

БАЙЕСОВСКОЕ СРАВНЕНИЕ МОДЕЛЕЙ

Сергей Николенко

СПбГУ — Санкт-Петербург

31 октября 2024 г.

Random facts:



- 31 октября во всём мире — Хэллоуин, ночь накануне Дня всех святых, а в России — День работников СИЗО и тюрем; 31 октября 1963 года были созданы первые СИЗО в СССР
- «31 октября» (Club Deportivo 31 de Octubre) — боливийский футбольный клуб из Ла Паса; 31 октября 1952 г. в Боливии национализировали горнорудную промышленность, а 31 октября 1954 г. там открылся съезд Первого национального конгресса трудящихся
- 31 октября 1517 г. Мартин Лютер прибил к двери церкви в Виттенберге 95 тезисов
- 31 октября 1811 г. был открыт Царскосельский лицей; среди первых воспитанников были не только Пушкин с Горчаковым, но и, например, полярный исследователь и адмирал Фёдор Матюшкин, тверской губернатор Александр Бакунин и российский посланник в Бразилии, Португалии и Нидерландах Сергей Ломоносов
- 31 октября 1905 г. была провозглашена Марковская республика, крестьянское самоуправление в Марковской волости Волоколамского уезда Московской губернии; казаки прекратили существование республики только в июне 1906 года
- 31 октября 2000 г. был остановлен последний компьютер, работавший под управлением операционной системы Multics, первый выпуск которой состоялся в 1965 году

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.

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

$$p(\mathcal{M}_i | D) \propto p(\mathcal{M}_i)p(D | \mathcal{M}_i).$$

- Если знать апостериорное распределение, то можно сделать предсказание:

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

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

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

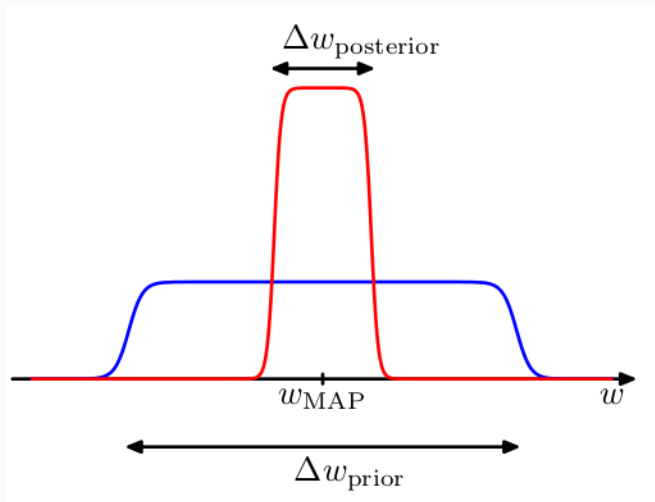
$$p(D | \mathcal{M}_i) = \int p(D | \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} | \mathcal{M}_i) d\mathbf{w}.$$

- 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:

$$p(\mathbf{w} | \mathcal{M}_i, D) = \frac{p(D | \mathbf{w}, \mathcal{M}_i) p(\mathbf{w} | \mathcal{M}_i)}{p(D | \mathcal{M}_i)}.$$

- Предположим, что у модели один параметр w , а апостериорное распределение – это острый пик вокруг w_{MAP} шириной $\Delta w_{\text{posterior}}$.
- Тогда можно приблизить $p(D) = \int p(D | w)p(w)dw$ 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

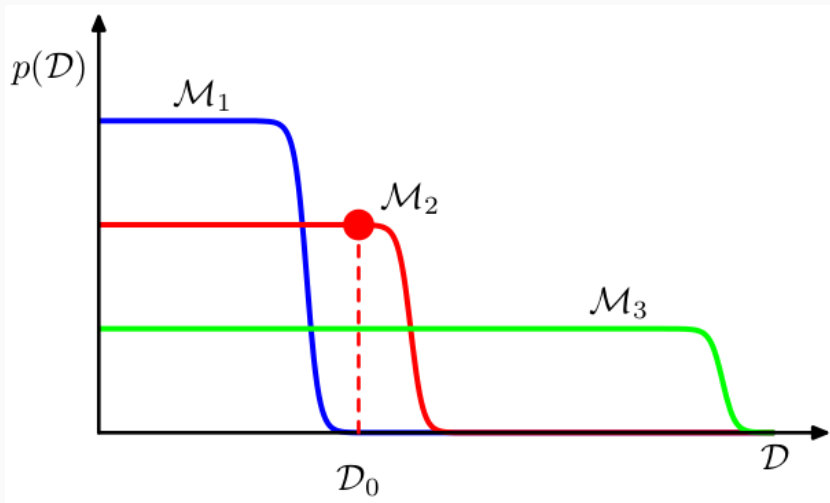
$$p(D) = \int p(D | w)p(w)dw \approx p(D | w_{\text{MAP}}) \frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}},$$
$$\ln p(D) \approx \ln p(D | w_{\text{MAP}}) + \ln \left(\frac{\Delta w_{\text{posterior}}}{\Delta w_{\text{prior}}} \right).$$

- 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

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

- 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 $p(D | \mathcal{M})$.
- Сложная модель (e.g., многочлен девятой степени) генерирует «много» разных датасетов, у неё низкая $p(D | \mathcal{M})$.
- Но сложная может хорошо выразить датасеты, которые не может выразить простая; поэтому в сумме надо выбирать «среднюю».

ПРИБЛИЖЕНИЕ $p(\mathcal{D})$



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

- ...то получится

$$\mathbb{E} \left[\ln \frac{p(D | \mathcal{M}_{\text{true}})}{p(D | \mathcal{M})} \right] = \int p(D | \mathcal{M}_{\text{true}}) \ln \frac{p(D | \mathcal{M}_{\text{true}})}{p(D | \mathcal{M})} dD.$$

- Это называется *расстоянием Кульбака-Лейблера* (Kullback-Leibler divergence) между распределениями $p(D | \mathcal{M}_{\text{true}})$ и $p(D | \mathcal{M})$.
- Но это только самый грубый результат о сравнении моделей...

БАЙЕСОВСКИЙ ИНФОРМАЦИОННЫЙ КРИТЕРИЙ

- Мы хотим сравнить несколько моделей $\mathcal{M}_1, \dots, \mathcal{M}_K$ с наборами параметров $\theta_1, \dots, \theta_K$ на наборе данных D , т.е. сравнить между собой $p(\mathcal{M}_k|D)$:

$$p(\mathcal{M}_k|D) \propto p(\mathcal{M}_k) p(D|\mathcal{M}_k).$$

- Будем полагать $p(\mathcal{M}_k)$ равномерными. А $p(D|\mathcal{M}_k)$ — это как раз знаменатель теоремы Байеса:

$$p(\theta_k|D, \mathcal{M}_k) = \frac{p(\theta_k|\mathcal{M}_k) p(D|\theta_k, \mathcal{M}_k)}{p(D|\mathcal{M}_k)}.$$

- Нам нужно оценить интеграл

$$p(D) = \int p(\theta) p(\theta|D) d\theta = \int p(\theta) e^{\ell(\theta)} d\theta,$$

где $\ell(\theta) = \log p(\theta|D)$.

- Применим лапласовскую аппроксимацию в окрестности точки максимума правдоподобия θ_{ML} :

$$\ell(\theta) \approx \ell(\theta_{\text{ML}}) - \frac{N}{2} (\theta - \theta_{\text{ML}})^\top J(\theta_{\text{ML}}) (\theta - \theta_{\text{ML}}),$$

где

$$J(\theta_{\text{ML}}) = -\frac{1}{N} \frac{\partial^2 \log p(\theta|D)}{\partial \theta \partial \theta^\top} \Big|_{\theta_{\text{ML}}}.$$

- Аналогічно можна розкласти априорне розподілення окрестности θ_{ML} , но там не пропадєть член первого порядка, поэтому давайте им и ограничимся:

$$p(\theta) \approx p(\theta_{\text{ML}}) + (\theta - \theta_{\text{ML}})^{\top} \nabla_{\theta} p(\theta)|_{\theta_{\text{ML}}}.$$

- Итого получается, что

$$p(D) \approx \int \left(p(\theta_{\text{ML}}) + (\theta - \theta_{\text{ML}})^{\top} \nabla_{\theta} p(\theta)|_{\theta_{\text{ML}}} \right) \times \\ \times e^{\ell(\theta_{\text{ML}}) - \frac{N}{2} (\theta - \theta_{\text{ML}})^{\top} J(\theta_{\text{ML}}) (\theta - \theta_{\text{ML}})} d\theta.$$

- Но теперь можно заметить, что

$$\int (\theta - \theta_{\text{ML}}) e^{-\frac{N}{2}(\theta - \theta_{\text{ML}})^\top J(\theta_{\text{ML}})(\theta - \theta_{\text{ML}})} d\theta = 0,$$

потому что это величина, пропорциональная матожиданию $(\theta - \theta_{\text{ML}})$ по гауссиану со средним $(\theta - \theta_{\text{ML}})$ и матрицей ковариаций $J(\theta_{\text{ML}})^{-1}$.

- А значит, наша аппроксимация превращается в

$$p(D) \approx e^{\ell(\theta_{\text{ML}})} p(\theta_{\text{ML}}) \int e^{-\frac{N}{2}(\theta - \theta_{\text{ML}})^\top J(\theta_{\text{ML}})(\theta - \theta_{\text{ML}})} d\theta.$$

- Интеграл теперь можно взять — из него получится нормировочная константа для того же самого гауссиана:

$$p(D) \approx e^{\ell(\theta_{\text{ML}})} p(\theta_{\text{ML}}) (2\pi)^{\frac{d}{2}} N^{-\frac{d}{2}} (\det J(\theta_{\text{ML}}))^{-\frac{1}{2}}, \quad \text{или}$$

$$\log p(D) \approx \ell(\theta_{\text{ML}}) - \frac{d}{2} \log N + \log p(\theta_{\text{ML}}) - \frac{1}{2} \log \det J(\theta_{\text{ML}}) + \frac{d}{2} \log(2\pi).$$

- Выбросим всё, что не растёт с N , и умножим на -2 ; получится *байесовский информационный критерий* (Bayesian information criterion, BIC), он же *критерий Шварца* (Schwartz criterion):

$$\text{BIC}(\mathcal{M}) = -2 \log p(D|\theta_{\text{ML}}, \mathcal{M}) + d \log N,$$

где d — это размерность вектора θ , или число свободных параметров в модели $\mathcal{M}$.

ЭМПИРИЧЕСКИЙ БАЙЕС

- Откуда берутся гиперпараметры?
- Оказывается, их тоже можно оптимизировать!
- У линейной регрессии, например, два гиперпараметра:
 $\beta = \frac{1}{\sigma^2}$ и α (точность регуляризатора, пусть гребневого).
- Давайте просто попробуем оптимизировать $p(D \mid \alpha, \beta)$ (marginal likelihood).

- Получается:

$$p(D \mid \alpha, \beta) = \int p(\mathbf{w})p(D \mid \mathbf{w})d\mathbf{w},$$
$$\ln p(D \mid \alpha, \beta) = \left(\frac{\beta}{2\pi}\right)^{\frac{N}{2}} \left(\frac{\alpha}{2\pi}\right)^{\frac{d}{2}} \int e^{-\frac{\beta}{2}\|\mathbf{y}-X\mathbf{w}\|^2 - \frac{\alpha}{2}\mathbf{w}^\top\mathbf{w}}d\mathbf{w}.$$

- Выделяем полный квадрат так же, как раньше:

$$A = \beta X^\top X + \alpha \mathbf{I},$$
$$\mu_N = \beta A^{-1} X^\top \mathbf{y}.$$

- Теперь

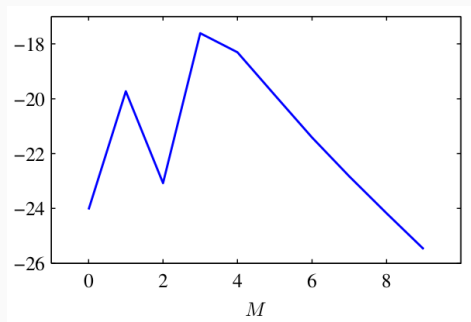
$$\int e^{-\frac{1}{2}(\mathbf{w}-\mu_N)^\top A(\mathbf{w}-\mu_N)} d\mathbf{w} = (2\pi)^{\frac{d}{2}} \sqrt{\det A^{-1}}.$$

- Получается:

$$\ln p(D \mid \alpha, \beta) = \frac{d}{2} \ln \alpha + \frac{N}{2} \ln \beta - \frac{\beta}{2} \|\mathbf{y} - X\mu_N\|^2 - \frac{\alpha}{2} \mu_N^\top \mu_N - \frac{1}{2} \ln \det A - \frac{N}{2} \ln(2\pi).$$

- Это теперь надо максимизировать по α и β , а можно и разные d перебирать, если речь идёт о том, как выбрать оптимальное число признаков.

- Пример графика по числу параметров:



- А как оптимизировать?

- Обозначим через λ_i собственные числа матрицы $\beta \mathbf{X}^\top \mathbf{X}$:

$$(\beta \mathbf{X}^\top \mathbf{X}) \mathbf{u}_i = \lambda_i \mathbf{u}_i$$

для некоторых собственных векторов $\mathbf{u}_i$.

- Тогда $\mathbf{u}_i$ будут являться и собственными векторами матрицы $\mathbf{A}$, с собственными числами $\alpha + \lambda_i$:

$$\mathbf{A} \mathbf{u}_i = (\beta \mathbf{X}^\top \mathbf{X} + \alpha \mathbf{I}) \mathbf{u}_i = \lambda_i \mathbf{u}_i + \alpha \mathbf{u}_i.$$

- Теперь будем оптимизировать логарифм маргинального правдоподобия $\log p(\mathbf{y}|\mathbf{X}, \alpha, \beta)$, взяв производную по α :

$$\frac{\partial \log \det \mathbf{A}}{\partial \alpha} = \frac{\partial \log \prod_i (\alpha + \lambda_i)}{\partial \alpha} = \sum_i \frac{\partial \log(\alpha + \lambda_i)}{\partial \alpha} = \sum_i \frac{1}{\alpha + \lambda_i},$$

и вся производная будет равна

$$\frac{\partial \log p(\mathbf{y}|\mathbf{X}, \alpha, \beta)}{\partial \alpha} = \frac{M}{2\alpha} - \frac{1}{2} \mu_N^\top \mu_n - \frac{1}{2} \sum_i \frac{1}{\alpha + \lambda_i}.$$

- Приравняем производную нулю:

$$\alpha \mu_N^\top \mu_N = M - \alpha \sum_i \frac{1}{\alpha + \lambda_i} = \sum_i \frac{\lambda_i}{\alpha + \lambda_i}.$$

- Получается итеративный процесс

$$\alpha^{(k+1)} = \frac{1}{\mu_N(\alpha^{(k)})^\top \mu_N(\alpha^{(k)})} \sum_i \frac{\lambda_i}{\alpha^{(k)} + \lambda_i}.$$

- Аналогично по β : можно заметить, что

$$\frac{\partial \lambda_i}{\partial \beta} = \frac{\lambda_i}{\beta}.$$

- Значит,

$$\frac{\partial \log \det \mathbf{A}}{\partial \beta} = \frac{\partial \log \prod_i (\alpha + \lambda_i)}{\partial \beta} = \sum_i \frac{\partial \log(\alpha + \lambda_i)}{\partial \beta} = \frac{1}{\beta} \sum_i \frac{\lambda_i}{\alpha + \lambda_i}, \text{ и}$$

$$\frac{\partial \log p(\mathbf{y}|\mathbf{X}, \alpha, \beta)}{\partial \beta} = \frac{N}{2\beta} - \frac{1}{2} \sum_{n=1}^N (t_n - \mu_n^\top \mathbf{x}_n)^2 - \frac{1}{2\beta} \sum_i \frac{\lambda_i}{\alpha + \lambda_i}.$$

- Итого по β получаем

$$\beta^{(k+1)} = \frac{N - \sum_i \frac{\lambda_i}{\alpha^{(k)} + \lambda_i}}{\sum_{n=1}^N \left(t_n - \mu_n \left(\alpha^{(k)}, \beta^{(k)} \right)^\top \mathbf{x}_n \right)^2}.$$

СПАСИБО!

Спасибо за внимание!

