V}\hat{\theta}.$$

Теорема Рао–Блэквелла–Колмогорова

Доказательство.

- Оценка — это функция от выборки $x_1, \ldots, x_n$. Она не должна зависеть от θ.

  Проверим, что $\hat{\theta}_T = \mathbb{E}\left(\hat{\theta} \mid T\right)$ *не зависит от* θ.

Теорема Рао–Блэквелла–Колмогорова

Доказательство.

- Оценка — это функция от выборки $x_1, \ldots, x_n$. Она не должна зависеть от θ.

  Проверим, что $\hat{\theta}_T = \mathbb{E}\left(\hat{\theta} \mid T\right)$ *не зависит от* θ.

  Действительно, по определению условного матожидания
  $$\mathbb{E}\left(\hat{\theta} \mid T\right) = \int \hat{\theta}(x) p(x \mid T = t, \theta) dx.$$
  Поскольку T — достаточная статистика, $p(x \mid T = t, \theta) = p(x \mid T = t)$.

  Поэтому интеграл (и $\hat{\theta}_T$) не зависит от θ.

Теорема Рао–Блэквелла–Колмогорова

- *Докажем несмещённость*
$$\mathbb{E}\hat{\theta}_T = \theta.$$

Q: Мы встречались с выражением вида $\mathbb{E}(\mathbb{E}(X \mid Y))$?

Теорема Рао–Блэквелла–Колмогорова

- *Докажем несмещённость*

$$\mathbb{E}\hat{\theta}_T = \theta.$$

Q: Мы встречались с выражением вида $\mathbb{E}(\mathbb{E}(X \mid Y))$?

A: Вспоминаем: “матожидание убирает условие”. Используем несмещённость $\hat{\theta}$:

$$\mathbb{E}\hat{\theta}_T = \mathbb{E}\left[\mathbb{E}\left(\hat{\theta} \mid T\right)\right] = \mathbb{E}\hat{\theta} = \theta.$$

Теорема Рао–Блэквелла–Колмогорова

- *Дисперсия не увеличивается*
$$\mathbb{V}\hat{\theta} \geq \mathbb{V}\hat{\theta}_T.$$

Q: Мы встречались с выражением вида $\mathbb{V}(\mathbb{E}(X \mid Y))$?

Теорема Рао–Блэквелла–Колмогорова

- *Дисперсия не увеличивается*
$$\mathbb{V}\hat{\theta} \geq \mathbb{V}\hat{\theta}_T.$$

Q: Мы встречались с выражением вида $\mathbb{V}(\mathbb{E}(X \mid Y))$?

A: Да, утверждение следует из закона полной дисперсии
$$\mathbb{V}(\mathbb{E}[X \mid Y]) = \mathbb{V}(X) - \mathbb{E}[\mathbb{V}(X \mid Y)].$$

Теорема Рао–Блэквелла–Колмогорова доказана.

Перерыв

Перерыв

Байес в ML

Во 2ой части лекции поговорим про Байес в ML:

1. Формулировка и схема решения ML-задач на байесовском языке.
2. Принцип максимума обоснованности.
   Объясним как сравнивать модели = перебирать гиперпараметры.

Байесовский подход к ML

ML-модели

ML-модели — по сути отображения

$$\text{Data} \xrightarrow{F_\theta} \text{Target}\,.$$

Скажем, по картинке предсказываем метку класса

Вероятностный подход к ML

Мы рассматриваем вероятностные модели — совместные распределения на интересующие нас переменные.

Будет 3 группы переменных X, T, θ:

- X — наблюдаемые переменные (Data).
- T — это целевые переменные (Target).
- θ — параметры модели.

2 типа моделей

В курсе встречается 2 типа моделей.

Ниже обсудим дискриминативные модели, для генеративных рассуждения аналогичны.

Дискриминативные	**Генеративные**
$p(t, \theta \mid x) = p(t \mid x, \theta) p(\theta)$	$p(x, t, \theta) = p(x \mid t, \theta) p(t \mid \theta) p(\theta)$
При известном x можно предсказывать значение t.	Можно генерировать x.

2 типа моделей

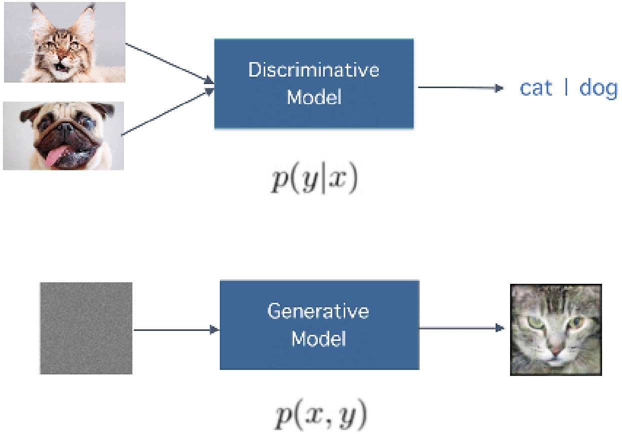
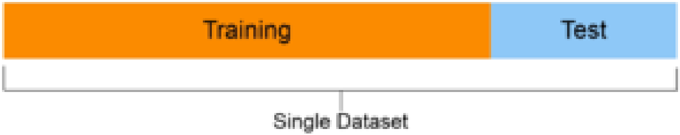


Рис.: Мы с ними уже сталкивались

2 типа моделей

Как обычно, при работе с ML-моделями, данные поделены на Train и Test.

- Обучаем модель на Train (т.е. на X_{tr}, T_{tr}).
- Затем предсказываем ответы для Test (по x_{test} угадываем t_{test}).



Этап обучения

Классический подход.

Дискриминативная модель — это задание функции правдоподобия

$$p(t \mid x, \theta).$$

Если знаем данные x и параметры модели θ, то можем найти метку t.

Дискриминативные модели. Классика

Классический подход.

Дискриминативная модель — это задание функции правдоподобия

$$p(t \mid x, \theta).$$

Если знаем данные x и параметры модели θ, то можем найти метку t.

Q: Как найти параметры модели θ?

Дискриминативные модели. Классика

Классический подход.

Дискриминативная модель — это задание функции правдоподобия

$$p(t \mid x, \theta).$$

Если знаем данные x и параметры модели θ, то можем найти метку t.

Q: Как найти параметры модели θ?

A: Например, оценкой максимума правдоподобия:

$$\hat{\theta}_{\mathrm{ML}} = \operatorname*{argmax}_{\theta \in \Theta} p(T_{tr} \mid X_{tr}, \theta) = \operatorname*{argmax}_{\theta \in \Theta} \prod_{i=1}^{N} p(t_i \mid x_i, \theta)$$

Байесовский подход.

- Известна функция правдоподобия $p(t \mid x, \theta)$.
- Задаём априорное распределение $p(\theta)$.

Q: Что дальше делают байесиане?

Байесовский подход.

- Известна функция правдоподобия $p(t \mid x, \theta)$.
- Задаём априорное распределение $p(\theta)$.

Q: Что дальше делают байесиане?

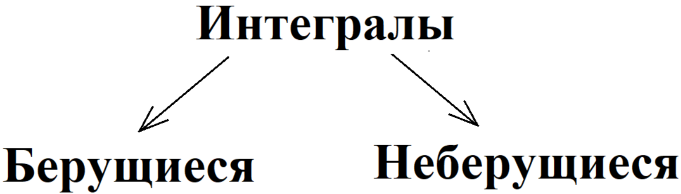
A: Ну конечно, байесовский вывод на неизвестные параметры:

$$p(\theta \mid X_{tr}, T_{tr}) = \frac{p(T_{tr} \mid X_{tr}, \theta) p(\theta)}{\int p(T_{tr} \mid X_{tr}, \theta) p(\theta) d\theta}.$$

Немезида аналитиков

$$p(\theta \mid X_{tr}, T_{tr}) = \frac{p(T_{tr} \mid X_{tr}, \theta) p(\theta)}{\int p(T_{tr} \mid X_{tr}, \theta) p(\theta) d\theta}.$$

А вот и он — *архивраг всех аналитических байесиан* — неберущийся интеграл.



Worst Case Scenario

Рассмотрим худший случай — интеграл не считается.

Мы знаем распределение с точностью до константы

$$p(\theta \mid X_{tr}, T_{tr}) = \frac{p(T_{tr} \mid X_{tr}, \theta) p(\theta)}{Z}.$$

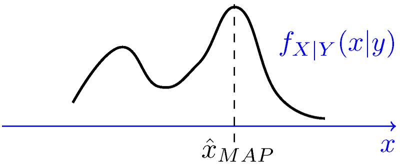
Q: Что можно сделать? (Что не меняется у функции при умножении на константу?)

MAP-оценка

При умножении на константу мода не меняется.

MAP-оценка (Maximum a posteriori) — это **апостериорный максимум**:

$$\theta_{MP} = \arg\max_{\theta} p(\theta \mid X_{tr}, T_{tr})$$



MAP-оценка

Замечание. MAP-оценка — это поправка MLE θ_{ML}:

$$\begin{aligned}\theta_{MP} &= \arg\max_{\theta} p(T_{tr} \mid X_{tr}, \theta) p(\theta) = \\ &= \arg\max_{\theta} \left[\ln p(T_{tr} \mid X_{tr}, \theta) + \ln p(\theta)\right].\end{aligned}$$

Этап тестирования

Тестирование. Классика

Классический подход.

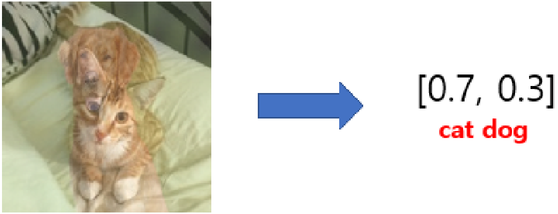
- Выбраны параметры модели θ_{ML}.
- Дан объект x_{test}.
- Нужно оценить неизвестное t_{test}.

Тестирование. Классика

Классический подход.

- Выбраны параметры модели θ_{ML}.
- Дан объект x_{test}.
- Нужно оценить неизвестное t_{test}.

Всё просто — предсказываем $p(t_{test} \mid x_{test}, \theta_{ML})$.



Тестирование. Байес

Байесовский подход.

Наша цель — вычислить вероятность

$$p(t_{test} \mid x_{test}, X_{tr}, T_{tr}).$$

- Видели обучающую выборку X_{tr}, T_{tr}, знаем объект x_{test}.
- Хотим знать вероятности меток t_{test}.

Q: Как это сделать? Какие распределения мы знаем?

Тестирование. Байес

Известны:

- апостериорное распределение на параметры модели:
$$p(\theta \mid X_{tr}, T_{tr})$$
- функция правдоподобия
$$p(t_{test} \mid x_{test}, \theta)$$

Q: Как вычислить $p(t_{test} \mid x_{test}, X_{tr}, T_{tr})$?

Тестирование. Байес

Известны:

- апостериорное распределение на параметры модели:

$$p(\theta \mid X_{tr}, T_{tr})$$

- функция правдоподобия

$$p(t_{test} \mid x_{test}, \theta)$$

Q: Как вычислить $p(t_{test} \mid x_{test}, X_{tr}, T_{tr})$?

В байесовском случае усредняем предсказания моделей по апостериорному распределению на параметр:

$$p(t_{test} \mid x_{test}, X_{tr}, T_{tr}) = \int p(t_{test} \mid x_{test}, \theta) p(\theta \mid X_{tr}, T_{tr}) d\theta.$$

Метод Монте-Карло

Замечание. Даже если распределение известно с точностью до константы, из него можно сэмплировать.

Предсказание модели можно оценить методом Монте-Карло (см. Лекцию 3):

$$p(t_{test} \mid x_{test}, X_{tr}, T_{tr}) = \int p(t_{test} \mid x_{test}, \theta) p(\theta \mid X_{tr}, T_{tr}) d\theta \approx \frac{1}{K} \sum_{k=1}^{K} p(t_{test} \mid x_{test}, \theta_k),$$

где $\theta_k \sim p(\theta \mid X_{tr}, T_{tr})$, т.е. сгенерирована из апостериорного распределения.

Принцип максимума обоснованности

Выбор моделей

Очень кратко обсудим байесовский способ отбора моделей.

Подробнее — см. Главу 3.4.

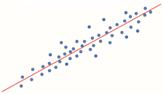
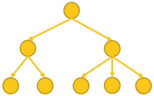
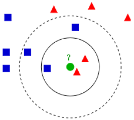
Bishop C.M.
Pattern Recognition and Machine Learning.

Выбор модели

Что нам дано?

Есть несколько моделей (занумеруем их $\mathcal{M} = 1, \ldots, M$).

У каждой из них — свои параметры w.

	Линейная регрессия	Decision Tree	KNN
Модели			
Параметры	Веса w	Глубина	Число соседей **N**

Выбор модели

Нам даны:

- Модели $\mathcal{M}$.
- Обучающая выборка (X_{tr}, Y_{tr}).

Q: Какую модель выбрал бы “байесовский человек”?

Выбор модели

Нам даны:

- Модели $\mathcal{M}$.
- Обучающая выборка (X_{tr}, Y_{tr}).

Q: Какую модель выбрал бы “байесовский человек”?

A: Наиболее вероятную при таких данных $\arg\max \mathbb{P}(\mathcal{M} \mid Y_{tr}, X_{tr})$.

Принцип наибольшей обоснованности

Теорема Байеса — универсальное решение всех проблем.

Поступаем “по-байесовски”:

- Вводим априорное распределение на модели $\mathbb{P}(\mathcal{M})$.
- По формуле Байеса

$$\mathbb{P}\left(\mathcal{M} \mid Y_{tr}, X_{tr}\right) = \frac{\mathbb{P}\left(Y_{tr} \mid X_{tr}, \mathcal{M}\right)\mathbb{P}(\mathcal{M})}{\mathbb{P}\left(Y_{tr} \mid X_{tr}\right)}.$$

- **Q**: Если нам нужна конкретная модель, а не распределение — какую брать?

Принцип наибольшей обоснованности

Максимум обоснованности.

- Если нужна точка — берём MAP-оценку

$$\mathcal{M}_{MP} = \arg\max_{\mathcal{M}} \mathbb{P}\left(Y_{tr} \mid X_{tr}, \mathcal{M}\right) \mathbb{P}(\mathcal{M}).$$

- Первое слагаемое $\mathbb{P}\left(Y_{tr} \mid X_{tr}, \mathcal{M}\right)$ называется обоснованностью (evidence).

- Если для нас модели равновероятны, то получается принцип максимума обоснованности.

$$\mathcal{M}_{ME} = \arg\max_{\mathcal{M}} \mathbb{P}\left(Y_{tr} \mid X_{tr}, \mathcal{M}\right).$$

Принцип наибольшей обоснованности

У модели $\mathcal{M}$ параметры w.

Мы умеем считать вероятности при фиксированных параметрах $\mathbb{P}(Y_{tr} \mid X_{tr}, w)$.

Q: Как посчитать обоснованность модели $\mathbb{P}(Y_{tr} \mid X_{tr}, \mathcal{M})$?

A: Нужно усреднить предсказание по распределению на параметры:

$$\mathbb{P}(Y_{tr} \mid X_{tr}, \mathcal{M}) = \int \mathbb{P}(Y_{tr} \mid X_{tr}, w)\, \mathbb{P}(w \mid \mathcal{M})\, dw$$

Принцип наибольшей обоснованности

Подведём итоги.

Принцип максимума обоснованности

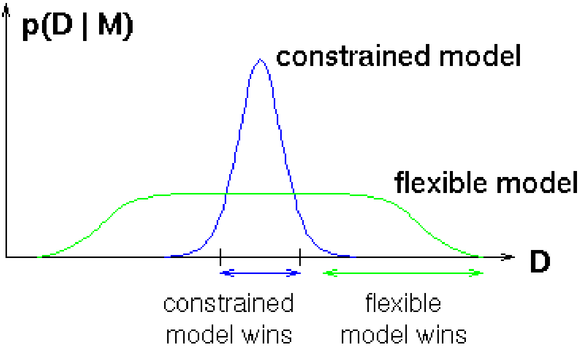
$$\mathcal{M}_{ME} = \arg\max_{\mathcal{M}} \mathbb{P}\,(Y_{tr} \mid X_{tr}, \mathcal{M})$$

$$\mathbb{P}\,(Y_{tr} \mid X_{tr}, \mathcal{M}) = \int \mathbb{P}\,(Y_{tr} \mid X_{tr}, w)\,\mathbb{P}\,(w \mid \mathcal{M})\,dw.$$

Принцип наибольшей обоснованности

В целом, если обозначить данные за D, то метод максимизирует $\mathbb{P}(D \mid \mathcal{M})$.

Метод предпочитает модели, которые *“дают ответы на меньшее число вопросов, но более уверенно”*.



Принцип наибольшей обоснованности

Простой пример "наиболее обоснованной модели".

Рассмотрим 3 кубика — с 4, 6 и 8 гранями. Выпало 5.

Q: Какая модель наиболее обоснованна?

Принцип наибольшей обоснованности

Простой пример "наиболее обоснованной модели".

Рассмотрим 3 кубика — с 4, 6 и 8 гранями. Выпало 5.

Q: Какая модель наиболее обоснованна?

A: Кубик с 6 гранями — у него наибольшая вероятность $\mathbb{P}(Data \mid \mathcal{M}) = \frac{1}{6}$.

Всё чудесатее и чудесатее!

Сегодня мы взглянули на ML “по-байесовски”. Чудно!

А может ли эта абстрактная теория применяться на практике?

Конкретный пример — RVM — обсудим на следующем занятии.