Лекция 9

Продвинутое глубокое обучение: байесовский взгляд

Дополнительные главы машинного обучения

Андрей Фильченков

07.05.2021

План лекции

- Переосмысление байесовского вывода
- Полный байесовский вывод
- Вариационный вывод
- Дропаут как байесовский вывод
- Выбор обоснованной модели

- В презентации используются материалы
  лекций школы «Deep Bayes School» и курса «Нейробайесовские методы машинного обучения Д.П. Ветрова
  лекций С.И. Николенко
  доклада Н.Е. Ханжиной на семинаре MLLab

- Слайды доступны: **shorturl.at/wGV59**
- Видео доступны: **shorturl.at/ovBTZ**

План лекции

- Переосмысление байесовского вывода
- Полный байесовский вывод
- Вариационный вывод
- Дропаут как байесовский вывод
- Выбор обоснованной модели

Задача оценки параметров

Дана выборка $\mathcal{D}$, необходимо оценить параметры распределения, по которому она была получена.

Как мы можем это сделать?

Частотный vs байесовский (1)

Частотный подход:

$$\hat{\theta}_{MLE} = \arg\max_{\theta} p(\mathcal{D}|\theta)$$

Теорема Байеса:

$$\underbrace{p(\theta|\mathcal{D})}_{\text{posterior}} = \frac{\overbrace{p(\mathcal{D}|\theta)}^{\text{likelihood}}\overbrace{p(\theta)}^{\text{prior}}}{\underbrace{p(\mathcal{D})}_{\text{evidence}}}$$

Байесовский подход:

$$\hat{\theta}_{MAP} = \arg\max_{\theta} \underbrace{p(\theta|\mathcal{D})}_{\text{posterior}}$$

$p(\theta|\mathcal{D})$

В байесовском подходе мы получаем распределение на параметры, что дает нам существенно больше информации.

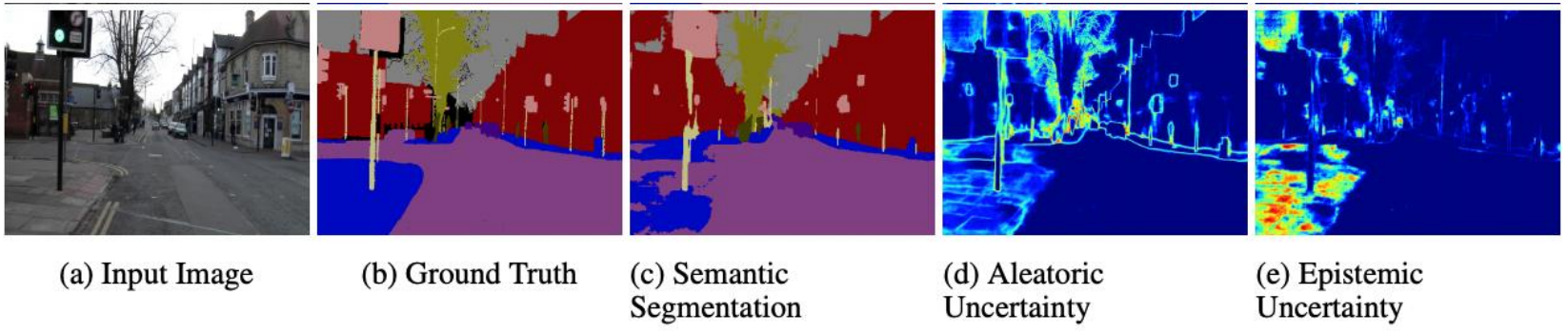
Благодаря этому мы можем оперировать оценками **неопределенности** прогнозов

Типы неопределенности

Неопределенность

Алеаторная (неопределенность данных)

Эпистемическая (неопределенность параметров модели)



(a) Input Image (b) Ground Truth (c) Semantic Segmentation (d) Aleatoric Uncertainty (e) Epistemic Uncertainty

Источники неопределенностей

Источники эпистемической неопределенности

1. Недостаток данных
   - Классы, которые модель не видела
   - Непохожие объекты известных классов

Источники аллеаторной неопределенности

1. Пересечение классов
2. Шум в данных:
   - Ошибки измерения
   - Ошибки разметки

Классический vs байесовский подход

	Классический	Байесовский
Интерпретация вероятности	Объективная неопределенность	Субъективное незнание детерминированного процесса
Величины	Случайные и детерминированные	Все случайные
Метод оценивания (вывод)	MLE	MAP
Оценки	Точные	Распределения
Применимость	Много данных	Любое число данных

Трактовка MAP

$$\hat{\theta}_{MAP} = \arg\max_{\theta} p(\theta|\mathcal{D}) =$$

$$= \arg\max_{\theta} \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \arg\max_{\theta} p(\mathcal{D}|\theta)p(\theta) =$$

$$= \arg\max_{\theta} \ln p(\mathcal{D}|\theta) + \ln p(\theta)$$

Компромисс между априорной информацией и правдоподобием

Связь классического и байесовского подходов

$$\lim_{n\to\infty} p(\theta|x_1, \dots x_n) = \delta(\theta - \theta_{ML})$$

Классический подход можно рассматривать как предельный случай байесовского подхода

Байесовский интерполирует между ситуациями, когда у нас есть все данные и когда у нас нет данных (априорная вероятность).

Анализ байесовского подхода

Преимущества:

- Борьба с переобучение за счет регуляризации
- Композитность — можно комбинировать разные модели
- Обработка данных на лету
- Модели с латентными переменными
- Масштабируемость

Недостатки

- Нет хорошего способа честно выбрать априорную вероятность
- В некоторых случаях проблемы с вычислимостью оценок

План лекции

- Переосмысление байесовского вывода
- Полный байесовский вывод
- Вариационный вывод
- Дропаут как байесовский вывод
- Выбор обоснованной модели

Байесовский вывод

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_\theta p(\mathcal{D}|\theta)p(\theta)\,dy}$$

В чем проблема?

Байесовский вывод

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_\theta p(\mathcal{D}|\theta)p(\theta)\,dy}$$

В чем проблема?

Интеграл обычно неаналитический.

И вычислительно оценить тоже невозможно из-за размерности данных.

Что можно делать?

- Особый вид распределений, когда интеграл аналитический
- Вариационный вывод (variational inference)
- Оценки марковской цепи Монте-Карло (MCMC)

Когда интеграл аналитический

Распределения $p(\theta) \sim \mathcal{A}(\alpha_0)$ и $p(x|\theta) \sim \mathcal{B}(\beta)$ **сопряженные**, если

$$p(\theta|x) = \mathcal{A}(\alpha_1)$$

Если априорное распределение выбрано из класса распределений, сопряженных правдоподобию, то апостериорное распределение можно выписать в явном виде.

Пример №1

Распределение Бернулли:

$$p(m|n,q) = C_n^m q^m (1-q)^{n-m} \sim \mathcal{B}(m|n,q)$$

Сопряженным является бета-распределение:

$$p(q|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1-q)^{b-1} \sim$$

$$\sim \text{Beta}(q|a,b)$$

Пример с монетой (1/3, напоминание)

G это распределение Бернулли:

$$\Pr(X = 1) = \theta, \Pr(X = 0) = 1 - \theta$$

$\mathcal{D} = \{x_1, \dots, x_m\}$ это IID сэмплы из p:

- Броски независимы
- Броски из одного распределения

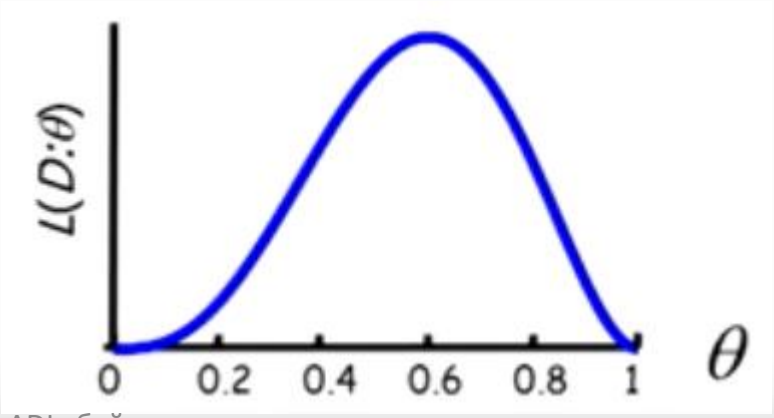
Пример с монетой (2/3, напоминание)

Цель: найти $\theta \in [0,1]$ которая хорошо предсказывает $\mathcal{D}$

Качество предсказания: правдоподобие $\mathcal{D}$ при известном θ

$$L(\theta : D) = p(\mathcal{D}|\theta) = \prod_{i=1}^{m} p(x_i|\theta)$$

$$L(\theta : \langle H, T, T, H, H \rangle)$$



Пример с монетой (3/3, напоминание)

Наблюдения: M_H орлов и M_T решек

Ищем θ, максимизируя правдоподобие

$$\mathrm{L}(\theta: M_H, M_T) = \theta^{M_H}(1-\theta)^{M_T}$$

Сводим к максимизации log-likelihood

$$l(\theta: M_H, M_T) = M_H \log\theta + M_T \log(1-\theta)$$

Дифференцируем log-likelihood и решаем для заданного θ:

$$\hat{\theta} = \frac{M_H}{M_H + M_T}$$

Пример №2

Мультиномиальное распределение:

$$\Pr(X = i) = \beta_i$$

Сопряженным является распределение Дирихле.

Для априорного $\mathrm{Dir}(\alpha)$ апостериорным будет $\mathrm{Dir}(\alpha + \beta)$.

Использование сопряженных распределений

Для большинства известных распределений существуют сопряженные.

Если правдоподобие представляет собой распределение, для которого существует сопряженное, именно его нужно стараться брать в качестве сопряженного.

План лекции

- Переосмысление байесовского вывода
- Полный байесовский вывод
- Вариационный вывод
- Дропаут как байесовский вывод
- Выбор обоснованной модели

Вариационный метод

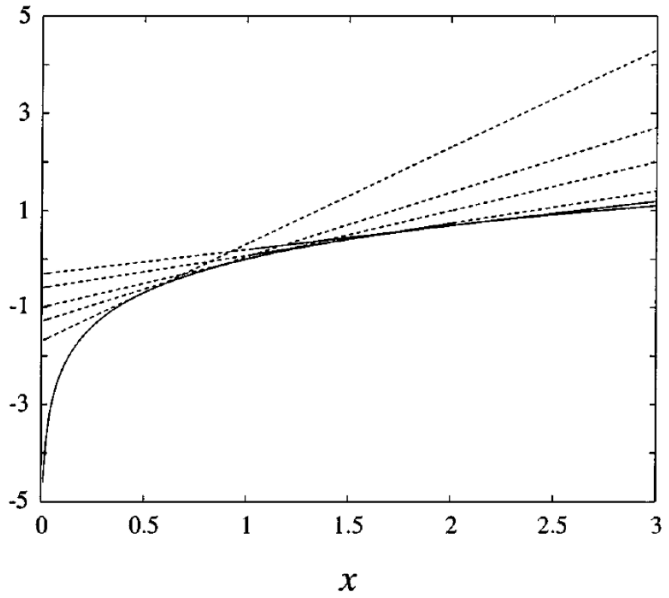
Вариационный метод — простое приближение сложной функции будем искать среди некоторого класса функций за счет выбора в этом классе элемента, который наиболее похож на приближаемую функцию.

Вариационный логарифм

Функцию логарифма можно представить в вариационном виде:

$$\ln x = \min_{\lambda}\{\lambda x - \ln \lambda - 1\}$$

Это семейство линейных функций, каждая из которых дает верхнюю оценку на $\ln x$



Принцип двойственности

Принцип двойственности: если $f(x)$ вогнута, то ее можно представить в виду

$$f(x) = \min_{\lambda}\{\lambda x - f^*(\lambda)\},$$

где $f^*(\lambda)$ — сопряженная функция.

Для вогнутых функций сначала ищем сопряженную, а если нет, то сначала делаем обратимое преобразование, превращающее ее в вогнутую.

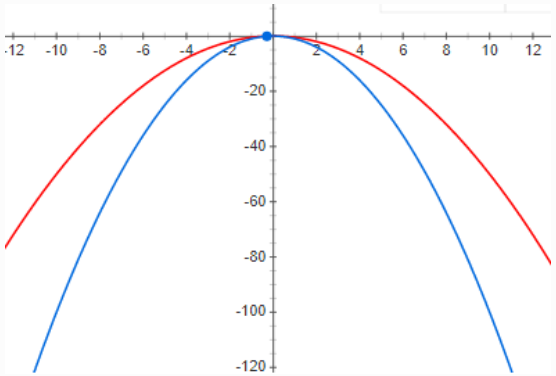
Близость границ

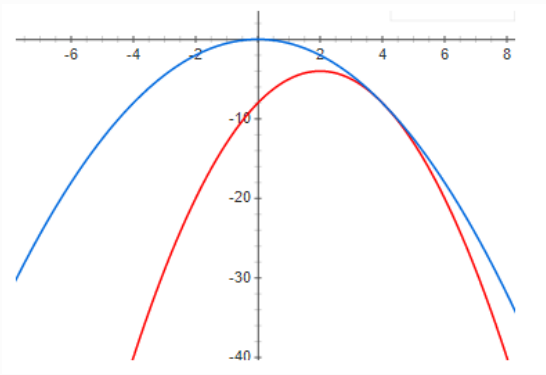
У любой функции существует бесчетное множество нижних (и верхних) границ:

$$f(x) \geq g(x)$$

Близость границ (tightness):

$$\tau = \max_x f(x) - \max_x g(x)$$





Дополнительное распределение

Хотим оценить

$$p(\theta|\mathcal{D}) = \frac{p(\mathcal{D}|\theta)p(\theta)}{p(\mathcal{D})} = \frac{p(\mathcal{D}|\theta)p(\theta)}{\int_\theta p(\mathcal{D}|\theta)p(\theta)\,dy}$$

Пусть $Z_p = \int_\theta p(\mathcal{D}|\theta)p(\theta)\,dy$

Введем дополнительное распределение

$$q(\theta),$$

которое будет максимально похоже на $p(\theta|\mathcal{D})$.

Переход к оптимизации

Похожесть определяем через KL-дивергенцию:

$$\mathrm{KL}(p||q) = \int p(x) \ln \frac{p(x)}{q(x)} dx = -\int p(x) \ln \frac{q(x)}{p(x)} dx$$

Будем решать $\mathrm{KL}(p||q) \to \min_q$, зная p с точностью до нормировочной константы Z_p.

Переход

$$\log Z_p = \int q(x) \log Z_p \, dx = \int q(x) \log \frac{\tilde{p}(x)}{p(x)} dx =$$

$$= \int q(x) \log \frac{\tilde{p}(x) q(x)}{q(x) p(x)} dx =$$

$$= \int q(x) \log \frac{\tilde{p}(x)}{q(x)} dx - \int q(x) \log \frac{p(x)}{q(x)} dx =$$

$$\boxed{L(q) + \mathrm{KL}(q||p)}$$

$\log Z_p \geq L(q)$, причем равенство только при $p(x) = q(x)$

ELBO

$L(q)$ называется вариационной нижней оценкой **(variational lower bound, evidence lower bound, ELBO)**

$$KL(q||p) \to \min_q \quad \Leftrightarrow \quad L(q) \to \max_q$$

Как решать максимизацию?

Оптимизация ELBO

Будем рассматривать факторизованные распределения $q(x) = \prod_i q_i(x_i)$

Решение будем градиентным спуском, причем батчевым.

Однако $L(q) \to \max_q$ в семействе факторизованных распределений не обладает свойством выпуклости.

План лекции

- Переосмысление байесовского вывода
- Полный байесовский вывод
- Вариационный вывод
- Дропаут как байесовский вывод
- Выбор обоснованной модели

Dropout (напоминание)

На каждом шаге обучения будем каждый вес случайным образом «выкидывать».

Это эквивалентно тому, что на каждом шаге обучения будем каждый вес домножать на случайную величину, распределенную по **чему?**

Dropout

На каждом шаге обучения будем каждый вес случайным образом «выкидывать».

Это эквивалентно тому, что на каждом шаге обучения будем каждый вес домножать на случайную величину, распределенную по Бернулли:

$$w_{ijl} \cdot \beta_{ijl}$$
$$\beta_{ijl} \sim \mathcal{B}(\beta_{ijl}|r)$$

Gaussian Dropout

На каждом шаге обучения будем каждый вес домножать на случайную величину, распределенную по Гауссу

$$w_{ijl} \cdot \delta_{ijl}$$

$$\delta_{ijl} \sim \mathcal{N}(\delta_{ijl} | 1, \alpha)$$

$$\alpha = \frac{1 - r}{r}$$

Откуда нормальное распределение?

Сумма бернуллиевских случайных величин сходится к нормальному распределению.

Сумма гауссиан с указанными коэффициентами дает то же самое.

Замена функционала

Какую функцию мы оптимизируем, когда используем дропаут?

Очевидно, что не логарифм правдоподобия.

Оптимизируемый функционал

$$\nabla_k Q = n \frac{\partial}{\partial w_{ijl}} \log p(y_k | x_k w_{ijl} \circ \delta)$$

$$k \sim U(1; l)$$

$$\nabla Q = \sum_{k=1}^{n} \int r(\delta) \frac{\partial}{\partial w_{ijl}} \log p\left(y_k \middle| x_k w_{ijl} \circ \delta\right) d\delta$$

$$Q = \sum_{k=1}^{n} \int r(\delta) \log p\left(y_k \middle| x_k w_{ijl} \circ \delta\right) d\delta$$

Трактовка

Мы пытаемся оптимизировать матожидание логарифма правдоподобия по дополнительному шумовому распределению.

Оптимизация по этому функционалу это фактически зашумление градиента при обучении.

Другие варианты

- DropLayer
- DropChannel
- DropBlock
- DropPath

План лекции

- Переосмысление байесовского вывода
- Полный байесовский вывод
- Вариационный вывод
- Дропаут как байесовский вывод
- Выбор обоснованной модели

Вероятностная схема обучения

Пусть $\mathcal{D} = \{(x_i, y_i)\}$

При обучении оценивается

$$p(\theta|\mathcal{D}) = \frac{p(Y|X,\theta)p(\theta)}{p(Y|X)} = \frac{\prod_{i=1}^{n} p(y_i|x_i,\theta)p(\theta)}{\int \prod_{i=1}^{n} p(y_i|x_i,\theta)p(\theta)d\theta}$$

Для вывода используется

$$p(y_*|x_*,\mathcal{D}) = \int p(y_*|x_*,\theta)p(\theta|\mathcal{D})d\theta$$

Это взвешенный ансамбль моделей.

Использование MAP-оценки

Интегрирование не всегда аналитично

Альтернатива — использование θ_{MAP}

$$p(y_*|x_*, \mathcal{D}) = \int p(y_*|x_*, \theta)p(\theta|\mathcal{D})d\theta \approx$$
$$\approx \int p(y_*|x_*, \theta)\delta_{\theta_{MAP}}(\theta)d\theta = p(y_*|x_*, \theta_{MAP})$$

Когда такое приближение осмысленно?

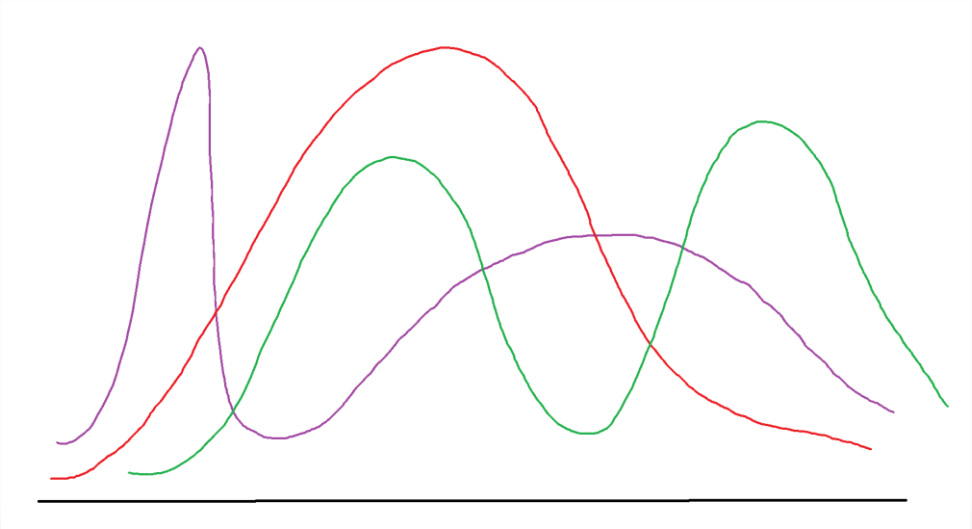
Использование MAP-оценки

Интегрирование не всегда аналитично

Альтернатива — использование θ_{MAP}

$$p(y_*|x_*, \mathcal{D}) = \int p(y_*|x_*, \theta)p(\theta|\mathcal{D})d\theta \approx$$

$$\approx \int p(y_*|x_*, \theta)\delta_{\theta_{MAP}}(\theta)d\theta = p(y_*|x_*, \theta_{MAP})$$



Выбор априорного распределения

Пусть $p(\theta) = p(\theta|\alpha)$.

Чтобы определить $p(\theta|\mathcal{D}, \alpha)$, необходимо определить α

Как это сделать?

Обоснованность

Пусть $p(\theta) = p(\theta|\alpha)$.

Чтобы определить $p(\theta|\mathcal{D}, \alpha)$, необходимо определить α

$$p(\alpha|\mathcal{D}) = \frac{p(Y|X,\alpha)p(\alpha)}{p(Y|X)} = \frac{\prod_{i=1}^{n} p(y_i|x_i,\alpha)p(\alpha)}{\int \prod_{i=1}^{n} p(y_i|x_i,\alpha)p(\alpha)d\alpha}$$

$$p(Y|X,\alpha) = \int p(Y|X,\theta)p(\theta,\alpha)d\theta$$

$p(Y|X,\alpha)$ — **обоснованность (evidence)**

Принцип максимальной обоснованности модели

Принцип максимальной обоснованности модели — для выбора априорного распределения на скрытые переменные следует оптимизировать обоснованность.

Слайд

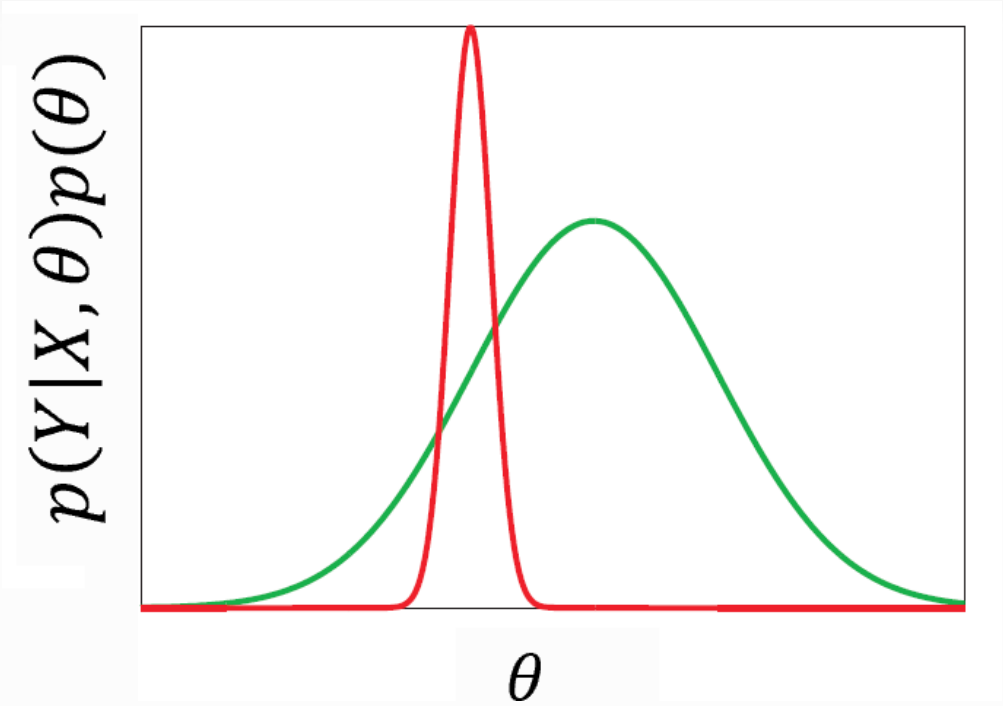
Схему можно (но обычно не нужно) масштабировать до произвольного числа уровней.

Если нет, то $p(\alpha) = const$

Тогда $p(\alpha|X, Y) \propto p(Y|x, \alpha)$

$$p(y_*|x_*, \mathcal{D}) = \int p(y_*|x_*, \theta)p(\theta|\alpha_{ME}, \mathcal{D})d\theta$$

Обоснованность двух моделей



Доля «хороших» алгоритмов зеленой модели больше, чем у красной, поэтому ее обоснованность выше

Сравнение моделей

Пик	Высокий	Низкий
Узкий	Модель слишком чувствительна, обоснованность низкая	Модель неадекватна, обоснованность низкая
Широкий	Модель позволяет предложить множество хороших объяснений, обоснованность высокая	Модель слишком простая, но ее обоснованность высокая