ЭКВИВАЛЕНТНОЕ ЯДРО И СРАВНЕНИЕ МОДЕЛЕЙ

Сергей Николенко

СПбГУ — Санкт-Петербург
23 сентября 2020 г.

Random facts:

- 23 сентября — День празднования бисексуальности (Celebrate Bisexuality Day), впервые отмечавшийся в 1999 году
- 23 сентября в Японии — День осеннего равноденствия, государственный праздник, отмечающийся в память о предках и усопших
- 23 сентября 1784 г. Людовик XVI издал указ о том, чтобы все носовые платки были квадратными
- 23 сентября 1846 г. в Берлинской обсерватории Иоганн Готтфрид Галле, руководствуясь указаниями Урбена Леверье, обнаружил Нептун
- 23 сентября 1862 г. граф Лев Толстой женился на Софье Андреевне Берс; за 27 лет у них родилось 13 детей
- 23 сентября 1873 г. на Одесской улице Санкт-Петербурга зажглись первые опытные электрические фонари с угольными лампами накаливания системы Лодыгина

Эквивалентное ядро и сравнение моделей

Эквивалентное ядро

- Вспомним наши байесовские предсказания:

$$p(t \mid \mathbf{t}, \alpha, \beta) = N(t \mid \mu_N^\top \phi(\mathbf{x}), \sigma_N^2),$$
$$\text{где } \sigma_N^2 = \frac{1}{\beta} + \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}).$$

- Давайте перепишем среднее апостериорного распределения в другой форме (вспомним, что $\mu_N = \beta \Sigma_N \Phi^\top \mathbf{t}$):

$$y(\mathbf{x}, \mu_N) = \mu_N^\top \phi(\mathbf{x}) = \beta \phi(\mathbf{x})^\top \Sigma_N \Phi^\top \mathbf{t} =$$
$$= \sum_{n=1}^{N} \beta \phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}_n) t_n.$$

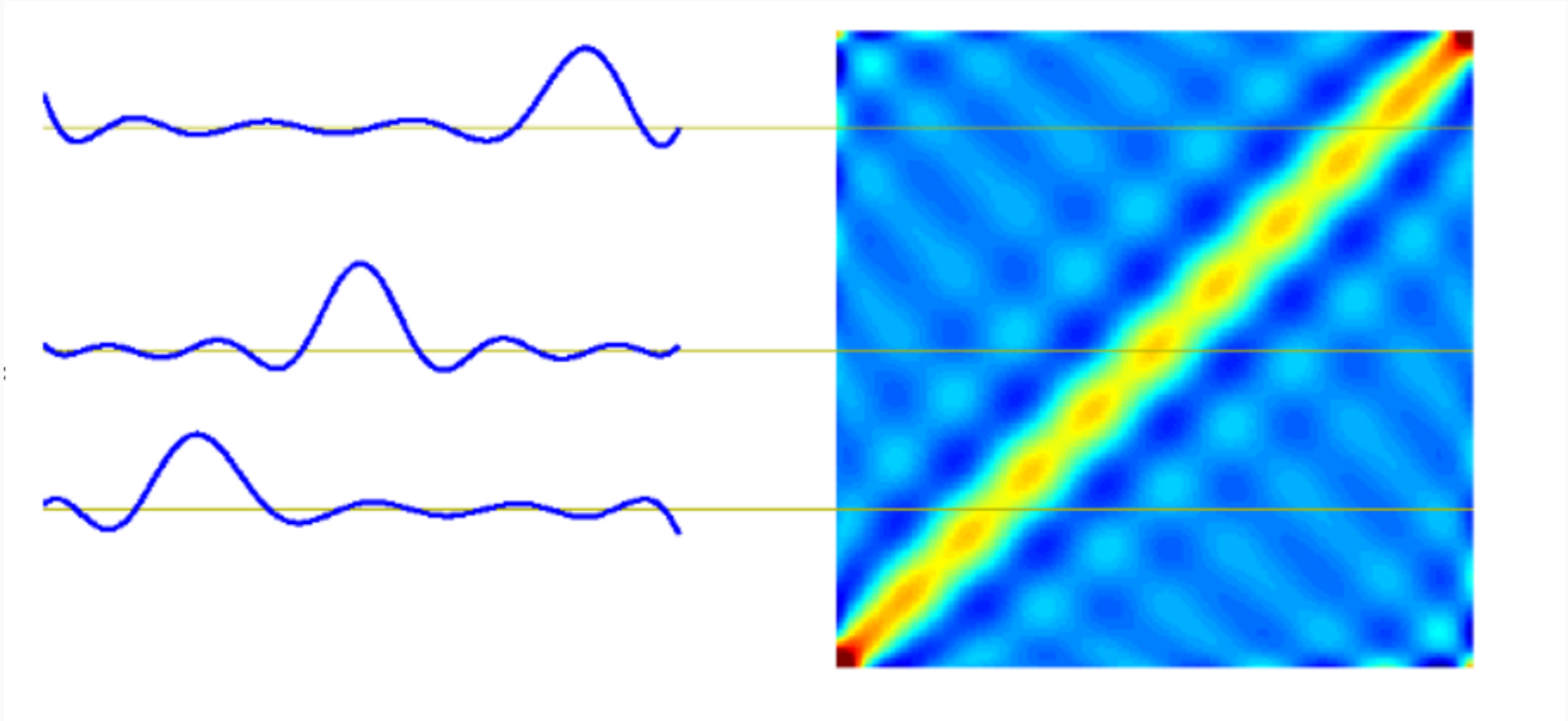
Эквивалентное ядро

- $y(\mathbf{x}, \mu_N) = \sum_{n=1}^{N} \beta\phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}_n) t_n$.
- Это значит, что предсказание можно переписать как

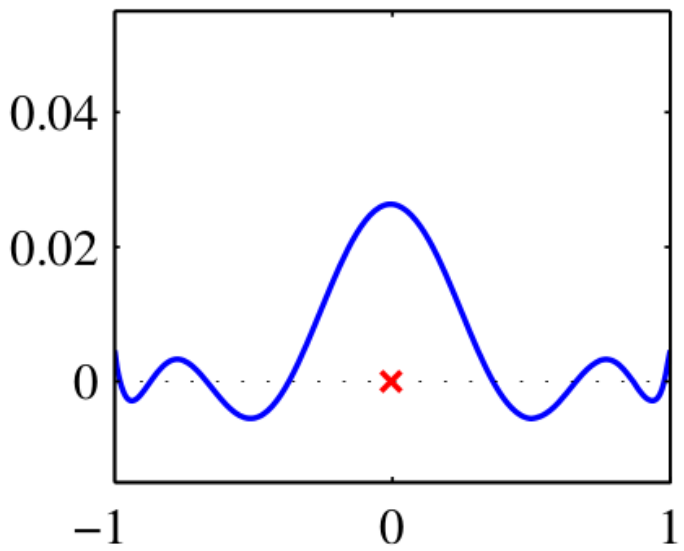
$$y(\mathbf{x}, \mu_N) = \sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) t_n.$$

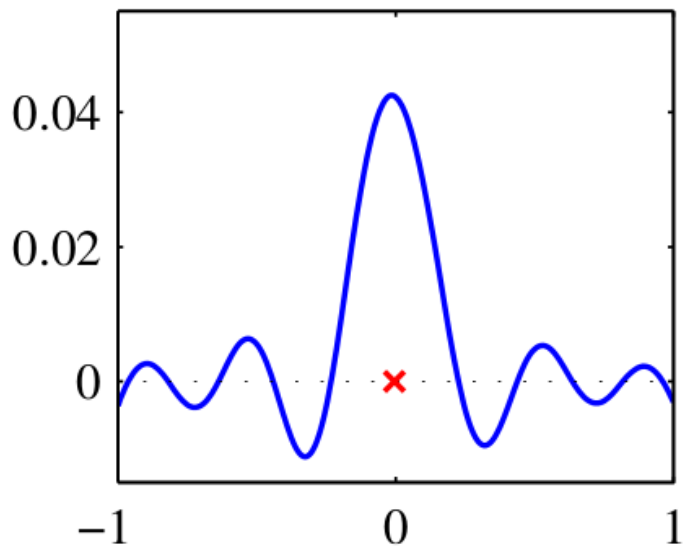
- Т.е. мы предсказываем следующую точку как линейную комбинацию значений в известных точках.
- Функция $k(\mathbf{x}, \mathbf{x}') = \beta\phi(\mathbf{x})^\top \Sigma_N \phi(\mathbf{x}')$ называется *эквивалентным ядром* (equivalent kernel).

Эквивалентное ядро



Эквивалентное ядро





Выводы про эквивалентное ядро

- Эквивалентное ядро $k(\mathbf{x}, \mathbf{x}')$ локализовано вокруг $\mathbf{x}$ как функция $\mathbf{x}'$, т.е. каждая точка оказывает наибольшее влияние около себя и затухает потом.
- Можно было бы с самого начала просто определить ядро и предсказывать через него, безо всяких базисных функций ϕ – такой подход мы ещё будем рассматривать.

Упражнение. Докажите, что $\sum_{n=1}^{N} k(\mathbf{x}, \mathbf{x}_n) = 1$.

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

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

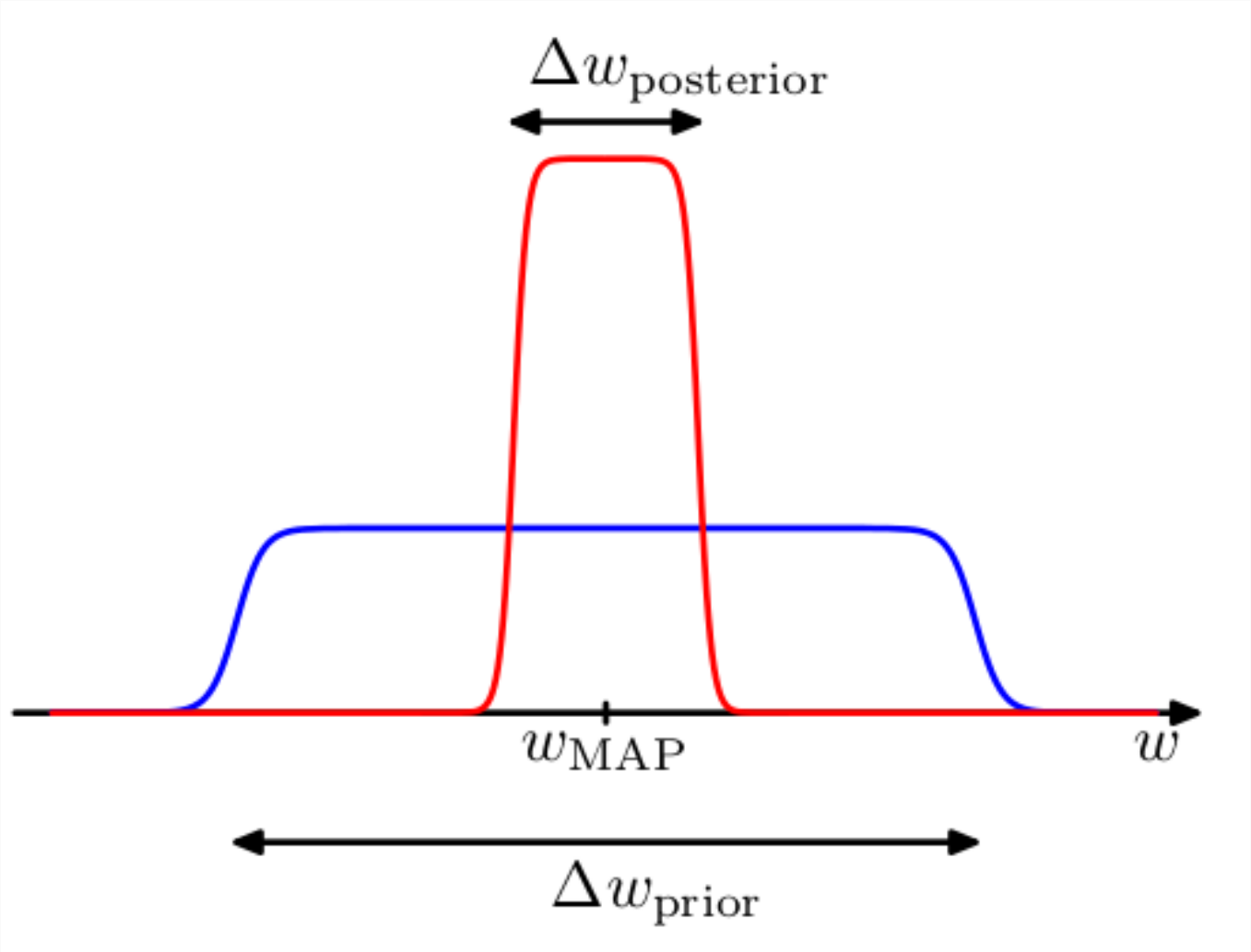
- Т.е. это вероятность сгенерировать D, если выбирать параметры модели по её априорному распределению, а потом накидывать данные.
- Это, кстати, в точности знаменатель из теоремы Байеса:

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

Байесовское сравнение моделей

- Предположим, что у модели один параметр w, а апостериорное распределение – это острый пик вокруг w_{MAP} шириной $\Delta w_{\text{posterior}}$.
- 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, $p(w) = \frac{1}{\Delta w_{\text{prior}}}$.

Приближение $p(D)$



Приближение $p(D)$

- Тогда получится

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

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

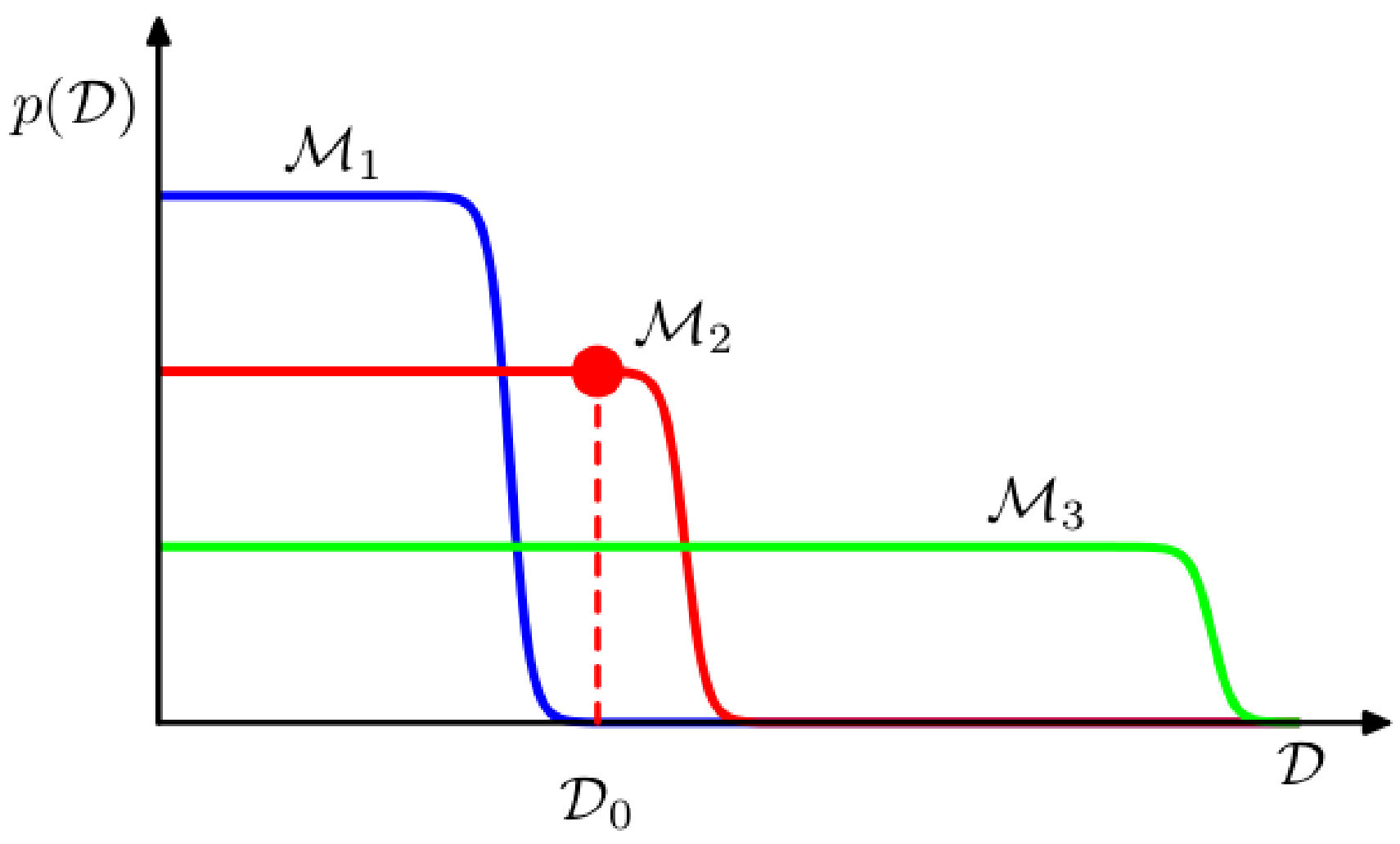
- Это значит, что мы добавляем штраф за «слишком узкое» апостериорное распределение – то есть в точности штраф за оверфиттинг!
- Для модели из M параметров, если предположить, что у них одинаковые $\Delta w_{\text{posterior}}$, получим

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

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

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

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

Пример: линейная регрессия и коронавирус

- Пример: линейная регрессия и коронавирус
- Давайте посмотрим на код...

Спасибо!

Спасибо за внимание!