ЛОГИСТИЧЕСКАЯ РЕГРЕССИЯ

Сергей Николенко

СПбГУ — Санкт-Петербург

20 октября 2018 г.

Random facts:

- 20 октября — Всемирный день статистики; празднуют его более ста стран, но почему-то лишь один раз в пять лет, так что следующий будет через два года
- 20 октября — день поминовения святой Магдалены из Нагасаки; она родилась в семье японских христиан в 1611 году; в 1620 её родителей казнили за веру, после чего Магдалена перешла в орден августинцев под руководство братьев Франциска и Винсента; тех сожгли заживо в 1632 году, и Магдалена перешла в ученицы к братьям Мельхиору и Мартену; когда и тех казнили, Магдалена сама предала себя в руки властей и объявила себя христианкой, что закончилось тоже довольно предсказуемо
- 20 октября 1955 г. было опубликовано «Возвращение короля»
- 20 октября 1964 г. Rolling Stones дали первый концерт в Париже; в результате беспорядков было арестовано 150 человек
- 20 октября 1968 г. Дик Фосбери победил в Мехико со своим флопом
- 20 октября 1982 г. на матче «Спартака» и голландского клуба Haarlem произошла «трагедия в Лужниках»; в давке погибло 66 человек

Логистическая регрессия

В прошлый раз

- Итак, мы рассмотрели логистический сигмоид:

$$p(C_1 \mid x) = \frac{p(x \mid C_1)p(C_1)}{p(x \mid C_1)p(C_1) + p(x \mid C_2)p(C_2)} = \frac{1}{1 + e^{-a}} = \sigma(a),$$

$$\text{где } a = \ln \frac{p(x \mid C_1)p(C_1)}{p(x \mid C_2)p(C_2)}, \qquad \sigma(a) = \frac{1}{1 + e^{-a}}.$$

- Вывели из него LDA и QDA, обучили их методом максимального правдоподобия, а потом отвлеклись на naive Bayes.

Два класса

- Возвращаемся к задаче классификации.
- Два класса, и апостериорное распределение – логистический сигмоид на линейной функции:

$$p(C_1 \mid \phi) = y(\phi) = \sigma(w^\top \phi), \quad p(C_2 \mid \phi) = 1 - p(C_1 \mid \phi).$$

- *Логистическая регрессия* – это когда мы напрямую оптимизируем w.

Два класса

- Для датасета $\{\phi_n, t_n\}$, $t_n \in \{0, 1\}$, $\phi_n = \phi(x_n)$:

$$p(t \mid w) = \prod_{n=1}^{N} y_n^{t_n} (1 - y_n)^{1 - t_n}, \quad y_n = p(C_1 \mid \phi_n).$$

- Ищем параметры максимального правдоподобия, минимизируя $-\ln p(t \mid w)$:

$$E(w) = -\ln p(t \mid w) = -\sum_{n=1}^{N} \left[t_n \ln y_n + (1 - t_n) \ln(1 - y_n) \right].$$

Два класса

- Пользуясь тем, что $\sigma' = \sigma(1 - \sigma)$, берём градиент (похоже на перцептрон):

$$\nabla E(w) = \sum_{n=1}^{N} (y_n - t_n)\phi_n.$$

- Если теперь сделать градиентный спуск, получим как раз разделяющую поверхность.
- Заметим, правда, что если данные действительно разделимы, то может получиться жуткий оверфиттинг: $\|w\| \to \infty$, и сигмоид превращается в функцию Хевисайда. Надо регуляризовать.

IRLS

- В логистической регрессии не получается замкнутого решения из-за сигмоида.
- Но функция $E(w)$ всё равно выпуклая, и можно воспользоваться методом Ньютона-Рапсона – на каждом шаге использовать локальную квадратичную аппроксимацию к функции ошибки:

$$w^{\text{new}} = w^{\text{old}} - H^{-1}\nabla E(w),$$

где H (Hessian) – матрица вторых производных $E(w)$.

IRLS

- Замечание: давайте применим Ньютона-Рапсона к обычной линейной регрессии с квадратической ошибкой:

$$\nabla E(w) = \sum_{n=1}^{N} \left(w^\top \phi_n - t_n\right) \phi_n = \Phi^\top \Phi w - \Phi^\top t,$$

$$\nabla\nabla E(w) = \sum_{n=1}^{N} \phi_n \phi_n^\top = \Phi^\top \Phi,$$

и шаг оптимизации будет

$$w^{\text{new}} = w^{\text{old}} - \left(\Phi^\top \Phi\right)^{-1} \left[\Phi^\top \Phi w^{\text{old}} - \Phi^\top t\right] =$$

$$= \left(\Phi^\top \Phi\right)^{-1} \Phi^\top t,$$

т.е. мы за один шаг придём к решению.

IRLS

- Для логистической регрессии:

$$\nabla E(w) = \sum_{n=1}^{N} (y_n - t_n)\, \phi_n = \Phi^\top (y - t)\,,$$

$$H = \nabla\nabla E(w) = \sum_{n=1}^{N} y_n(1 - y_n)\phi_n \phi_n^\top = \Phi^\top R \Phi$$

для диагональной матрицы R с $R_{nn} = y_n(1 - y_n)$.

IRLS

- Формула шага оптимизации:

$$w^{\text{new}} = w^{\text{old}} - \left(\Phi^\top R \Phi\right)^{-1} \Phi^\top (y - t) = \left(\Phi^\top R \Phi\right)^{-1} \Phi^\top R z,$$

  где $z = \Phi w^{\text{old}} - R^{-1}(y - t)$.
- Получилось как бы решение взвешенной задачи минимизации квадратического отклонения с матрицей весов R.
- Отсюда название: iterative reweighted least squares (IRLS).

Несколько классов

- В случае нескольких классов

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

- 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 t_n и правдоподобие

$$p(T \mid w_1, \dots, w_K) = \prod_{n=1}^{N} \prod_{k=1}^{K} p(C_k \mid \phi_n)^{t_{nk}} = \prod_{n=1}^{N} \prod_{k=1}^{K} y_{nk}^{t_{nk}}$$

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

$$E(w_1, \dots, w_K) = -\ln p(T \mid w_1, \dots, w_K) = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}, \text{ и}$$

$$\nabla_{w_j} E(w_1, \dots, w_K) = -\sum_{n=1}^{N} (y_{nj} - t_{nj}) \phi_n.$$

Несколько классов

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

$$\nabla_{w_k} \nabla_{w_j} E(w_1, \dots, w_K) = -\sum_{n=1}^{N} y_{nk} \left([k = j] - y_{nj} \right) \phi_n \phi_n^\top.$$

Пробит-регрессия

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

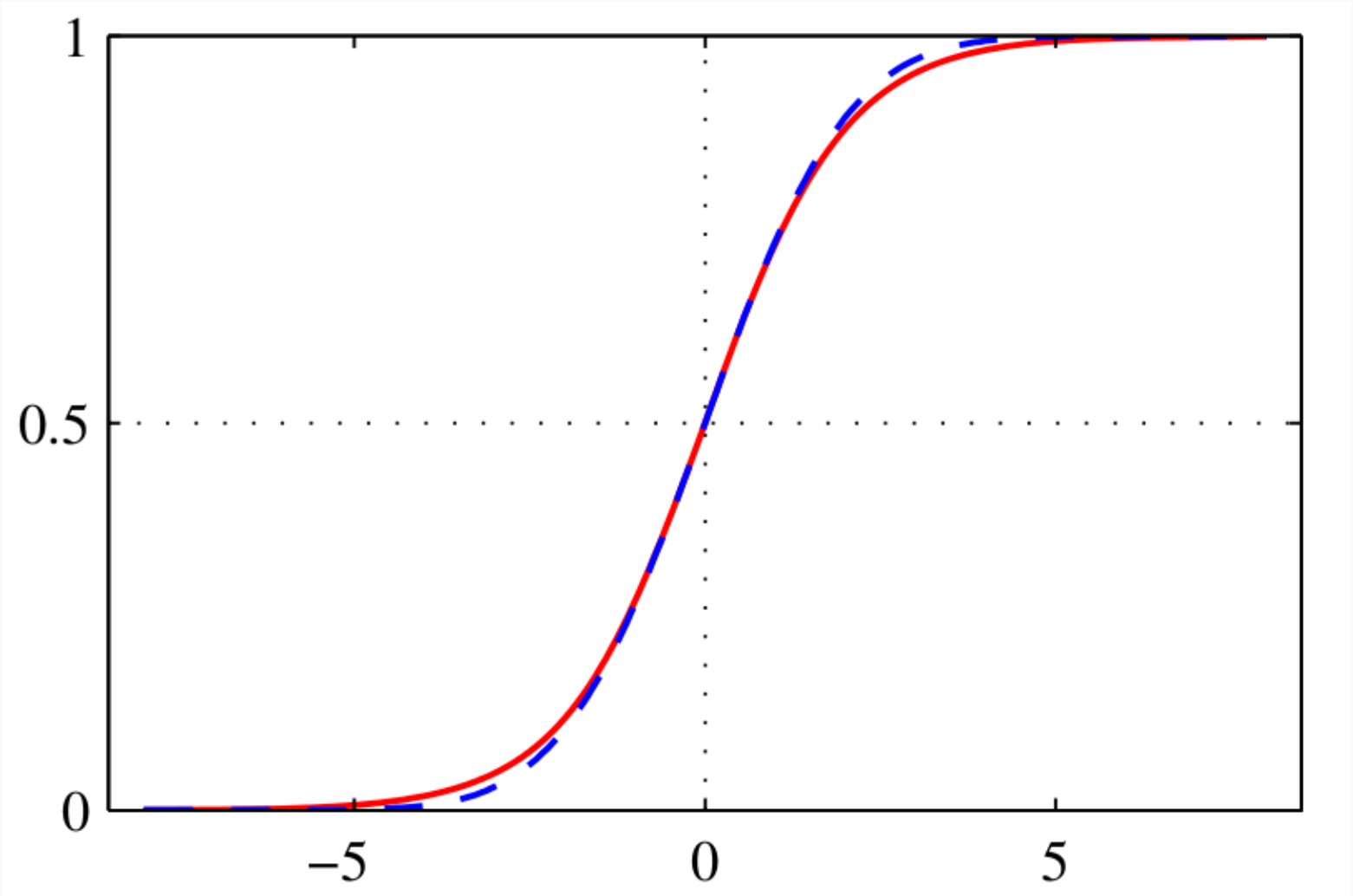
- Это называется *пробит-функцией* (probit); неэлементарная, но тесно связана с

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

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

- 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

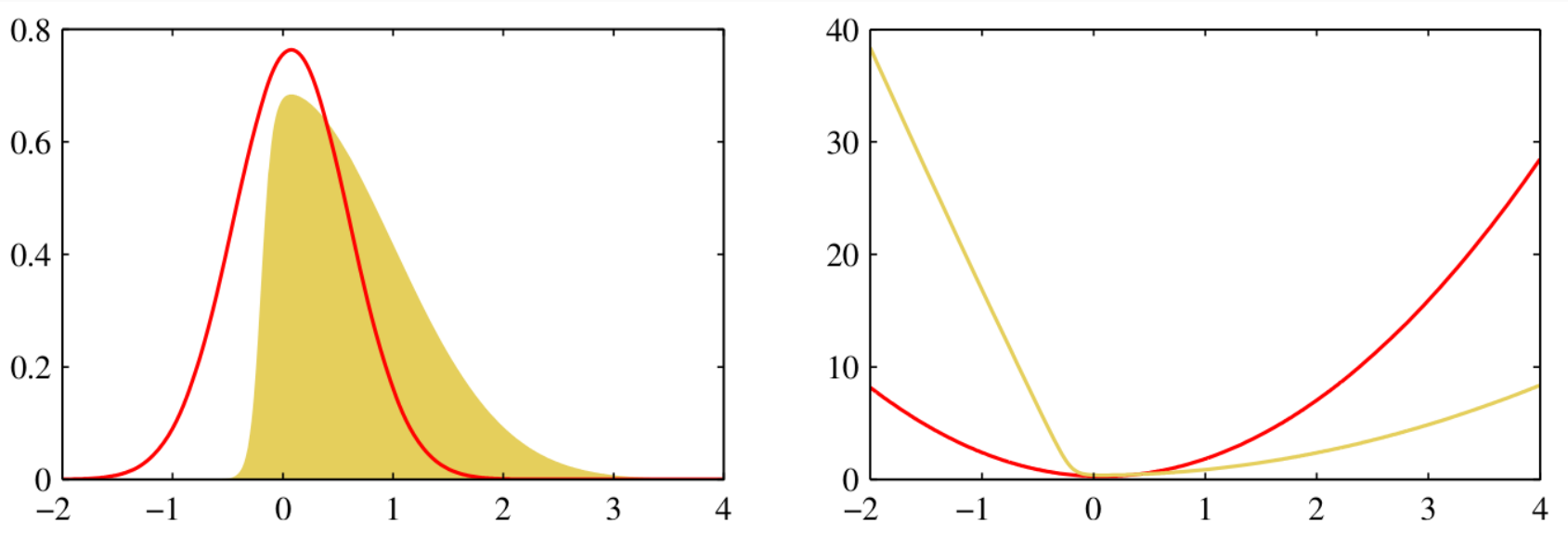
- Это можно обобщить на многомерное распределение $p(z) = \frac{1}{Z} f(z)$:

$$f(z) \approx f(z_0) e^{-\frac{1}{2}(z-z_0)^\top A (z-z_0)},$$

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

Упражнение. Какая здесь будет нормировочная константа?

Лапласовская аппроксимация



Сравнение моделей по Лапласу

- Вооружившись лапласовской аппроксимацией, давайте применим её сначала к выбору моделей.
- 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:

$$Z = \int f(z)dz \approx \int f(z_0)e^{-\frac{1}{2}(z-z_0)^\top A(z-z_0)}dz = f(z_0)\frac{(2\pi)^{M/2}}{|A|^{1/2}}.$$

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

Сравнение моделей по Лапласу

- Получаем

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

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

Сравнение моделей по Лапласу

- Получаем

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

- Если гауссовское априорное распределение $p(\theta)$ достаточно широкое, и 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).

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:

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

- Тогда апостериорное будет

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

Байесовская логистическая регрессия

- Чтобы приблизить, сначала находим максимум w_{MAP}, а потом матрица ковариаций – это матрица вторых производных

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

- Наше приближение – это

$$q(w) = N(w \mid w_{\mathrm{MAP}}, \Sigma_N).$$

Байесовская логистическая регрессия

- Теперь можно описать байесовское предсказание:

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

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

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

Байесовская логистическая регрессия

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

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

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

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

Байесовская логистическая регрессия

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

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

- Итого получили, что

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

Байесовская логистическая регрессия

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

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

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

Проклятие размерности

В предыдущих сериях...

- Теорема Байеса:

$$p(\theta|D) = \frac{p(\theta)p(D|\theta)}{p(D)}.$$

- Две основные задачи байесовского вывода:
  1. найти апостериорное распределение на гипотезах/параметрах:

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

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

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

В предыдущих сериях...

- Мы изучили метод наименьших квадратов для линейной регрессии и метод ближайших соседей...
- ...построили функцию регрессии

$$\hat{f}(x) = \mathrm{E}_{y|x'}(y \mid x' = x)$$

и оптимальный классификатор

$$\hat{g}(x) = \arg\min_g \sum_{k=1}^{K} L(g_k, \hat{g}(x)) p(g_k \mid x) \ldots$$

- ...и выяснили, что метод наименьших квадратов – это метод максимального правдоподобия для нормально распределённого шума.

Проклятие размерности

- В размерности 2-3 k-NN даёт гораздо более разумные результаты, чем линейная модель, особенно если хорошо выбрать k.
- Может быть, нам в этой жизни больше ничего и не нужно?
- Давайте посмотрим, как k-NN будет вести себя в более высокой размерности (что очень реалистично).

Проклятие размерности

- Давайте поищем ближайших соседей у точки в единичном гиперкубе. Предположим, что наше исходное распределение равномерное.
- Чтобы покрыть долю α тестовых примеров, нужно (ожидаемо) покрыть долю α объёма, и ожидаемая длина ребра гиперкуба-окрестности в размерности p будет $e_p(\alpha) = \alpha^{1/p}$.
- Например, в размерности 10 $e_{10}(0.1) = 0.8$, $e_{10}(0.01) = 0.63$, т.е. чтобы покрыть 1% объёма, нужно взять окрестность длиной больше половины носителя по каждой координате!
- Это скажется и на k-NN: трудно отвергнуть по малому числу координат, быстрые алгоритмы хуже работают.

Проклятие размерности

- Второе проявление the curse of dimensionality: пусть N точек равномерно распределены в единичном шаре размерности p. Тогда среднее расстояние от нуля до точки равно

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p},$$

  т.е., например, в размерности 10 для $N = 500$ $d \approx 0.52$, т.е. больше половины.
- Большинство точек в результаты ближе к границе носителя, чем к другим точкам, а это для ближайших соседей проблема – придётся не интерполировать внутри существующих точек, а экстраполировать наружу.

Проклятие размерности

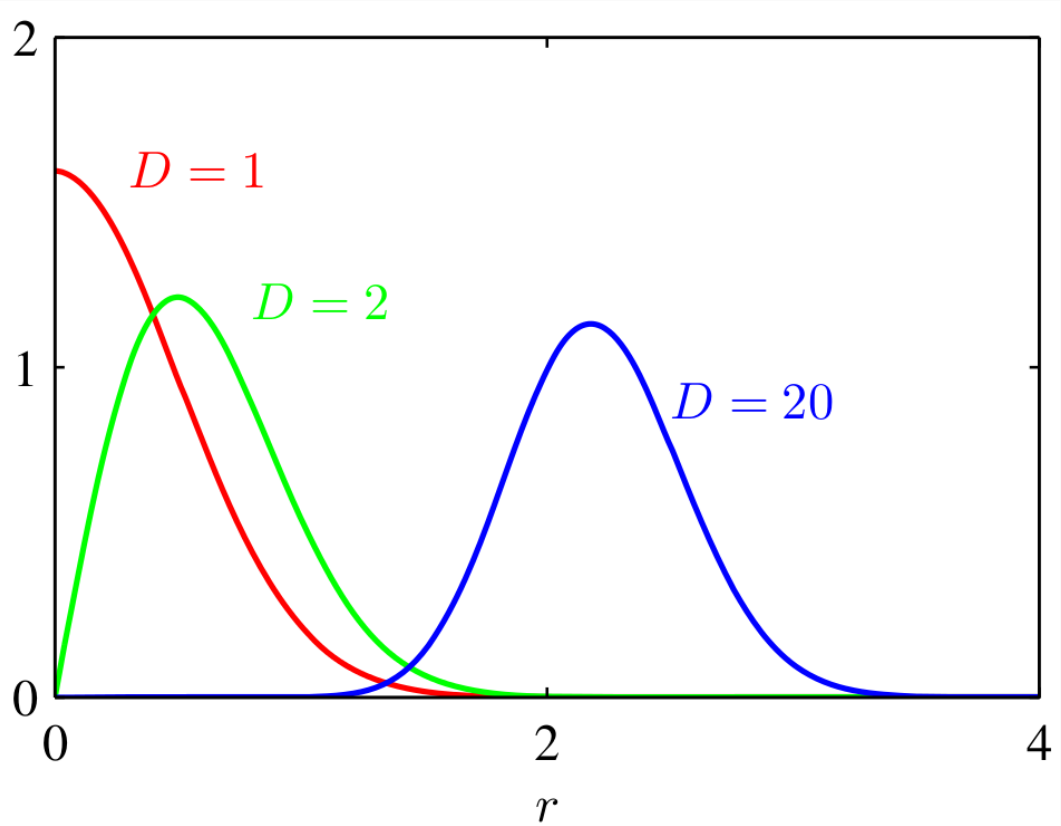
- Третье проявление: проблемы в оптимизации, которые и имел в виду Беллман.
- Если нужно примерно оптимизировать функцию от d переменных, на решётке с шагом ϵ понадобится примерно $\left(\frac{1}{\epsilon}\right)^d$ вычислений функции.
- В численном интегрировании – чтобы интегрировать функцию с точностью ϵ, нужно тоже примерно $\left(\frac{1}{\epsilon}\right)^d$ вычислений.

Проклятие размерности

- Плотные множества становятся очень разреженными. Например, чтобы получить плотность, создаваемую в размерности 1 при помощи $N = 100$ точек, в размерности 10 нужно будет 100^{10} точек.
- Поведение функций тоже усложняется с ростом размерности – чтобы строить регрессии в высокой размерности с той же точностью, может потребоваться экспоненциально больше точек, чем в низкой размерности.
- А у линейной модели ничего такого не наблюдается, она не подвержена проклятию размерности.

Проклятие размерности

- Ещё пример: нормально распределённая величина будет сосредоточена в тонкой оболочке.



Упражнение. Переведите плотность нормального распределения в полярные координаты и проверьте это утверждение.

Спасибо!

Спасибо за внимание!