БАЙЕСОВСКИЙ ВЫВОД ДЛЯ ГАУССИАНА

Сергей Николенко

образование

Академия больших данных MADE — VK

11 марта 2022 г.

Random facts:

- 11 марта 105 г. евнух Цай Лунь подал доклад об усовершенствовании технологии производства бумаги, научившись делать её из бамбука
- 11 марта 843 г. — «Торжество Православия»: императрица Феодора праздновала окончание Константинопольского собора, восстановившего иконопочитание
- 11 марта 1811 г. в Ноттингеме началось восстание луддитов, которые разрушили множество шерстяных и хлопкообрабатывающих фабрик
- 11 марта 1878 г. фонограф Эдисона демонстрировался «бессмертным» парижской Академии; когда из коробки раздался голос, профессор-филолог Буйо вскочил с кресла, подбежал к пригласившему инженеров физику Монселю, схватил его за воротник и в ярости стал душить, повторяя: «Негодяй! Плут! Вы думаете, что мы позволим чревовещателю надувать нас?!»
- 11 марта 1931 г. в СССР был введён физкультурный комплекс «Готов к труду и обороне СССР» (ГТО), а также запрещены продажа и ввоз Библии
- 11 марта 2020 г. Всемирная организация здравоохранения (ВОЗ) объявила, что вспышка болезни, вызванной коронавирусом нового типа (COVID-19), является пандемией

Логистическая регрессия

Несколько классов

- В случае нескольких классов

$$p(C_k \mid \phi) = y_k(\phi) = \frac{e^{a_k}}{\sum_j e^{a_j}} \text{ для } a_k = \mathbf{w}_k^\top \phi.$$

- Опять выпишем максимальное правдоподобие; во-первых,

$$\frac{\partial y_k}{\partial a_j} = y_k \left([k = j] - y_j \right).$$

Несколько классов

- Теперь запишем правдоподобие – для схемы кодирования 1-of-K будет целевой вектор $\mathbf{t}_n$ и правдоподобие

$$p(\mathbf{T} \mid \mathbf{w}_1, \dots, \mathbf{w}_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k \mid \phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

для $y_{nk} = y_k(\phi_n)$; берём логарифм:

$$E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\ln p(\mathbf{T} \mid \mathbf{w}_1, \dots, \mathbf{w}_K) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}, \text{ и}$$

$$\nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\sum_{n=1}^{N} \left(y_{nj} - t_{nj}\right) \phi_n.$$

Несколько классов

- Оптимизировать опять можно по Ньютону-Рапсону; гессиан получится как

$$\nabla_{\mathbf{w}_k} \nabla_{\mathbf{w}_j} E(\mathbf{w}_1, \dots, \mathbf{w}_K) = -\sum_{n=1}^{N} y_{nk} \left([k = j] - y_{nj} \right) \phi_n \phi_n^\top.$$

Пробит-регрессия

- А что если у нас другая форма сигмоида?
- Мы по-прежнему в той же постановке: два класса, $p(t = 1 \mid a) = f(a)$, $a = \mathbf{w}^\top \phi$, f – функция активации.
- Давайте установим функцию активации с порогом θ: для каждого ϕ_n, вычисляем $a_n = \mathbf{w}^\top \phi_n$, и

$$\begin{cases} t_n = 1, & \text{если } a_n \geq \theta, \\ t_n = 0, & \text{если } a_n < \theta. \end{cases}$$

Пробит-регрессия

- Если θ берётся по распределению $p(\theta)$, это соответствует

$$f(a) = \int_{-\infty}^{a} p(\theta) \mathrm{d}\theta.$$

- Пусть, например, $p(\theta)$ – гауссиан с нулевым средним и единичной дисперсией. Тогда

$$f(a) = \Phi(a) = \int_{-\infty}^{a} N\left(\theta \mid 0, 1\right) \mathrm{d}\theta.$$

Пробит-регрессия

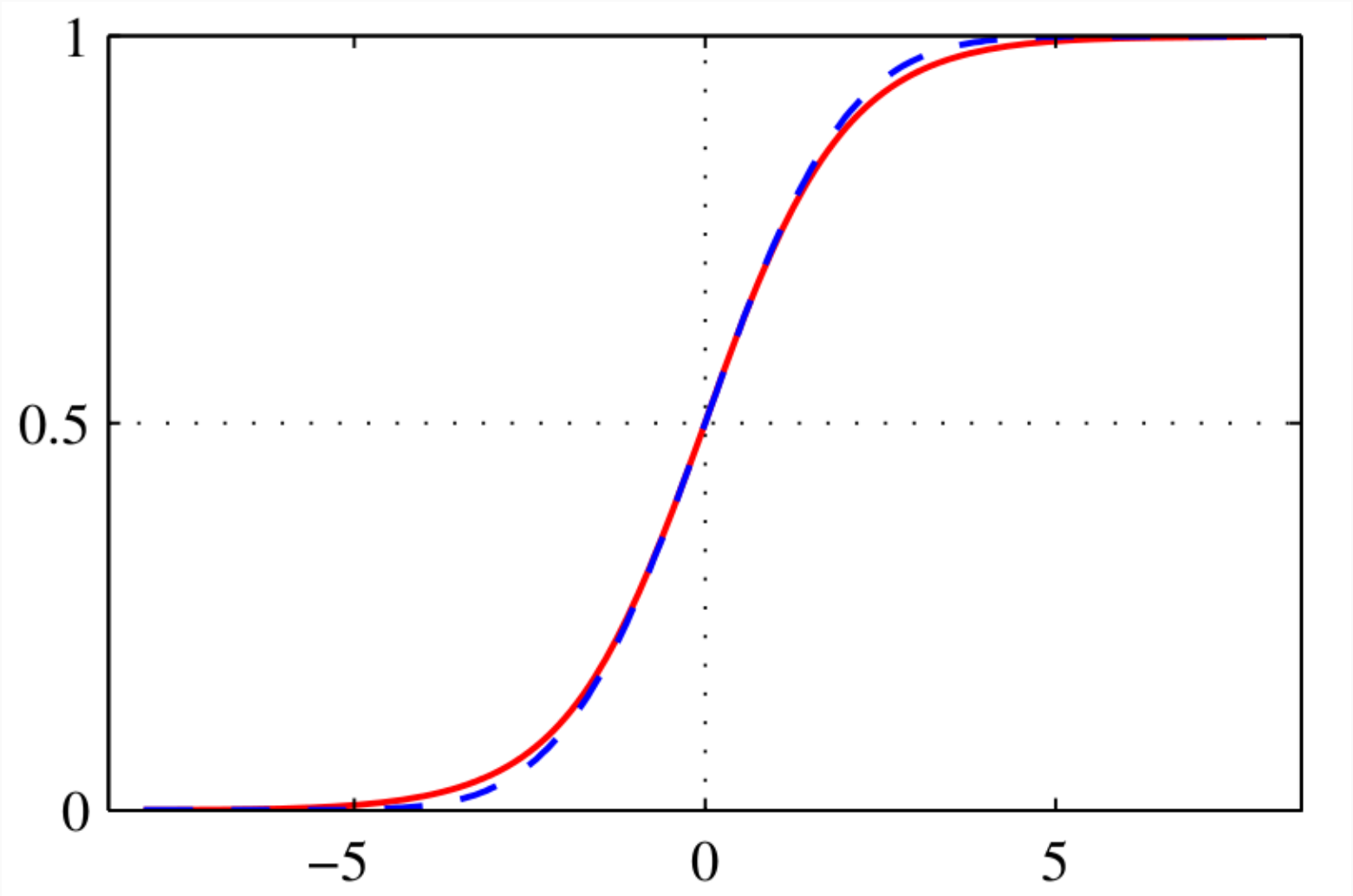
- Это называется *пробит-функцией* (probit); неэлементарная, но тесно связана с

$$\operatorname{erf}(a) = \frac{2}{\sqrt{\pi}} \int_0^a e^{-\frac{\theta^2}{2}} \mathrm{d}\theta :$$

$$\Phi(a) = \frac{1}{2}\left[1 + \frac{1}{\sqrt{2}}\operatorname{erf}(a)\right].$$

- Пробит-регрессия – это модель с пробит-функцией активации.

σ и Φ



Лапласовская аппроксимация и байесовская логистическая регрессия

Лапласовская аппроксимация

- Небольшое лирическое отступление: как приблизить сложное распределение простым?
- Например, как приблизить гауссианом возле максимума? (естественная задача)
- Рассмотрим пока распределение от одной непрерывной переменной $p(z) = \frac{1}{Z} f(z)$.

Лапласовская аппроксимация

- Первый шаг: найдём максимум z_0.
- Второй шаг: разложим в ряд Тейлора

$$\ln f(z) \approx \ln f(z_0) - \frac{1}{2} A(z - z_0)^2, \text{ где } A = -\frac{d^2}{dz^2} \ln f(z) \mid_{z=z_0} .$$

- Третий шаг: приблизим

$$f(z) \approx f(z_0) e^{-\frac{A}{2}(z-z_0)^2},$$

и после нормализации это будет как раз гауссиан.

Лапласовская аппроксимация

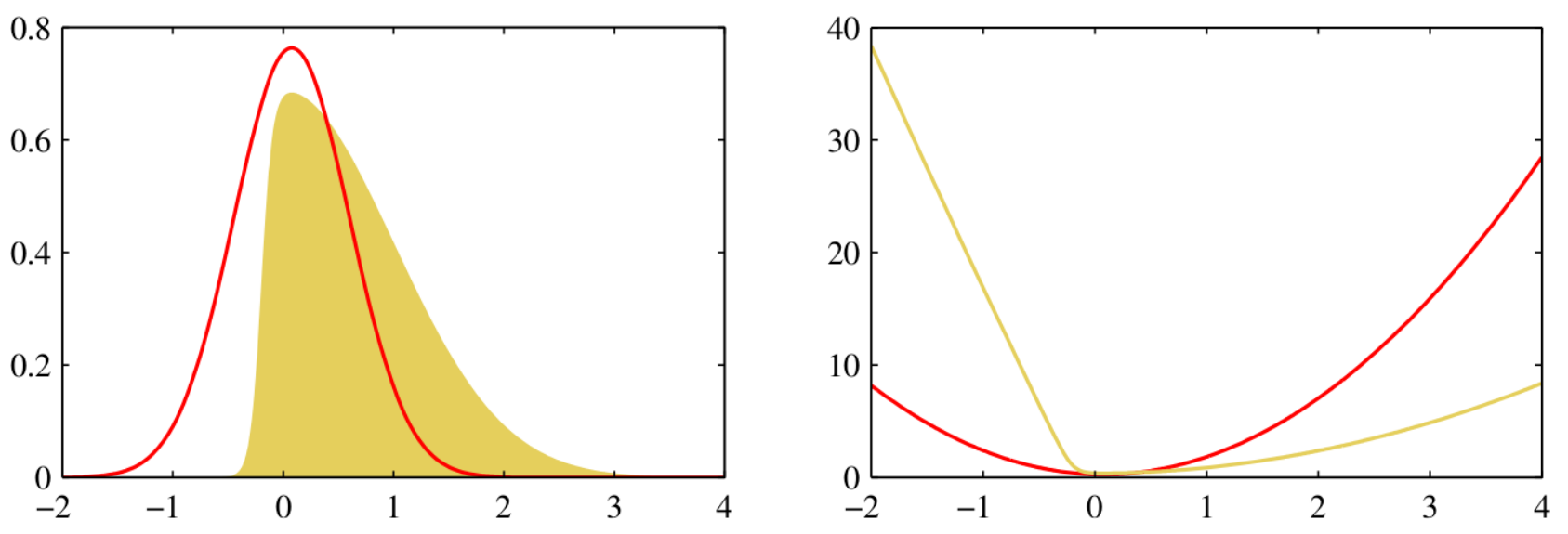
- Это можно обобщить на многомерное распределение $p(\mathbf{z}) = \frac{1}{Z} f(\mathbf{z})$:

$$f(\mathbf{z}) \approx f(\mathbf{z}_0) e^{-\frac{1}{2}(\mathbf{z}-\mathbf{z}_0)^\top \mathbf{A} (\mathbf{z}-\mathbf{z}_0)},$$

$$\text{где } \mathbf{A} = -\nabla\nabla \ln f(\mathbf{z}) \,|_{z=z_0} \,.$$

Упражнение. Какая здесь будет нормировочная константа?

Лапласовская аппроксимация



Байесовская логистическая регрессия

- Теперь давайте обработаем логистическую регрессию по-байесовски.
- Логистическую регрессию так просто не выпишешь, как линейную – точного ответа из произведения логистических сигмоидов не получается.
- Будем приближать по Лапласу.

Байесовская логистическая регрессия

- Априорное распределение выберем гауссовским:

$$p(\mathbf{w}) = N(\mathbf{w} \mid \mu_0, \Sigma_0).$$

- Тогда апостериорное будет

$$\begin{aligned} p(\mathbf{w} \mid \mathbf{t}) \propto & p(\mathbf{w}) p(\mathbf{t} \mid \mathbf{w}), \text{ и} \\ \ln p(\mathbf{w} \mid \mathbf{t}) = & -\frac{1}{2} \left(\mathbf{w} - \mu_0\right)^\top \Sigma_0^{-1} \left(\mathbf{w} - \mu_0\right) \\ & + \sum_{n=1}^{N} \left[t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\right] + \text{const}, \\ \text{где } y_n = & \sigma(\mathbf{w}^\top \phi_n). \end{aligned}$$

Байесовская логистическая регрессия

- Чтобы приблизить, сначала находим максимум $\mathbf{w}_{\text{MAP}}$, а потом матрица ковариаций – это матрица вторых производных

$$\Sigma_N = -\nabla\nabla \ln p(\mathbf{w} \mid \mathbf{t}) = \Sigma_0^{-1} + \sum_{n=1}^{N} y_n(1 - y_n)\phi_n\phi_n^\top.$$

- Наше приближение – это

$$q(\mathbf{w}) = N(\mathbf{w} \mid \mathbf{w}_{\text{MAP}}, \Sigma_N).$$

Байесовская логистическая регрессия

- Теперь можно описать байесовское предсказание:

$$p(C_1 \mid \phi, \mathbf{t}) = \int p(C_1 \mid \phi, \mathbf{w}) p(\mathbf{w} \mid \mathbf{t}) d\mathbf{w} \approx \int \sigma(\mathbf{w}^\top \phi) q(\mathbf{w}) \mathrm{d}\mathbf{w}.$$

- Заметим, что $\sigma(\mathbf{w}^\top \phi)$ зависит от $\mathbf{w}$ только через его проекцию на ϕ.
- Обозначим $a = \mathbf{w}^\top \phi$:

$$\sigma(\mathbf{w}^\top \phi) = \int \delta(a - \mathbf{w}^\top \phi) \sigma(a) \mathrm{d}a.$$

Байесовская логистическая регрессия

- $\sigma(\mathbf{w}^\top \phi) = \int \delta(a - \mathbf{w}^\top \phi)\sigma(a)\mathrm{d}a$, а значит,

$$\int \sigma(\mathbf{w}^\top \phi) q(\mathbf{w}) d\mathbf{w} = \int \sigma(a) p(a) \mathrm{d}a,$$

$$\text{где } p(a) = \int \delta(a - \mathbf{w}^\top \phi) q(\mathbf{w}) \mathrm{d}\mathbf{w}.$$

- $p(a)$ – это маргинализация гауссиана $q(\mathbf{w})$, где мы интегрируем по всему, что ортогонально ϕ.

Байесовская логистическая регрессия

- $p(a)$ – это маргинализация гауссиана $q(\mathbf{w})$, где мы интегрируем по всему, что ортогонально ϕ.
- Значит, $p(a)$ – тоже гауссиан; найдём его моменты:

$$\begin{aligned}\mu_a =& \mathbf{E}[a] = \int a p(a) \mathrm{d}a = \int q(\mathbf{w}) \mathbf{w}^\top \phi \mathrm{d}\mathbf{w} = \mathbf{w}_{\mathrm{MAP}}^\top \phi, \\ \sigma_a^2 =& \int \left(a^2 - \mathbf{E}[a]\right)^2 p(a) \mathrm{d}a = \\ =& \int q(\mathbf{w}) \left[(\mathbf{w}^\top \phi)^2 - (\mu_N^\top \phi)^2\right]^2 \mathrm{d}\mathbf{w} = \phi^\top \Sigma_N \phi.\end{aligned}$$

- Итого получили, что

$$p(C_1 \mid \mathbf{t}) = \int \sigma(a) p(a) \mathrm{d}a = \int \sigma(a) N(a \mid \mu_a, \sigma_a^2) \mathrm{d}a.$$

Байесовская логистическая регрессия

- $p(C_1 \mid \mathbf{t}) = \int \sigma(a) N(a \mid \mu_a, \sigma_a^2) \mathrm{d}a.$
- Этот интеграл так просто не взять, потому что сигмоид сложный, но можно приблизить, если приблизить $\sigma(a)$ через пробит: $\sigma(a) \approx \Phi(\lambda a)$ для $\lambda = \sqrt{\pi/8}$.

Упражнение. Докажите, что для $\lambda = \sqrt{\pi/8}$ у σ и Φ одинаковый наклон в нуле.

Байесовская логистическая регрессия

- А если мы перейдём к пробит-функции, то её свёртка с гауссианом будет просто другим пробитом:

$$\int \Phi(\lambda a) N(a \mid \mu, \sigma^2) \mathrm{d}a = \Phi\left(\frac{\mu}{\sqrt{\frac{1}{\lambda^2} + \sigma^2}}\right).$$

Упражнение. Докажите это.

Байесовская логистическая регрессия

- В итоге получается аппроксимация

$$\int \sigma(a) N(a \mid \mu, \sigma^2) \mathrm{d}a \approx \sigma\left(\kappa(\sigma^2)\mu\right),$$

$$\text{где } \kappa(\sigma^2) = \frac{1}{\sqrt{1 + \frac{\pi}{8}\sigma^2}}.$$

Байесовская логистическая регрессия

- И теперь, собирая всё вместе, мы получили распределение предсказаний:

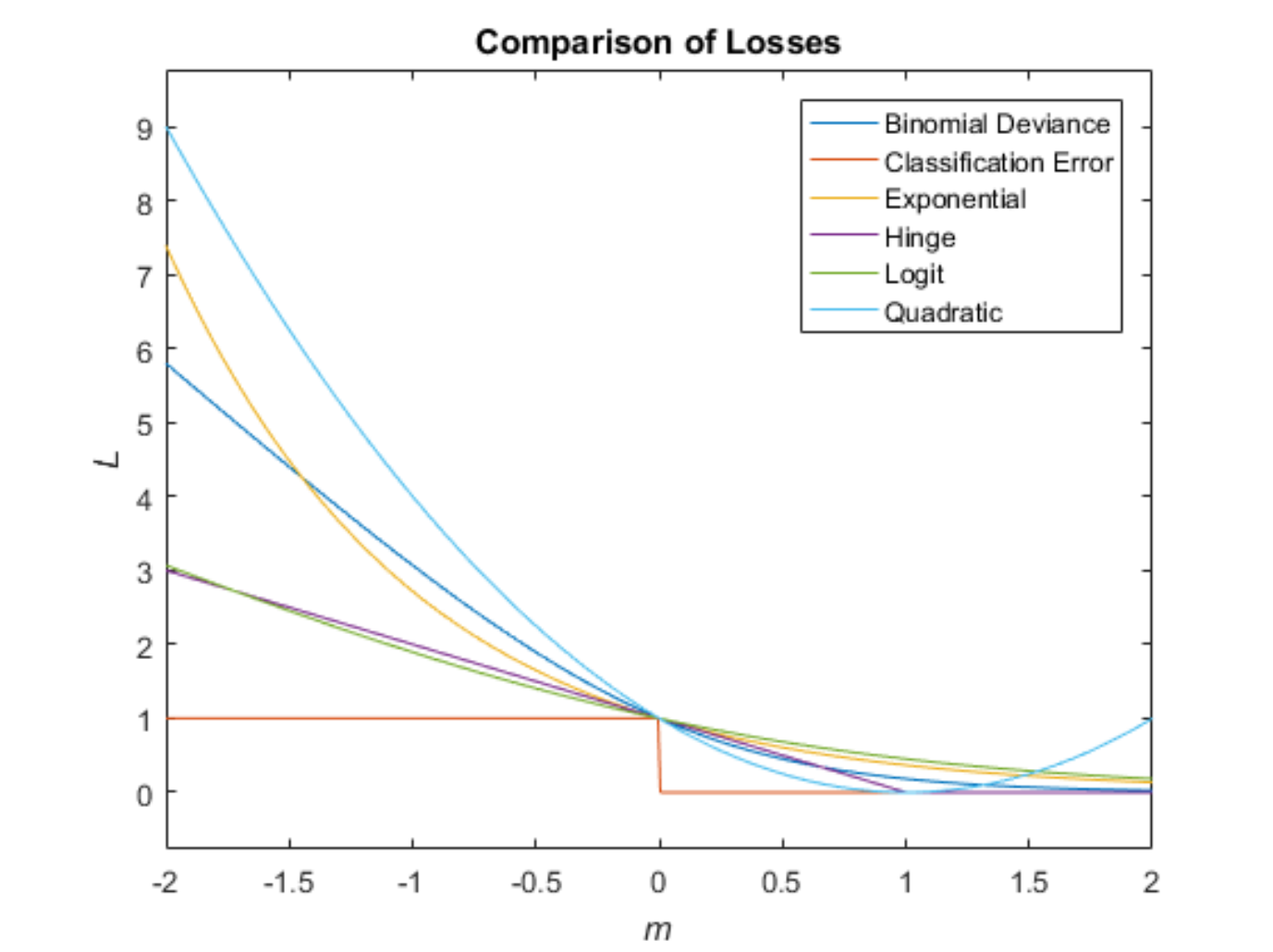
$$
\begin{aligned}
p(C_1 \mid \phi, \mathbf{t}) &= \sigma\left(\kappa(\sigma_a^2)\mu_a\right), \text{ где} \\
\mu_a &= \mathbf{w}_{\mathrm{MAP}}^\top \phi, \\
\sigma_a^2 &= \phi^\top \Sigma_N \phi, \\
\kappa(\sigma^2) &= \frac{1}{\sqrt{1 + \frac{\pi}{8}\sigma^2}}.
\end{aligned}
$$

- Кстати, разделяющая поверхность $p(C_1 \mid \phi, \mathbf{t}) = \frac{1}{2}$ задаётся уравнением $\mu_a = 0$, и тут нет никакой разницы с просто использованием $\mathbf{w}_{\mathrm{MAP}}$. Разница будет только для более сложных критериев.

Функции потерь в классификации

- И напоследок немножко другой взгляд: разные методы классификации отличаются друг от друга тем, какую функцию ошибки они оптимизируют.
- У классификации проблема с «правильной» функцией ошибки, то есть ошибкой собственно классификации:
  - она и не везде дифференцируема,
  - и производная её никому не нужна.
- Давайте посмотрим на разные функции потерь (loss functions); мы уже несколько видели, но ещё немало осталось.

Функции потерь в классификации



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

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

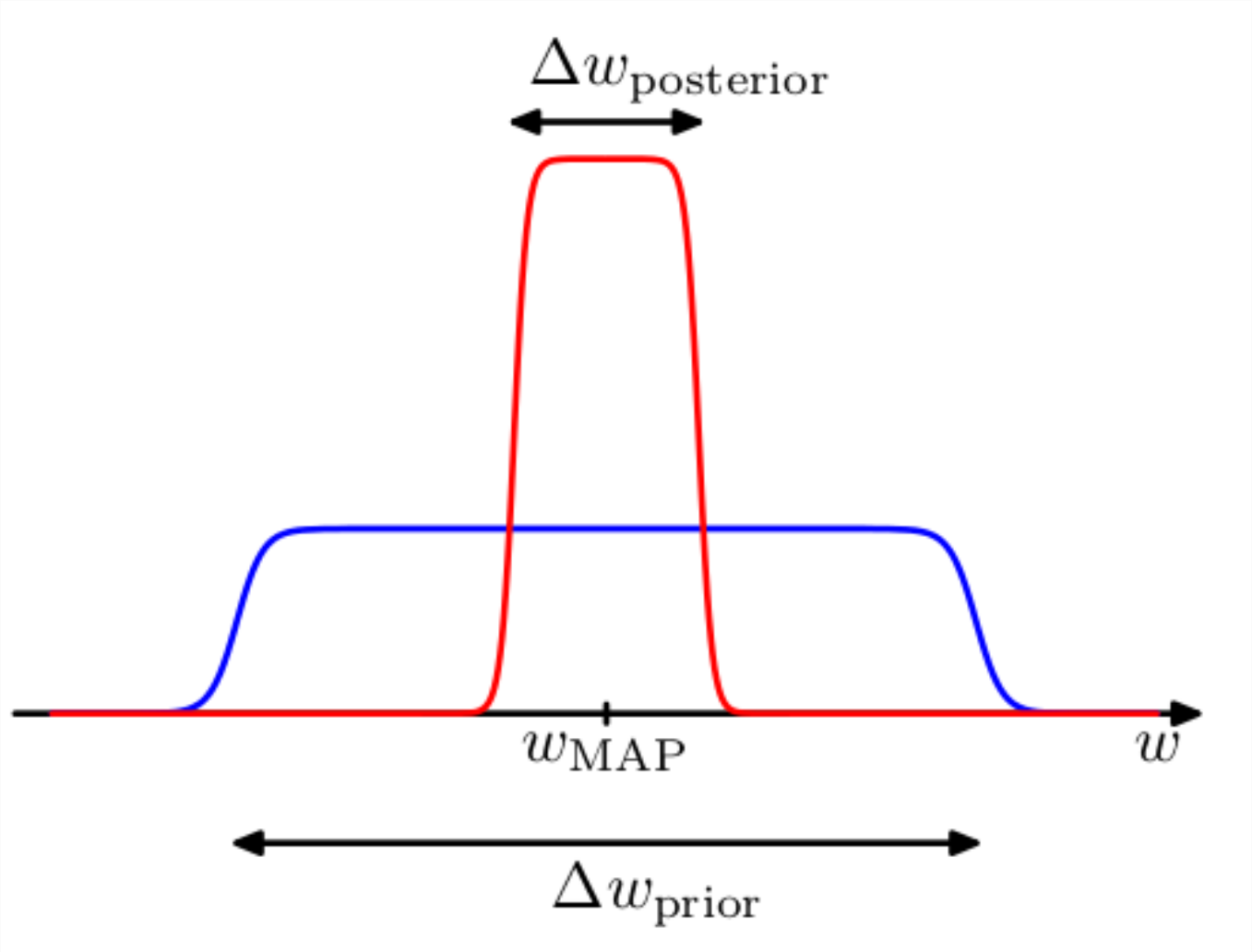
- Т.е. это вероятность сгенерировать D, если выбирать параметры модели по её априорному распределению, а потом накидывать данные.
- Это, кстати, в точности знаменатель из теоремы Байеса:

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

Байесовское сравнение моделей

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

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

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

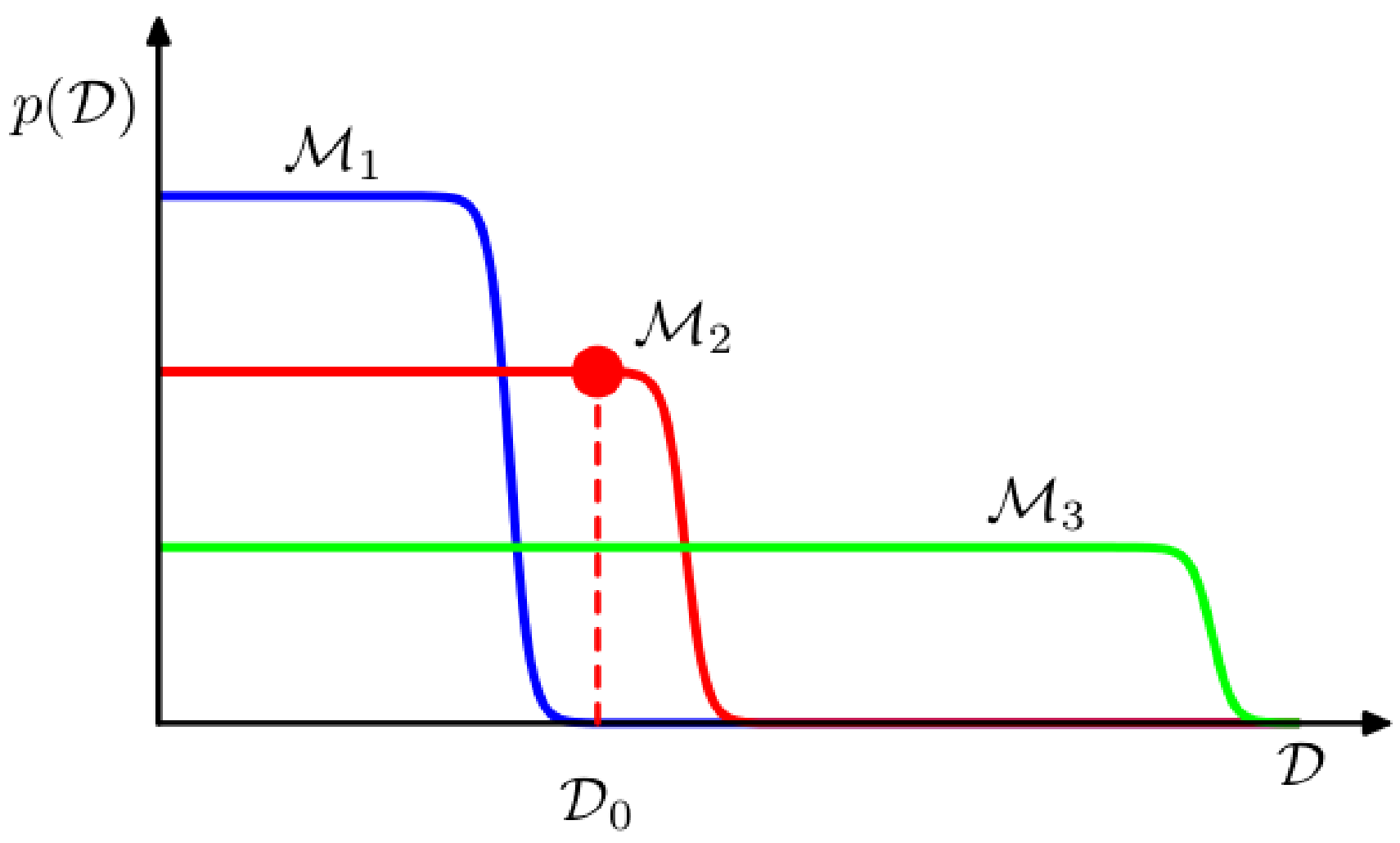
- Это значит, что мы добавляем штраф за «слишком узкое» апостериорное распределение – то есть в точности штраф за оверфиттинг!
- Для модели из M параметров, если предположить, что у них одинаковые $\Delta w_{\text{posterior}}$, получим

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

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

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

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

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

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

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

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

  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).

Байесовский информационный критерий

BIC в руки

- Мы хотим сравнить несколько моделей $M_1, \dots, M_K$ с наборами параметров $\theta_1, \dots, \theta_K$ на наборе данных D, т.е. сравнить между собой $p\left(M_k|D\right)$:

$$p\left(M_k|D\right) \propto p\left(M_k\right) p\left(D|M_k\right).$$

- Будем полагать $p\left(M_k\right)$ равномерными. А $p\left(D|M_k\right)$ — это как раз знаменатель теоремы Байеса:

$$p\left(\theta_k|D, M_k\right) = \frac{p\left(\theta_k|M_k\right) p\left(D|\theta_k, M_k\right)}{p\left(D|M_k\right)}.$$

BIC в руки

- Нам нужно оценить интеграл

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

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

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

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

где

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

BIC в руки

- Аналогично можно разложить априорное распределение окрестности θ_{ML}, но там не пропадёт член первого порядка, поэтому давайте им и ограничимся:

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

- Итого получается, что

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

BIC в руки

- Но теперь можно заметить, что

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

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

- А значит, наша аппроксимация превращается в

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

BIC в руки

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

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

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

- 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):

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

где d — это размерность вектора θ, или число *свободных параметров* в модели M.

Информационный критерий Акаике

Идея и вывод AIC

- Пусть данные $X = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$ были получены из истинного распределения $p_{\text{data}}(\mathbf{x})$, а мы пытаемся приблизить их некоторой параметрической моделью $p(\mathbf{x}|\theta)$, $\theta \in \mathbb{R}^d$.
- Предположим, что мы обучили модель методом максимального правдоподобия, получив $p(\mathbf{x}|\theta_{\text{ML}})$.
- Давайте попробуем оценить, насколько модель $p(\mathbf{x}|\theta_{\text{ML}})$ отличается от неизвестного истинного распределения $p_{\text{data}}(\mathbf{x})$:

$$\begin{aligned}\text{KL}\left(p_{\text{data}} \| p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right) &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log \frac{p_{\text{data}}(\mathbf{x})}{p\left(\mathbf{x}|\theta_{\text{ML}}\right)}\right] = \\ &= \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p_{\text{data}}(\mathbf{x})\right] - \mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right].\end{aligned}$$

Идея и вывод AIC

- Модель будет тем лучше, чем больше будет $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right]$, и для оценки расхождения между $p\left(\mathbf{x}|\theta_{\text{ML}}\right)$ и $p_{\text{data}}(\mathbf{x})$ нужно получить оценку ожидаемого логарифма правдоподобия.
- Во всех критериях важен логарифм правдоподобия в точке его максимума, ведь это как раз выборочная оценка ожидания:

$$\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right] = \int p_{\text{data}}(\mathbf{x}) \log p\left(\mathbf{x}|\theta_{\text{ML}}\right) \mathrm{d}\mathbf{x} \approx$$

$$\approx \frac{1}{N} \sum_{n=1}^{N} \log p\left(\mathbf{x}_n|\theta_{\text{ML}}\right).$$

- Но это смещённая оценка как минимум потому, что мы обучаем параметры максимального правдоподобия θ_{ML} на том же датасете $\mathbf{X}$, который используется в этой оценке.

Идея и вывод AIC

- Если истинная модель p_{data} тоже из семейства $p(\mathbf{x}|\theta)$ с некоторым истинным параметром θ_0, то

$$\theta_0 = \arg\max_\theta \mathbb{E}_{p_{\text{data}}(\mathbf{x})} \left[\log p(\mathbf{x}|\theta)\right],$$

  но это ожидание берётся по всему распределению.
- θ_0 — это «истинная» гипотеза максимального правдоподобия; при некоторых условиях регулярности можно доказать, что:
  - $\theta_{\text{ML}}(\mathbf{X}) \to \theta_0$ при $N \to \infty$;
  - для $\theta_{\text{ML}}(\mathbf{X})$ верна асимптотическая нормальность, т.е. распределение величины $\sqrt{N}(\theta_{\text{ML}} - \theta_0)$ сходится по вероятности к распределению $N(0, I(\theta_0)^{-1})$, где $I(\theta)$ — это матрица информации Фишера

$$I(\theta) = \int p(\mathbf{x}|\theta) \frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta} \frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta^\top} \mathrm{d}\mathbf{x}.$$

Идея и вывод AIC

- Более того, те формулы предполагали, что $p_{\text{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$, но аналогичные результаты можно получить и если $p_{\text{data}}(\mathbf{x})$ не принадлежит параметрическому семейству $p\left(\mathbf{x}|\theta\right)$.
- Пусть θ_0 — максимум ожидания логарифма правдоподобия по p_{data}, то есть решение системы

$$\int p_{\text{data}}(\mathbf{x}) \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta} \mathrm{d}\mathbf{x} = 0.$$

Идея и вывод AIC

- Тогда при тех же условиях можно доказать, что:
  - $\theta_{\mathrm{ML}}(\mathbf{X}) \to \theta_0$ при $N \to \infty$;
  - распределение величины $\sqrt{N}\,(\theta_{\mathrm{ML}} - \theta_0)$ сходится по вероятности к нормальному распределению

$$\sqrt{N}\,(\theta_{\mathrm{ML}} - \theta_0) \to_{N\to\infty} N\left(0, J^{-1}(\theta_0) I(\theta_0) J^{-1}(\theta_0)\right),$$

    где $I(\theta)$ — это та же матрица информации Фишера, только по распределению p_{data}:

$$I(\theta) = \int p_{\mathrm{data}}(\mathbf{x}) \frac{\partial \log p\,(\mathbf{x}|\theta)}{\partial \theta} \frac{\partial \log p\,(\mathbf{x}|\theta)}{\partial \theta^{\top}} \mathrm{d}\mathbf{x};$$

    а $J(\theta)$ — это ожидание матрицы вторых производных

$$J(\theta) = -\int p_{\mathrm{data}}(\mathbf{x}) \frac{\partial^2 \log p\,(\mathbf{x}|\theta)}{\partial \theta \partial \theta^{\top}} \mathrm{d}\mathbf{x}.$$

Идея и вывод AIC

- Иначе говоря, на позиции (i, j) у матрицы $I(\theta)$ стоит ожидание произведения $\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_i}$ и $\frac{\partial \log p(\mathbf{x}|\theta)}{\partial \theta_j}$, а у матрицы $J(\theta)$ — ожидание второй производной $\frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_i \partial \theta_j}$.
- И если всё-таки $p_{\text{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$, то

$$\frac{\partial^2 \log p\left(\mathbf{x}|\theta\right)}{\partial \theta_i \partial \theta_j} = \frac{\partial}{\partial \theta_i}\left(\frac{1}{p\left(x|\theta\right)}\frac{\partial p\left(\mathbf{x}|\theta\right)}{\partial \theta_j}\right) = \\ = \frac{1}{p\left(x|\theta\right)}\frac{\partial^2 p\left(\mathbf{x}|\theta\right)}{\partial \theta_i \partial \theta_j} - \frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta_i}\frac{\partial \log p\left(\mathbf{x}|\theta\right)}{\partial \theta_j},$$

а в ожидании по $p_{\text{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$ при подстановке $\theta = \theta_0$

$$\mathbb{E}_{p(\mathbf{x}|\theta_0)}\left[\frac{1}{p\left(\mathbf{x}|\theta_0\right)}\frac{\partial^2 p\left(\mathbf{x}|\theta_0\right)}{\partial \theta_i \partial \theta_j}\right] = \int \frac{p\left(\mathbf{x}|\theta_0\right)}{p\left(\mathbf{x}|\theta_0\right)}\frac{\partial^2 p\left(\mathbf{x}|\theta_0\right)}{\partial \theta_i \partial \theta_j}\mathrm{d}\mathbf{x} = \\ = \frac{\partial^2}{\partial \theta_i \partial \theta_j}\int p\left(\mathbf{x}|\theta_0\right) = 0, \quad \text{то есть здесь} \quad I(\theta_0) = J(\theta_0).$$

Идея и вывод AIC

- Различные информационные критерии для сравнения моделей оценивают смещение выборочной оценки для величины $\mathbb{E}_{p_{\text{data}}(\mathbf{x})}\left[\log p\left(\mathbf{x}|\theta_{\text{ML}}\right)\right]$

$$b(p_{\text{data}}) = \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_{\text{ML}}\right) - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\text{ML}}\right)\right]\right],$$

  где мы взяли ожидание по датасетам $\mathbf{X} \sim p_{\text{data}}$.
- Если мы сможем оценить смещение $b(p_{\text{data}})$, то информационный критерий можно будет построить, умножив на -2 аналогично BIC:

$$\begin{aligned}\text{IC}\left(\mathbf{X}, \theta\right) = \\ = -2\left(\text{логарифм правдоподобия } \mathbf{X} \text{ в } \theta_{\text{ML}} - \text{оценка смещения}\right) = \\ = -2\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_{\text{ML}}\right) + 2\left(\text{оценка } b(p_{\text{data}})\right).\end{aligned}$$

Идея и вывод AIC

- Давайте попробуем оценить $b(p_{\text{data}})$:

$$
\begin{aligned}
b(p_{\text{data}}) &= \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_{\text{ML}}\right) - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\text{ML}}\right)\right]\right] = \\
&= \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_{\text{ML}}\right) - \log p\left(\mathbf{X}|\theta_0\right)\right] + \\
&+ \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_0\right) - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right]\right] + \\
&+ \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right] - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\text{ML}}\right)\right]\right] = \\
&= B_1 + B_2 + B_3.
\end{aligned}
$$

- Будем оценивать слагаемые по отдельности.

Идея и вывод AIC

- Проще всего оценить B_2, потому что в нём нет θ_{ML}:

$$B_2 = \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\log p\left(\mathbf{X}|\theta_0\right) - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right]\right] =$$

$$= \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_0\right)\right] - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right] = 0.$$

- Это не значит, что B_2 всегда равно нулю; для конкретного датасета $\mathbf{X}$ значение B_2 будет ненулевым, но в ожидании получится ноль.

Идея и вывод AIC

- Чтобы оценить B_3, рассмотрим функцию $\eta(\theta_{\mathrm{ML}}) = \mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\mathrm{ML}}\right)\right]$ и разложим её по формуле Тейлора в окрестности точки θ_0 (её максимума):

$$\eta(\theta_{\mathrm{ML}}) \approx \eta(\theta_0) - \frac{1}{2}\left(\theta_{\mathrm{ML}} - \theta_0\right)^{\top} J(\theta_0)\left(\theta_{\mathrm{ML}} - \theta_0\right),$$

где

$$J(\theta_0) = -\mathbb{E}_{p_{\mathrm{data}}(\mathbf{z})}\left[\left.\frac{\partial^2 \log p\left(\mathbf{z}|\theta\right)}{\partial\theta\partial\theta^{\top}}\right|_{\theta_0}\right] =$$

$$= -\int p_{\mathrm{data}}(\mathbf{z})\left.\frac{\partial^2 \log p\left(\mathbf{z}|\theta\right)}{\partial\theta\partial\theta^{\top}}\right|_{\theta_0} \mathrm{d}\mathbf{z}.$$

Идея и вывод AIC

- А B_3 — это ожидание $\eta(\theta_0) - \eta(\theta_{\text{ML}})$ по распределению $p_{\text{data}}(\mathbf{X})$:

$$B_3 = \mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_0\right)\right] - N\mathbb{E}_{p_{\text{data}}(\mathbf{z})}\left[\log p\left(\mathbf{z}|\theta_{\text{ML}}\right)\right]\right] =$$
$$= \frac{N}{2}\mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\left(\theta_{\text{ML}} - \theta_0\right)^\top J(\theta_0)\left(\theta_{\text{ML}} - \theta_0\right)\right] =$$
$$= \frac{N}{2}\mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\text{Tr}\left(J(\theta_0)\left(\theta_{\text{ML}} - \theta_0\right)\left(\theta_{\text{ML}} - \theta_0\right)^\top\right)\right] =$$
$$= \frac{N}{2}\text{Tr}\left(J(\theta_0)\mathbb{E}_{\mathbf{X}\sim p_{\text{data}}}\left[\left(\theta_{\text{ML}} - \theta_0\right)\left(\theta_{\text{ML}} - \theta_0\right)^\top\right]\right).$$

- Теперь можно вместо ожидания матрицы ковариаций по датасету $\mathbf{X}$ подставить асимптотический результат:

$$B_3 = \frac{N}{2}\text{Tr}\left(J(\theta_0)\frac{1}{N}J(\theta_0)^{-1}I(\theta_0)J(\theta_0)^{-1}\right) = \frac{1}{2}\text{Tr}\left(I(\theta_0)J(\theta_0)^{-1}\right).$$

Идея и вывод AIC

- Для оценки B_1 нужно провернуть аналогичный трюк с $\ell(\theta) = \log p\left(X|\theta\right)$, разложив его вокруг своего максимума θ_{ML}:

$$\ell(\theta) = \ell(\theta_{\mathrm{ML}}) + \frac{1}{2}\left(\theta - \theta_{\mathrm{ML}}\right)^{\top} \left.\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\top}}\right|_{\theta_{\mathrm{ML}}} \left(\theta - \theta_{\mathrm{ML}}\right).$$

- Мы знаем, что $\theta_{\mathrm{ML}} \to \theta_0$ при $N \to \infty$; а по закону больших чисел можно получить, что

$$-\frac{1}{N}\left.\frac{\partial^2 \ell(\theta)}{\partial\theta\partial\theta^{\top}}\right|_{\theta} = -\frac{1}{N}\sum_{n=1}^{N}\left.\frac{\partial^2 \log p\left(\mathbf{x}_n|\theta\right)}{\partial\theta\partial\theta^{\top}}\right|_{\theta_0} \to J(\theta_0).$$

Идея и вывод AIC

- Следовательно, в нашу оценку можно подставить

$$\ell(\theta_{\mathrm{ML}}) - \ell(\theta_0) \approx -\frac{N}{2} \left(\theta - \theta_{\mathrm{ML}}\right)^\top J(\theta_0) \left(\theta - \theta_{\mathrm{ML}}\right).$$

- А затем и оценить B_1 так же, как оценивали B_3:

$$\begin{aligned} B_1 &= \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}} \left[\log p\left(\mathbf{X}|\theta_{\mathrm{ML}}\right) - \log p\left(\mathbf{X}|\theta_0\right)\right] = \\ &= \frac{N}{2} \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}} \left[\left(\theta - \theta_{\mathrm{ML}}\right)^\top J(\theta_0) \left(\theta - \theta_{\mathrm{ML}}\right)\right] = \\ &= \frac{N}{2} \mathrm{Tr}\left(J(\theta_0) \mathbb{E}_{\mathbf{X}\sim p_{\mathrm{data}}} \left[\left(\theta - \theta_{\mathrm{ML}}\right)^\top \left(\theta - \theta_{\mathrm{ML}}\right)\right]\right), \end{aligned}$$

то есть

$$B_3 = \frac{1}{2} \mathrm{Tr}\left(I(\theta_0) J(\theta_0)^{-1}\right). \tag{1}$$

Идея и вывод AIC

- Осталось только объединить три оценки:

$$b(p_{\text{data}}) = B_1 + B_2 + B_3 = \operatorname{Tr}\left(I(\theta_0)J(\theta_0)^{-1}\right).$$

- $I(\theta_0)$ и $J(\theta_0)$ нам неизвестны, т.к. зависят от p_{data}; если взять оценки $\hat{I}$ и $\hat{J}$, это приведёт нас к *информационному критерию Такеучи* (Takeuchi information criterion, TIC):

$$\text{TIC} = -2\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_{\text{ML}}\right) + 2\text{Tr}\left(\hat{I}\hat{J}^{-1}\right).$$

- В качестве $\hat{I}$ и $\hat{J}$ можно подставить просто усреднённые значения по датасету в точке максимума правдоподобия:

$$\hat{I}_{i,j} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial \log p\left(\mathbf{x}_n|\theta\right)}{\partial\theta_i}\frac{\partial \log p\left(\mathbf{x}_n|\theta\right)}{\partial\theta_j}\bigg|_{\theta_{\text{ML}}}, \quad \hat{J}_{i,j} = \frac{1}{N}\sum_{n=1}^{N}\frac{\partial^2 \log p\left(\mathbf{x}_n|\theta\right)}{\partial\theta_i\partial\theta_j}\bigg|_{\theta_{\text{ML}}}.$$

Идея и вывод AIC

- А если можно всё-таки предполагать, что истинное распределение данных p_{data} лежит в параметрическом семействе $p\left(\mathbf{x}|\theta\right)$, то есть $p_{\text{data}}(\mathbf{x}) = p\left(\mathbf{x}|\theta_0\right)$, то, как мы обсуждали выше, $I(\theta_0) = J(\theta_0)$, и информационный критерий Такеучи превращается в *информационный критерий Акаике* (Akaike information criterion, AIC):

$$\text{AIC} = -2\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_{\text{ML}}\right) + 2\text{Tr}\left(I_d\right) =$$

$$= -2\sum_{n=1}^{N}\log p\left(\mathbf{x}_n|\theta_{\text{ML}}\right) + 2d.$$

- Очень простая формула!

Идея и вывод AIC

- Пример: вернёмся к полиномиальной регрессии с логарифмом правдоподобия

$$\ell(\mathbf{w}) = -\frac{N}{2}\log\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top \mathbf{x}_n\right)^2.$$

- Давайте в этом разделе для разнообразия будем дисперсию тоже обучать:

$$\mathbf{w}_{\mathrm{ML}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{t}, \quad {\sigma_{\mathrm{ML}}}^2 = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}_{\mathrm{ML}}^\top \mathbf{x}_n\right)^2.$$

- Тогда при подстановке гипотезы максимального правдоподобия получится

$$\ell(\mathbf{w}_{\mathrm{ML}}) = -\frac{N}{2}\log\left(2\pi{\sigma_{\mathrm{ML}}}^2\right) - \frac{N}{2}.$$

Идея и вывод AIC

- AIC и BIC в таком примере будут, скорее всего, выбирать примерно одну и ту же модель, хотя разница между ними всё-таки есть (см. пример)

Байесовский вывод для гауссиана

Нормальное распределение: фиксируем σ

- На самом деле всё это — байесовский вывод для нормального распределения:

$$p(x_1, \dots, x_n \mid \mu, \sigma^2) \propto \frac{1}{\sigma^n} \exp\left(-\frac{1}{2\sigma^2} \sum_i (x_i - \mu)^2\right).$$

- Хотим: найти сопряжённое априорное распределение, подсчитать правдоподобие, решить задачу предсказания.
- Для начала зафиксируем σ^2 и будем в качестве параметра рассматривать только μ.

Нормальное распределение: фиксируем σ

- Сопряжённое априорное распределение для μ при фиксированном σ^2 тоже нормальное и выглядит как

$$p(\mu \mid \mu_0, \sigma_0^2) \propto \frac{1}{\sigma_0^n} \exp\left(-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right).$$

- Обычно выбирают $\mu_0 = 0$, $\sigma_0^2 \to \infty$ (порой буквально).
- Давайте рассмотрим сначала случай ровно одного наблюдения x и найдём $p(\mu \mid x)$.

Нормальное распределение: фиксируем σ

- При нашем априорном распределении у μ и x совместное нормальное распределение:

$$x = \mu + \sigma\epsilon, \quad \mu = \mu_0 + \sigma_0\delta, \quad \epsilon, \delta \sim \mathcal{N}(0, 1).$$

Упражнение. Пусть (z_1, z_2) – случайные величины с совместным нормальным распределением. Докажите, что случайная величина $z_1 \mid z_2$ распределена нормально с параметрами

$$E(z_1 \mid z_2) = E(z_1) + \frac{Cov(z_1, z_2)}{Var(z_2)}\left(z_2 - E(z_2)\right),$$

$$Var(z_1 \mid z_2) = Var(z_1) - \frac{Cov^2(z_1, z_2)}{Var(z_2)}$$

$(Var(x) = E\left[(x - Ex)^2\right], Cov(x, y) = E\left[(x - Ex)(y - Ey)\right])$.

Нормальное распределение: фиксируем σ

- В нашем случае:

$$
\begin{aligned}
x = \mu + \sigma\epsilon, \quad \mu = \mu_0 + \sigma_0\delta, \quad \epsilon, \delta \sim \mathcal{N}(0,1),\\
E(x) = \mu_0,\\
Var(x) = E(Var(x \mid \mu)) + Var(E(x \mid \mu)) = \sigma^2 + \sigma_0^2,\\
Cov(x, \mu) = E\left[(x - \mu_0)(\mu - \mu_0)\right] = \sigma_0^2.
\end{aligned}
$$

- Применив упражнение, получаем:

$$
\begin{aligned}
E(\mu \mid x) = \mu_0 + \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}(x - \mu_0) = \frac{\sigma_0^2}{\sigma_0^2 + \sigma^2}x + \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\mu_0,\\
Var(\mu \mid x) = \frac{\sigma^2\sigma_0^2}{\sigma_0^2 + \sigma^2} = \frac{1}{\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}}.
\end{aligned}
$$

Нормальное распределение: фиксируем σ

- Итого:

$$p(\mu \mid x) \sim \mathcal{N}\left(\frac{\sigma_0^2}{\sigma_0^2 + \sigma^2} x + \frac{\sigma^2}{\sigma_0^2 + \sigma^2}\mu_0, \left(\frac{1}{\sigma_0^2} + \frac{1}{\sigma^2}\right)^{-1}\right).$$

- Опять же, сложные вычисления можно забыть и пользоваться этими формулами.
- Замечание: часто используют $\tau = \frac{1}{\sigma^2}$ как параметр нормального распределения (precision). Тогда

$$\tau_{\mu|x} = \tau_\mu + \tau.$$

Нормальное распределение: фиксируем σ

- А что, если данных больше, $x_1, \dots, x_n$?
- Тогда можно повторить всё то же самое, а можно заметить, что набор данных описывается своим средним.

Упражнение. Докажите, что если $p(x_i \mid \mu) \sim \mathcal{N}(\mu, \sigma^2)$ и x_i независимы, то $p(\bar{x} \mid \mu) \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$.

Нормальное распределение: фиксируем σ

- Для апостериорной вероятности будет

$$p(\mu \mid x_1, \dots, x_n) \propto p(x_1, \dots, x_n \mid \mu) p(\mu) \propto p(\bar{x} \mid \mu) p(\mu) \propto p(\mu \mid \bar{x}).$$

- Подставляя в наш предыдущий результат, получим:

$$p(\mu \mid x_1, \dots, x_n) \sim \mathcal{N}\left(\frac{\sigma_0^2}{\sigma_0^2 + \frac{\sigma^2}{n}} x + \frac{\sigma^2}{n\sigma_0^2 + \sigma^2}\mu_0, \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2}\right)^{-1}\right).$$

Нормальное распределение: фиксируем μ

- Если зафиксировать μ и менять σ^2, то сопряжённым априорным распределением будет обратное гамма-распределение:

$$p(\sigma^2 \mid \alpha, \beta) \propto IG(\alpha, \beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} z^{-\alpha-1} \exp\left(\frac{-\beta}{z}\right).$$

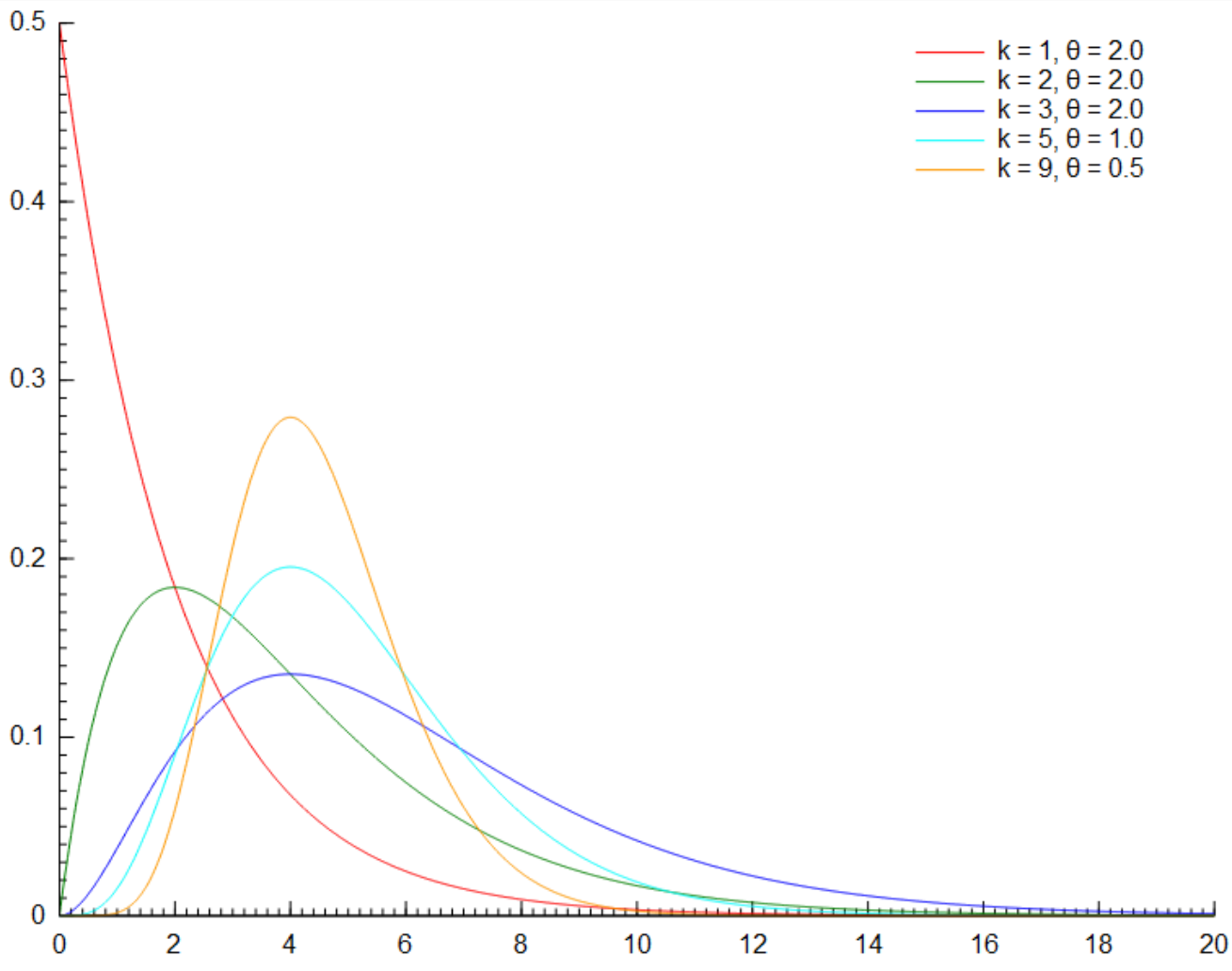
- Тогда в апостериорном распределении будет

$$p(\sigma^2 \mid x_1, \dots, x_n, \alpha, \beta) \propto IG\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i - \mu)\right).$$

- А в терминах $\tau = \frac{1}{\sigma^2}$ будет обычное гамма-распределение:

$$p(\tau \mid x_1, \dots, x_n, \alpha, \beta) \propto \text{Gamma}\left(\alpha + \frac{n}{2}, \beta + \frac{1}{2}\sum(x_i - \mu)\right).$$

Гамма--распределение



Когда и μ, и σ^2 меняются

- Что делать, когда и μ, и σ^2 меняются?
- Можно было бы предположить, что μ и σ^2 независимы; тогда просто априорное распределение будет

$$p(\mu, \sigma \mid \mu_0, \sigma_0, \alpha, \beta) \propto \mathcal{N}(\mu_0, \sigma_0^2) \cdot IG(\alpha, \beta).$$

- К сожалению, это распределение не будет сопряжённым к нормальному. Почему?

Когда и μ, и σ^2 меняются

- Что делать, когда и μ, и σ^2 меняются?
- Можно было бы предположить, что μ и σ^2 независимы; тогда просто априорное распределение будет

$$p(\mu, \sigma \mid \mu_0, \sigma_0, \alpha, \beta) \propto \mathcal{N}(\mu_0, \sigma_0^2) \cdot IG(\alpha, \beta).$$

- К сожалению, это распределение не будет сопряжённым к нормальному. Почему?
- Потому что μ и σ^2 зависимы. :) Новая точка x вводит зависимость между ними.
- В результате получается распределение Стьюдента.

Экспоненциальное семейство

- Вообще говоря, всё, о чём мы говорили – частные случаи *экспоненциального семейства* распределений:

$$p(\mathbf{x} \mid \eta) = h(\mathbf{x})g(\eta)e^{\eta^\top \mathbf{u}(\mathbf{x})}.$$

- η называются *естественными параметрами* (natural parameters).

Экспоненциальное семейство

- Например, распределение Бернулли:

$$p(x \mid \mu) = \mu^x (1-\mu)^{1-x} = e^{x \ln \mu + (1-x)\ln(1-\mu)} = \\ = (1-\mu) e^{\ln\left(\frac{\mu}{1-\mu}\right) x},$$

и естественный параметр получился $\eta = \ln\left(\frac{\mu}{1-\mu}\right)$:

$$p(x \mid \eta) = \sigma(-\eta) e^{-\eta x},$$

где $\sigma(y) = \frac{1}{1+e^{-y}}$ – *сигмоид-функция*.

Экспоненциальное семейство

- Для мультиномиального распределения с параметрами $\mu_1, \ldots, \mu_{M-1}$ получаются

$$\eta_k = \ln\left(\frac{\mu_k}{1 - \sum_j \mu_j}\right) \quad \text{и}$$

$$p(\mathbf{x} \mid \eta) = \left(1 + \sum_{k=1}^{M-1} e^{\eta_k}\right)^{-1} e^{\eta^\top \mathbf{x}}.$$

Упражнение. Проверьте!

Экспоненциальное семейство

- Так вот, для распределений из экспоненциального семейства

$$p(\mathbf{x} \mid \eta) = h(\mathbf{x})g(\eta)e^{\eta^\top \mathbf{u}(\mathbf{x})}$$

можно сразу оптом найти сопряжённые априорные распределения:

$$p(\eta \mid \chi, \nu) = f(\chi, \nu)g(\eta)^\nu e^{\nu\eta^\top \chi},$$

где χ – гиперпараметры, а g то же самое, что в исходном распределении.

Упражнение. Проверьте это и получите вышеописанные примеры как частные случаи.

Когда и μ, и σ^2 меняются

- В настоящем сопряжённом априорном распределении будут:

$$\begin{aligned} x \mid \mu, \tau &\sim \mathcal{N}(\mu, \tau), \\ \mu \mid \tau &\sim \mathcal{N}(\mu_0, n_0\tau), \\ \tau &\sim G(\alpha, \beta). \end{aligned}$$

- Давайте выясним, как изменятся параметры, и заодно докажем.

Когда и μ, и σ^2 меняются

- Самое простое – это, по уже известным результатам,

$$\mu \mid x, \tau \sim \mathcal{N}\left(\frac{n\tau}{n\tau + n_0\tau}\bar{x} + \frac{n_0\tau}{n\tau + n_0\tau}\mu_0, n\tau + n_0\tau\right).$$

- Затем давайте разберёмся с $\tau \mid x$:

$$p(\tau, \mu \mid x) \propto p(\tau) \cdot p(\mu \mid \tau) \cdot p(x \mid \tau, \mu),$$

  и мы хотим это распределение маргинализовать по μ...

Когда и μ, и σ^2 меняются

- Подсчитаем:

$$
\begin{aligned}
p(\tau, \mu \mid x) &\propto p(\tau) \cdot p(\mu \mid \tau) \cdot p(x \mid \tau, \mu) \\
&\propto \tau^{\alpha-1} e^{-\tau\beta} \cdot \tau^{\frac{1}{2}} e^{-\frac{n_0 \tau}{2}(\mu-\mu_0)^2} \cdot \tau^{\frac{n}{2}} e^{-\frac{\tau}{2}\sum(x_i-\mu)^2} \\
&\propto \tau^{\alpha+\frac{n}{2}-\frac{1}{2}} e^{-\tau\left(\beta+\frac{1}{2}\sum(x_i-\bar{x})^2\right)} e^{-\frac{\tau}{2}\left(n_0(\mu-\mu_0)^2+n(\bar{x}-\mu)^2\right)}
\end{aligned}
$$

(простой трюк: $x_i - \mu = x_i - \bar{x} + \bar{x} - \mu$).

Когда и μ, и σ^2 меняются

- Теперь надо проинтегрировать

$$\int_{\mu} e^{-\frac{\tau}{2}\left(n_0(\mu-\mu_0)^2+n(\bar{x}-\mu)^2\right)} d\mu.$$

Упражнение. Проинтегрируйте. :) Должна получиться нормировочная константа

$$\tau^{-\frac{1}{2}} e^{\frac{-nn_0\tau}{2(n+n_0)}(\bar{x}-\mu_0)^2}.$$

Когда и μ, и σ^2 меняются

- Таким образом, получается апостериорное распределение

$$p(\tau \mid x) \propto \tau^{\alpha+\frac{n}{2}-1} e^{-\tau\left(\beta+\frac{1}{2}\sum(x_i-\bar{x})^2+\frac{nn_0}{2(n+n_0)}(\bar{x}-\mu_0)^2\right)}.$$

- Итого результаты такие:

$$\begin{aligned}
\mu \mid \tau, x &\sim \mathcal{N}\left(\tfrac{n\tau}{n\tau+n_0\tau}\bar{x}+\tfrac{n_0\tau}{n\tau+n_0\tau}\mu_0, n\tau+n_0\tau\right), \\
\tau \mid x &\sim G\left(\alpha+\tfrac{n}{2}, \beta+\tfrac{1}{2}\sum(x_i-\bar{x})^2+\tfrac{nn_0}{2(n+n_0)}(\bar{x}-\mu_0)^2\right).
\end{aligned}$$

Предсказание

- Теперь предсказание нового x_{new}:

$$p(x_{\text{new}} \mid x) = \int\int \underbrace{\text{Gamma}}_{\tau|x} \cdot \underbrace{\text{Gaussian}}_{\mu|\tau,x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,\mu} d\tau d\mu =$$

$$= \int \underbrace{\text{Gamma}}_{\tau|x} \int \underbrace{\text{Gaussian}}_{\mu|\tau,x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,\mu} d\tau d\mu =$$

$$= \int \underbrace{\text{Gamma}}_{\tau|x} \cdot \underbrace{\text{Gaussian}}_{x_{\text{new}}|\tau,x} d\tau = \ldots$$

- В результате получится распределение Стьюдента.

Спасибо!

Спасибо за внимание!