БАЙЕСОВСКОЕ СРАВНЕНИЕ МОДЕЛЕЙ

Сергей Николенко

СПбГУ — Санкт-Петербург

25 октября 2022 г.

Random facts:

- 25 октября 1147 г. прошло одно из главных сражений Реконкисты — осада Лиссабона, когда крестоносцы и португальцы под командованием Афонсу Энрикеша захватили Лиссабон у Альмохадов, а 25 октября 1415 г. — одно из главных сражений Столетней войны, битва при Азенкуре, когда имевшие многократное превосходство в численности французы были наголову разбиты вооружёнными длинными луками англичанами под командованием Генриха V
- 25 октября 1854 г. в Балаклавском сражении, крупнейшей битве Крымской войны, 93-й шотландский пехотный полк выстроился в «тонкую красную линию» (thin red line), отражая атаку 1-го Уральского казачьего полка, а затем русские пушки попыталась отбить кавалерия под командованием лорда Кардигана (Charge of the Light Brigade)
- 25 октября 1955 г. американская фирма «Tappan Company» впервые представила микроволновую печь
- 25 октября 1949 г. началась битва за Цзиньмэнь, одна из последних в Гражданской войне в Китае; победа войск Гоминьдана позволила им сохранить за собой Тайвань
- 25 октября 1990 г. в СССР был принят закон о свободе вероисповедания, а также вышел первый выпуск телеигры Влада Листьева «Поле чудес»

Байесовское сравнение моделей

Байесовское сравнение моделей

- Мы говорили о том, что при увеличении числа параметров модели возникает оверфиттинг.
- Как этого избежать? Как сравнить модели с разным числом параметров?
- Теория байесовского вывода предлагает такой выход: давайте будем не точечные оценки параметров модели рассматривать, а тоже интегрировать по параметрам модели.

Байесовское сравнение моделей

- Пусть мы хотим сравнить модели из множества $\{M_i\}_{i=1}^{L}$.
- Модель – это распределение вероятностей над данными D.
- По тестовому набору D можно оценить апостериорное распределение

$$p(M_i \mid D) \propto p(M_i)p(D \mid M_i).$$

Байесовское сравнение моделей

- Если знать апостериорное распределение, то можно сделать предсказание:

$$p(t \mid \mathbf{x}, D) = \sum_{i=1}^{L} p(t \mid \mathbf{x}, M_i, D) p(M_i \mid D).$$

- *Model selection* (выбор модели) – это когда мы приближаем предсказание, выбирая просто самую (апостериорно) вероятную модель.

Байесовское сравнение моделей

- Если модель определена параметрически, через $\mathbf{w}$, то

$$p(D \mid M_i) = \int p(D \mid \mathbf{w}, M_i) p(\mathbf{w} \mid M_i) d\mathbf{w}.$$

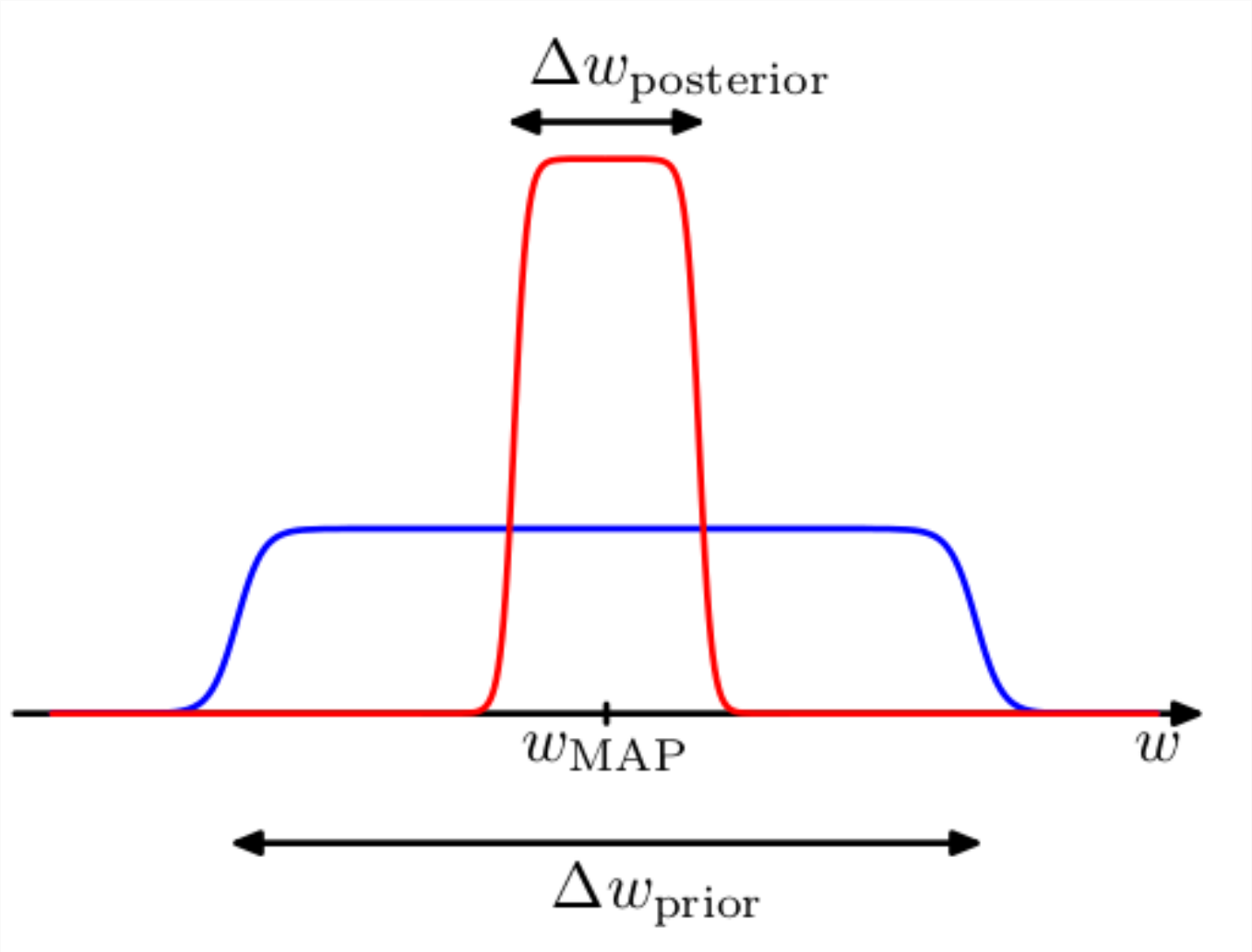
- Т.е. это вероятность сгенерировать D, если выбирать параметры модели по её априорному распределению, а потом накидывать данные.
- Это, кстати, в точности знаменатель из теоремы Байеса:

$$p(\mathbf{w} \mid M_i, D) = \frac{p(D \mid \mathbf{w}, M_i) p(\mathbf{w} \mid M_i)}{p(D \mid M_i)}.$$

Байесовское сравнение моделей

- Предположим, что у модели один параметр w, а апостериорное распределение – это острый пик вокруг w_{MAP} шириной $\Delta w_{\mathrm{posterior}}$.
- Тогда можно приблизить $p(D) = \int p(D \mid w)p(w)dw$ как значение в максимуме, умноженное на ширину.
- Предположим ещё, что априорное распределение тоже плоское, $p(w) = \frac{1}{\Delta w_{\mathrm{prior}}}$.

Приближение $p(D)$



Приближение $p(D)$

- Тогда получится

$$p(D) = \int p(D \mid w)p(w)dw \approx p\left(D \mid w_{\text{MAP}}\right) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}},$$

$$\ln p(D) \approx \ln p\left(D \mid w_{\text{MAP}}\right) + \ln\left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}\right).$$

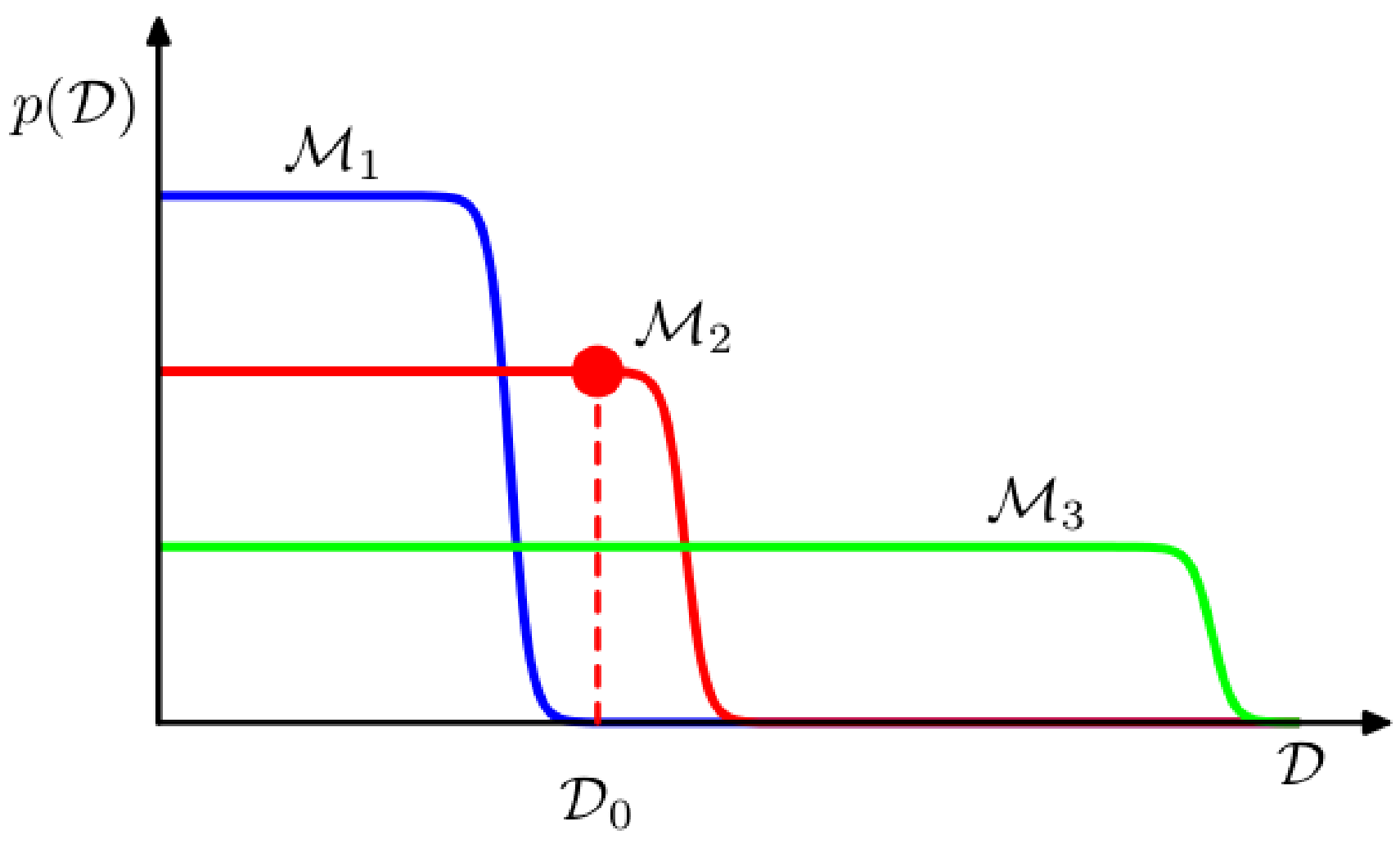
- Это значит, что мы добавляем штраф за «слишком узкое» апостериорное распределение – то есть в точности штраф за оверфиттинг!
- Для модели из M параметров, если предположить, что у них одинаковые $\Delta w_{\text{posterior}}$, получим

$$\ln p(D) \approx \ln p\left(D \mid w_{\text{MAP}}\right) + M \ln\left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}}\right).$$

Другой взгляд

- Другими словами: давайте посмотрим, какие датасеты может генерировать та или иная модель.
- Простая модель (e.g., линейная) генерирует похожие датасеты, «мало» разных датасетов, у неё высокая $p(D \mid M)$.
- Сложная модель (e.g., многочлен девятой степени) генерирует «много» разных датасетов, у неё низкая $p(D \mid M)$.
- Но сложная может хорошо выразить датасеты, которые не может выразить простая; поэтому в сумме надо выбирать «среднюю».

Приближение $p(D)$



Правильный ответ лучше

- Sanity check: тут какие-то штрафы мы навводили; будет ли истинный правильный ответ $p(D \mid M_{\text{true}})$ всегда оптимальным в этом смысле?
- Конечно, для конкретного датасета может так повезти, что не будет.
- Но если усреднить по всем датасетам, выбранным по $p(D \mid M_{\text{true}})$...

Правильный ответ лучше

- ...то получится

$$\mathrm{E}\left[\ln \frac{p(D \mid M_{\mathrm{true}})}{p(D \mid M)}\right] = \int p(D \mid M_{\mathrm{true}}) \ln \frac{p(D \mid M_{\mathrm{true}})}{p(D \mid M)} dD.$$

- Это называется *расстоянием Кульбака-Лейблера* (Kullback-Leibler divergence) между распределениями $p(D \mid M_{\mathrm{true}})$ и $p(D \mid M)$.

Сравнение моделей по Лапласу

- А ещё мы можем сравнивать модели при помощи лапласовской аппроксимации.
- Напомним: чтобы сравнить модели из множества $\{M_i\}_{i=1}^{L}$, по тестовому набору D оценим апостериорное распределение

$$p(M_i \mid D) \propto p(M_i)p(D \mid M_i).$$

- Если модель определена параметрически, то $p(D \mid M_i) = \int p(D \mid \theta, M_i)p(\theta \mid M_i)d\theta$.
- Это вероятность сгенерировать D, если выбирать параметры модели по её априорному распределению; знаменатель из теоремы Байеса:

$$p(\theta \mid M_i, D) = \frac{p(D \mid \theta, M_i)p(\theta \mid M_i)}{p(D \mid M_i)}.$$

Сравнение моделей по Лапласу

- Мы раньше приближали фактически кусочно-постоянной функцией.
- Теперь давайте гауссианом приблизим; возьмём интеграл:

$$Z = \int f(\mathbf{z})d\mathbf{z} \approx \int f(\mathbf{z}_0)e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top \mathbf{A}(\mathbf{z}-\mathbf{z}_0)}d\mathbf{z} = f(\mathbf{z}_0)\frac{(2\pi)^{M/2}}{|\mathbf{A}|^{1/2}}.$$

- А у нас $Z = p(D)$, $f(\theta) = p(D \mid \theta)p(\theta)$.

Сравнение моделей по Лапласу

- Получаем

$$\ln p(D) \approx \ln p(D \mid \theta_{\text{MAP}}) + \ln P(\theta_{\text{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|.$$

- $\ln P(\theta_{\text{MAP}}) + \frac{M}{2} \ln(2\pi) - \frac{1}{2} \ln |\mathbf{A}|$ – *фактор Оккама.*
- $\mathbf{A} = -\nabla\nabla \ln p(D \mid \theta_{\text{MAP}}) p(\theta_{\text{MAP}}) = -\nabla\nabla \ln p(\theta_{\text{MAP}} \mid D).$

Сравнение моделей по Лапласу

- Получаем

$$\ln p(D) \approx \ln p(D \mid \theta_{\mathrm{MAP}}) + \ln P(\theta_{\mathrm{MAP}}) + \frac{M}{2}\ln(2\pi) - \frac{1}{2}\ln|\mathbf{A}|.$$

- Если гауссовское априорное распределение $p(\theta)$ достаточно широкое, и $\mathbf{A}$ полного ранга, то можно грубо приблизить (докажите это!)

$$\ln p(D) \approx \ln p(D \mid \theta_{\mathrm{MAP}}) - \frac{1}{2}M\ln N,$$

  где M – число параметров, N – число точек в D, а аддитивные константы мы опустили.
- Это *байесовский информационный критерий* (Bayesian information criterion, BIC), он же *критерий Шварца* (Schwarz criterion).

Спасибо!

Спасибо за внимание!