Сопряженные априорные распределения

Сергей Николенко

Академия больших данных MADE — VK

31 января 2022 г.

Random facts:

- 31 января 1606 г. лидеры Порохового заговора были hanged, drawn, and quartered; приговорённых к этой казни последовательно вешали (не до смерти), кастрировали, потрошили, четвертовали и обезглавливали; впрочем, сам Гай Фокс умудрился спрыгнуть с эшафота и удачно сломать себе шею ещё на первом этапе
- 31 января 1714 г. Пётр I учредил «Государев кабинет», будущую Кунсткамеру, а также издал указ об обучении дворянских детей «цыфири и геометрии»; «не постигшим основ знаний» запрещалось жениться
- 31 января 1968 г. была провозглашена независимость Науру, самой маленькой независимой республики на Земле (21.3 км^2, 10000 человек)
- 31 января 1990 г. на Пушкинской площади в Москве открылся первый (и единственный) McDonald's в СССР
- 31 января 2020 г. Великобритания потеряла представительство и право голоса в органах власти ЕС; в полночь с 31 января на 1 февраля было прекращено членство Великобритании в ЕС, продолжавшееся с 1973 года

Еще немного про вероятности

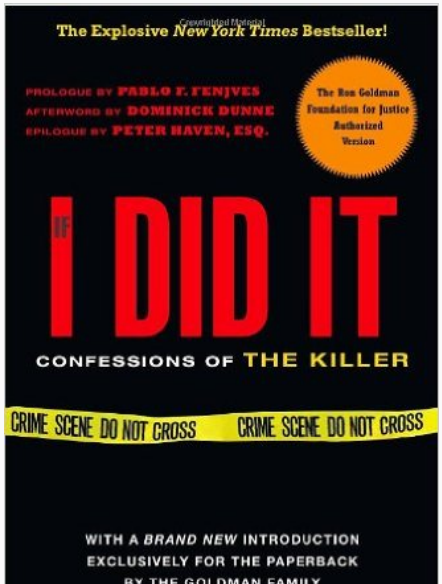
Зачем нужны вероятностные модели

- Зачем нужны вероятностные модели? Апостериорные вероятности помогают:
  - добавить опцию «я не знаю»;
  - минимизировать риск, учесть разные веса ошибок;
  - перебалансировать классы или по-другому добавить априорные вероятности;
  - комбинировать модели (например, наивным Байесом)...
- Понимание смысла помогает:
  - понимать границы применимости, предположения, которые делают модели;
  - обобщать и переносить идеи моделей на другие задачи;
  - содержательно интерпретировать происходящее.

Байес в жизни

(1) Прокурор указал, что O.J. Simpson уже бил жену в прошлом. Адвокат ответил: «Убивают только одну из 2500 женщин, подвергавшихся семейному насилию, так что это вообще нерелевантно». Суд согласился с адвокатом; верно ли это рассуждение?

(2) У Салли Кларк погибли два младенца; прокурор указал, что вероятность двух случаев SIDS в одной семье, которую он получил из статистики одиночных случаев, — около 1 из 73 миллионов; в чём он не прав?



Байесовский вывод для монетки

ML vs. MAP

- Итак, в статистике обычно ищут *гипотезу максимального правдоподобия* (maximum likelihood):

$$\theta_{ML} = \arg\max_\theta p(D \mid \theta).$$

- В байесовском подходе ищут *апостериорное распределение* (posterior)

$$p(\theta|D) \propto p(D|\theta)p(\theta)$$

  и, возможно, *максимальную апостериорную гипотезу* (maximum a posteriori):

$$\theta_{MAP} = \arg\max_\theta p(\theta \mid D) = \arg\max_\theta p(D \mid \theta)p(\theta).$$

Постановка задачи

- Простая задача вывода: дана нечестная монетка, она подброшена N раз, имеется последовательность результатов падения монетки. Надо определить её «нечестность» и предсказать, чем она выпадет в следующий раз.

Постановка задачи

- Если у нас есть вероятность p_h того, что монетка выпадет решкой (вероятность орла $p_t = 1 - p_h$), то вероятность того, что выпадет последовательность s, которая содержит n_h решек и n_t орлов, равна

$$p(s|p_h) = p_h^{n_h}(1 - p_h)^{n_t}.$$

- Сделаем предположение: будем считать, что монетка выпадает равномерно, т.е. у нас нет априорного знания p_h.
- Теперь нужно использовать теорему Байеса и вычислить скрытые параметры.

Пример применения теоремы Байеса

- Правдоподобие: $p(p_h|s) = \frac{p(s|p_h)p(p_h)}{p(s)}$.
- Здесь $p(p_h)$ следует понимать как непрерывную случайную величину, сосредоточенную на интервале $[0, 1]$, коей она и является. Наше предположение о равномерном распределении в данном случае значит, что априорная вероятность $p(p_h) = 1$, $p_h \in [0, 1]$ (т.е. априори мы не знаем, насколько нечестна монетка, и предполагаем это равновероятным). А $p(s|p_h)$ мы уже знаем.
- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1 - p_h)^{n_t}}{p(s)}.$$

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

- $p(s)$ можно подсчитать как

$$p(s) = \int_0^1 p_h^{n_h}(1-p_h)^{n_t} dp_h = \\ = \frac{\Gamma(n_h+1)\Gamma(n_t+1)}{\Gamma(n_h+n_t+2)} = \frac{n_h! n_t!}{(n_h+n_t+1)!},$$

  но найти $\arg\max_{p_h} p(p_h \mid s) = \frac{n_h}{n_h+n_t}$ можно и без этого.

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

- Но это ещё не всё. Чтобы предсказать следующий исход, надо найти $p(\text{heads}|s)$:

$$\begin{aligned} p(\text{heads}|s) &= \int_0^1 p(\text{heads}|p_h)p(p_h|s)dp_h = \\ &= \int_0^1 \frac{p_h^{n_h+1}(1-p_h)^{n_t}}{p(s)}dp_h = \\ &= \frac{(n_h+1)!n_t!}{(n_h+n_t+2)!} \cdot \frac{(n_h+n_t+1)!}{n_h!n_t!} = \frac{n_h+1}{n_h+n_t+2}. \end{aligned}$$

- Получили правило Лапласа.

Пример применения теоремы Байеса

- Итого получается:

$$p(p_h|s) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s)}.$$

- Это была иллюстрация двух основных задач байесовского вывода:
  1. найти апостериорное распределение на гипотезах/параметрах:

$$p(\theta \mid D) \propto p(D|\theta)p(\theta)$$

     (и/или найти гипотезу максимального правдоподобия $\arg\max_\theta p(\theta \mid D)$);
  2. найти апостериорное распределение исходов дальнейших экспериментов:

$$p(x \mid D) \propto \int_{\theta\in\Theta} p(x \mid \theta)p(D|\theta)p(\theta)\mathrm{d}\theta.$$

Сопряжённые априорные распределения

Напоминание

- Напоминаю, что основная наша задача – как обучить параметры распределения и/или предсказать следующие его точки по имеющимся данным.
- В байесовском выводе участвуют:
  - $p(x \mid \theta)$ – правдоподобие данных;
  - $p(\theta)$ – априорное распределение;
  - $p(x) = \int_\Theta p(x \mid \theta) p(\theta) d\theta$ – маргинальное правдоподобие;
  - $p(\theta \mid x) = \frac{p(x|\theta)p(\theta)}{p(x)}$ – апостериорное распределение;
  - $p(x' \mid x) = \int_\Theta p(x' \mid \theta) p(\theta \mid x) d\theta$ – предсказание нового x'.
- Задача обычно в том, чтобы найти $p(\theta \mid x)$ и/или $p(x' \mid x)$.

Априорные распределения

- Когда мы проводим байесовский вывод, у нас, кроме правдоподобия, должно быть ещё *априорное распределение* (prior distribution) по всем возможным значениям параметров.
- Мы раньше к ним специально не присматривались, но они очень важны.
- Задача байесовского вывода – как подсчитать $p(\theta \mid x)$ и/или $p(x' \mid x)$.
- Но чтобы это сделать, сначала надо выбрать $p(\theta)$. Как выбирать априорные распределения?

Сопряжённые априорные распределения

- Разумная цель: давайте будем выбирать распределения так, чтобы они оставались такими же и *a posteriori*.
- До начала вывода есть априорное распределение $p(\theta)$.
- После него есть какое-то новое апостериорное распределение $p(\theta \mid x)$.
- Я хочу, чтобы $p(\theta \mid x)$ тоже имело тот же вид, что и $p(\theta)$, просто с другими параметрами.

Сопряжённые априорные распределения

- Не слишком формальное определение: семейство распределений $p(\theta \mid \alpha)$ называется семейством *сопряжённых априорных распределений* для семейства правдоподобий $p(x \mid \theta)$, если после умножения на правдоподобие апостериорное распределение $p(\theta \mid x, \alpha)$ остаётся в том же семействе: $p(\theta \mid x, \alpha) = p(\theta \mid \alpha')$.
- α называются *гиперпараметрами* (hyperparameters), это «параметры распределения параметров».
- Тривиальный пример: семейство всех распределений будет сопряжённым чему угодно, но это не очень интересно.

Сопряжённые априорные распределения

- Разумеется, вид хорошего априорного распределения зависит от вида распределения собственно данных, $p(x \mid \theta)$.
- Сопряжённые априорные распределения подсчитаны для многих распределений, мы приведём несколько примеров.

Испытания Бернулли

- Каким будет сопряжённое априорное распределение для бросания нечестной монетки (испытаний Бернулли)?
- Ответ: это будет бета-распределение; плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

Испытания Бернулли

- Плотность распределения нечестности монетки θ

$$p(\theta \mid \alpha, \beta) = \frac{\theta^{\alpha-1}(1-\theta)^{\beta-1}}{B(\alpha, \beta)}.$$

- Тогда, если мы посэмплируем монетку, получив s орлов и f решек, получится

$$p(s, f \mid \theta) = \binom{s+f}{s} \theta^s (1-\theta)^f, \text{ и}$$

$$p(\theta|s,f) = \frac{\binom{s+f}{s}\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}/B(\alpha,\beta)}{\int_0^1 \binom{s+f}{s}x^{s+\alpha-1}(1-x)^{f+\beta-1}/B(\alpha,\beta)dx} =$$

$$= \frac{\theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}}{B(s+\alpha, f+\beta)}.$$

Испытания Бернулли

- Итого получается, что сопряжённое априорное распределение для параметра нечестной монетки θ – это

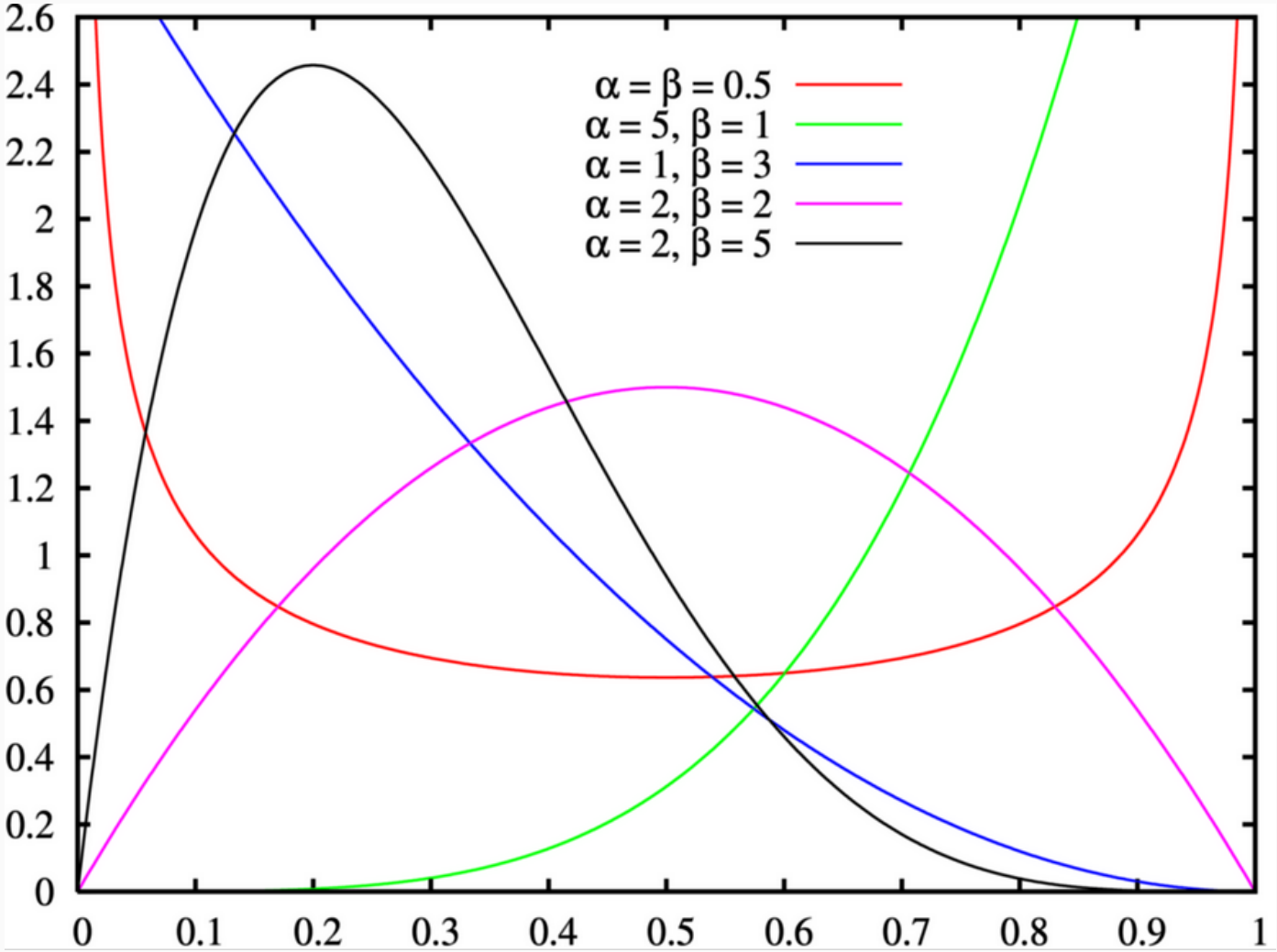
$$p(\theta \mid \alpha, \beta) \propto \theta^{\alpha-1}(1-\theta)^{\beta-1}.$$

- После получения новых данных с s орлами и f решками гиперпараметры меняются на

$$p(\theta \mid s+\alpha, f+\beta) \propto \theta^{s+\alpha-1}(1-\theta)^{f+\beta-1}.$$

- На этом этапе можно забыть про сложные формулы и выводы, получилось очень простое правило обучения (под обучением теперь понимается изменение гиперпараметров).

Бета-распределение



Мультиномиальное распределение

- Простое обобщение: рассмотрим мультиномиальное распределение с n испытаниями, k категориями и по x_i экспериментов дали категорию i.
- Параметры θ_i показывают вероятность попасть в категорию i:

$$p(x \mid \theta) = \binom{n}{x_1, \ldots, x_n} \theta_1^{x_1} \theta_2^{x_2} \ldots \theta_k^{x_k}.$$

- Сопряжённым априорным распределением будет распределение Дирихле:

$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \ldots \theta_k^{\alpha_k - 1}.$$

Мультиномиальное распределение

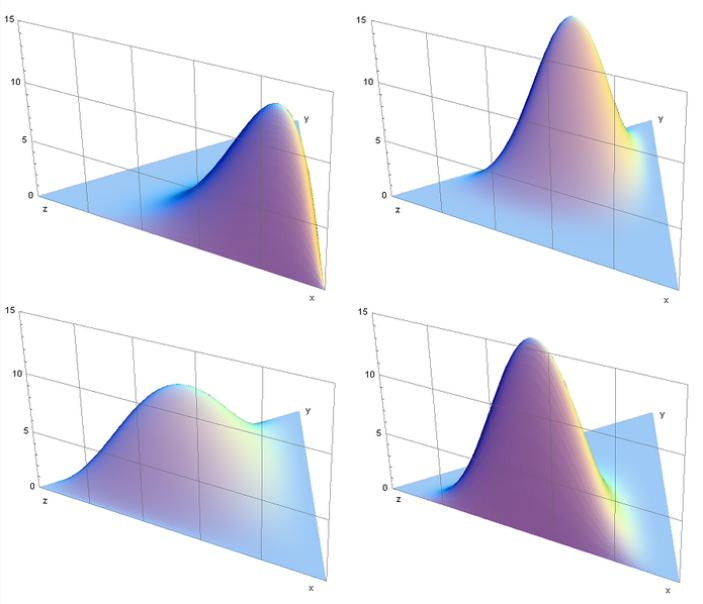
- Сопряжённым априорным распределением будет распределение Дирихле:

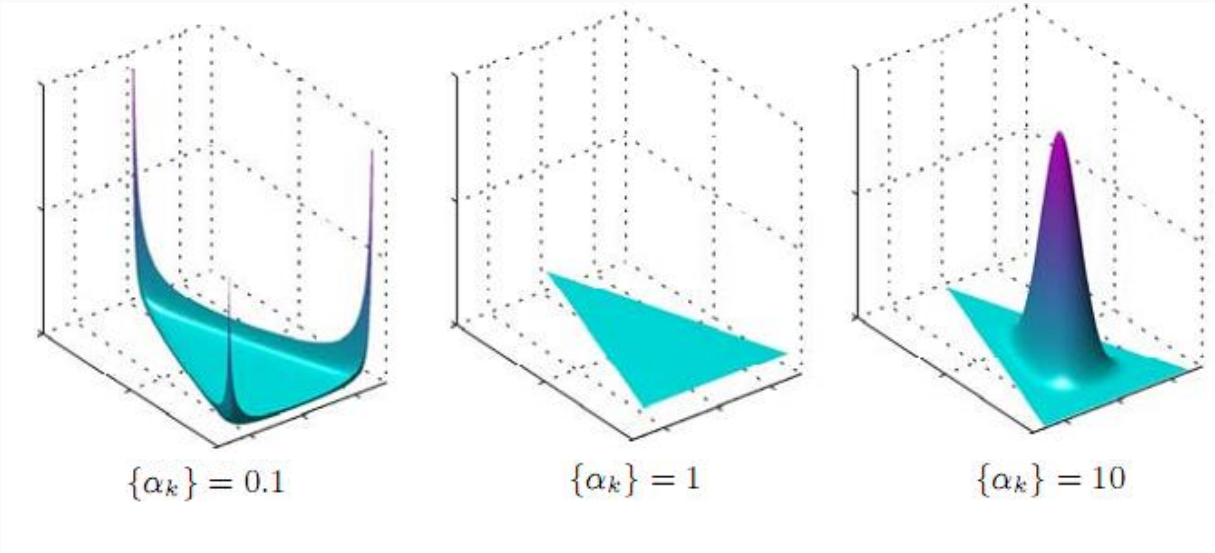
$$p(\theta \mid \alpha) \propto \theta_1^{\alpha_1 - 1} \theta_2^{\alpha_2 - 1} \dots \theta_k^{\alpha_k - 1}.$$

Упражнение. Докажите, что при получении данных $x_1, \dots, x_k$ гиперпараметры изменятся на

$$p(\theta \mid x, \alpha) = p(\theta \mid x + \alpha) \propto \theta_1^{x_1 + \alpha_1 - 1} \theta_2^{x_2 + \alpha_2 - 1} \dots \theta_k^{x_k + \alpha_k - 1}.$$

Распределение Дирихле





Линейная регрессия

Метод наименьших квадратов

- Линейная регрессия: рассмотрим линейную функцию

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{p} x_j w_j = \mathbf{x}^\top \mathbf{w}, \quad \mathbf{x} = (1, x_1, \ldots, x_p).$$

- Таким образом, по вектору входов $\mathbf{x}^\top = (x_1, \ldots, x_p)$ мы будем предсказывать выход y как

$$\hat{y}(\mathbf{x}) = \hat{w}_0 + \sum_{j=1}^{p} x_j \hat{w}_j = \mathbf{x}^\top \hat{\mathbf{w}}.$$

Метод наименьших квадратов

- Как найти оптимальные параметры $\hat{\mathbf{w}}$ по тренировочным данным вида $(\mathbf{x}_i, y_i)_{i=1}^N$?
- Метод наименьших квадратов: будем минимизировать

$$\mathrm{RSS}(\mathbf{w}) = \sum_{i=1}^{N} (y_i - \mathbf{x}_i^\top \mathbf{w})^2.$$

- Как минимизировать?

Метод наименьших квадратов

- Можно на самом деле решить задачу точно – записать как

$$\mathrm{RSS}(\mathbf{w}) = (\mathbf{y} - \mathbf{X}\mathbf{w})^\top (\mathbf{y} - \mathbf{X}\mathbf{w}),$$

  где $\mathbf{X}$ – матрица $N \times p$, продифференцировать по $\mathbf{w}$, получится

$$\hat{\mathbf{w}} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y},$$

  если матрица $\mathbf{X}^\top \mathbf{X}$ невырожденная.

- Замечание: $\left(\mathbf{X}^\top \mathbf{X}\right)^{-1} \mathbf{X}^\top$ называется *псевдообратной матрицей Мура–Пенроуза* (Moore–Penrose pseudo-inverse) матрицы $\mathbf{X}$; это обобщение понятия обратной матрицы на неквадратные матрицы.

Байесовская регрессия

- Теперь давайте поговорим о линейной регрессии по-байесовски.
- Основное наше предположение – в том, что шум (ошибка в данных) распределён нормально, т.е. переменная t, которую мы наблюдаем, получается как

$$t = y(\mathbf{x}, \mathbf{w}) + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

  Иными словами,

$$p(t \mid \mathbf{x}, \mathbf{w}, \sigma^2) = \mathcal{N}(t \mid y(\mathbf{x}, \mathbf{w}), \sigma^2).$$

- Здесь пока y – любая функция.

Байесовская регрессия

- Чтобы не повторять совсем уж то же самое, мы рассмотрим не в точности линейную регрессию, а её естественное обобщение – линейную модель с базисными функциями:

$$y(\mathbf{x}, \mathbf{w}) = w_0 + \sum_{j=1}^{M-1} w_j \phi_j(\mathbf{x}) = \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x})$$

(M параметров, $M-1$ базисная функция, $\phi_0(\mathbf{x}) = 1$).

Байесовская регрессия

- Базисные функции ϕ_i – это, например:
  - результат feature extraction;
  - расширение линейной модели на нелинейные зависимости (например, $\phi_j(x) = x^j$);
  - локальные функции, которые существенно не равны нулю только в небольшой области (например, гауссовские базисные функции $\phi_j(\mathbf{x}) = e^{-\frac{(x-\mu_j)^2}{2s^2}}$);
  - ...

Байесовская регрессия

- Рассмотрим набор данных $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ со значениями $\mathbf{t} = \{t_1, \ldots, t_N\}$.
- Будем предполагать, что данные взяты независимо по одному и тому же распределению:

$$p(\mathbf{t} \mid \mathbf{X}, \mathbf{w}, \sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(t_n \mid \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n), \sigma^2\right).$$

- Прологарифмируем (опустим $\mathbf{X}$, т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right)^2.$$

Байесовская регрессия

- Прологарифмируем (опустим $\mathbf{X}$, т.к. по нему всегда условная вероятность будет):

$$\ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = -\frac{N}{2} \ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right)^2.$$

- И вот мы получили, что для максимизации правдоподобия по $\mathbf{w}$ нам нужно как раз минимизировать среднеквадратичную ошибку!

$$\nabla_{\mathbf{w}} \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = \frac{1}{\sigma^2} \sum_{n=1}^{N} \left(t_n - \mathbf{w}^\top \boldsymbol{\phi}(\mathbf{x}_n)\right) \boldsymbol{\phi}(\mathbf{x}_n).$$

Байесовская регрессия

- Решая систему уравнений $\nabla \ln p(\mathbf{t} \mid \mathbf{w}, \sigma^2) = 0$, получаем то же самое, что и раньше:

$$\mathbf{w}_{ML} = \left(\mathbf{\Phi}^\top \mathbf{\Phi}\right)^{-1} \mathbf{\Phi}^\top \mathbf{t}.$$

- Здесь $\mathbf{\Phi} = (\phi_j(\mathbf{x}_i))_{i,j}$.

Байесовская регрессия

- Теперь можно и относительно σ^2 максимизировать правдоподобие; получим

$$\sigma^2_{ML} = \frac{1}{N}\sum_{n=1}^{N}\left(t_n - \mathbf{w}^\top_{ML}\boldsymbol{\phi}(\mathbf{x}_n)\right)^2,$$

  т.е. как раз выборочная дисперсия имеющихся данных вокруг предсказанного значения.

Спасибо!

Спасибо за внимание!