Байесовский вывод

Сергей Николенко

Машинное обучение — ИТМО, осень 2006

Outline

Задача

- Why have sex?

Задача

- Why have sex?
- Этим мы займёмся на семинаре.

Вероятностное пространство

Definition

Вероятностное пространство — это тройка $(\Omega, \mathcal{F}, p)$, где

- Ω — множество, элементы которого называются элементарными событиями, исходами или точками;
- $\mathcal{F}$ — сигма-алгебра подмножеств Ω, называемых (случайными) событиями;
- P — вероятностная мера или вероятность, т.е. такая σ–аддитивная конечная мера, что $p(\Omega) = 1$.

Definition

Случайная величина $X : \Omega \longrightarrow \mathbb{R}$ индуцирует *распределение случайной величины* X на борелевских $P^X(B) = P(X \in B)$.

Операции над вероятностями

- Совместная вероятность $p(x, y)$, $x \in \Omega_X$, $y \in \Omega_Y$ определяется на прямом произведении $\Omega_X \times \Omega_Y$.
- Маргинализация из совместной вероятности:

$$p(x = a) = \sum_{y \in \Omega_Y} p(x = a, y).$$

- Условная вероятность:

$$p(x = a | y = b) = \frac{p(x = a, y = b)}{p(y = b)}, \text{ если } p(y = b) \neq 0.$$

- В жизни все вероятности условные; в нашем курсе они тоже будут обычно условными.

Общие правила

- Правило произведения:

$$p(x, y|H) = p(x|y, H)p(y|H) = p(y|x, H)p(x|H).$$

- Правило суммирования:

$$p(x|H) = \sum_y p(x, y|H) = \sum_y p(x|y, H)p(y|H).$$

- Теорема Байеса:

$$p(y|x, H) = \frac{p(x|y, H)p(y|H)}{p(x|H)} = \frac{p(x|y, H)p(y|H)}{\sum_{y'} p(x|y', H)p(y'|H)}.$$

- *Независимость* определим просто: две случайные переменные X и Y независимы, если

$$p(x, y) = p(x)p(y).$$

О болезнях и вероятностях

- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как позитивной, так и негативной ошибки).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?

О болезнях и вероятностях

- Приведём классический пример из классической области применения статистики — медицины.
- Пусть некий тест на какую-нибудь болезнь имеет вероятность успеха 95% (т.е. 5% — вероятность как позитивной, так и негативной ошибки).
- Всего болезнь имеется у 1% респондентов (отложим на время то, что они разного возраста и профессий).
- Пусть некий человек получил позитивный результат теста (тест говорит, что он болен). С какой вероятностью он действительно болен?
- Ответ: 16%.

Доказательство

- Обозначим через t результат теста, через d — наличие болезни.
- $p(d = 1) = p(d = 1|t = 1)p(t = 1) + p(d = 1|t = 0)p(t = 0)$.
- Используем теорему Байеса:

$$p(d = 1|t = 1) = \\ = \frac{p(t = 1|d = 1)p(t = 1)}{p(d = 1|t = 1)p(t = 1) + p(d = 1|t = 0)p(t = 0)} = \\ = \frac{0.95 \times 0.01}{0.95 \times 0.01 + 0.05 \times 0.99} = 0.16.$$

Вывод

- Вот такие задачи составляют суть вероятностного вывода (probabilistic inference).
- Поскольку они обычно основаны на теореме Байеса, вывод часто называют байесовским (Bayesian inference).
- Но не только поэтому.

Вероятность как частота

- Обычно в классической теории вероятностей, происходящей из физики, вероятность понимается как предел отношения количества определённого результата эксперимента к общему количеству экспериментов.
- Стандартный пример: бросание монетки.

Вероятность как степень доверия

- Мы можем рассуждать о том, «насколько вероятно» то, что В. В. Путин назначит преемником С. В. Иванова или то, что «Одиссею» написала женщина. Но о «количестве экспериментов» говорить бессмысленно — эксперимент ровно один.
- Здесь вероятности уже выступают как *степени доверия* (degrees of belief). Это байесовский подход к вероятностям (Томас Байес так понимал).
- К счастью, и те, и другие вероятности подчиняются одним и тем же законам, а если не подчиняются — значит, не вероятности и были.

Прямые и обратные задачи

- Прямая задача: в урне лежат 10 шаров, из них 3 чёрных. Какова вероятность выбрать чёрный шар?
- Или: в урне лежат 10 шаров с номерами от 1 до 10. Какова вероятность, что номера трёх последовательно выбранных шаров дадут в сумме 12?
- Обратная задача: перед нами две урны, в каждой по 10 шаров, но в одной 3 чёрных, а в другой — 6. Кто-то взял из какой-то урны шар, и он оказался чёрным. Какова вероятность, что он брал шар из первой урны?

Прямые и обратные задачи

- Иначе говоря, прямые задачи теории вероятностей описывают некий вероятностный процесс или модель и просят подсчитать ту или иную вероятность (т.е. фактически по модели предсказать поведение).
- Обратные задачи содержат *скрытые переменные* (в примере — номер урны, из которой брали шар). Они часто просят по известному поведению построить вероятностную модель.
- Задачи машинного обучения обычно являются задачами второй категории.

Определения

- Запишем теорему Байеса:

$$p(x|y,H) = \frac{p(x)p(y|x,H)}{p(y|H)}.$$

- Здесь $p(x)$ — *априорная вероятность* (prior probability), $p(y|x,H)$ — *правдоподобие* (likelihood), $p(x|y)$ — *апостериорная вероятность* (aposterior probability), $p(y|H)$ — вероятность данных (evidence).
- Вообще, *функция правдоподобия* имеет вид

$$a \mapsto p(y|x=a)$$

для некоторой случайной величины y.

Likelihood principle

- Принцип правдоподобия (likelihood principle): если дана модель данных d при условии параметров θ, дана функция правдоподобия $p(d|\theta)$ и мы пронаблюдали некоторый набор данных d_1, в дальнейшем вывод и предсказания должны зависеть только от $p(d_1|\theta)$.
- Иными словами, нам нужно только знать, как вероятность данных конкретных тестовых примеров d зависит от гипотезы.

Outline

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.

Первые замечания

- 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

$$p(s|p_h, H_1) = p_h^{n_h}(1 - p_h)^{n_t}.$$

- 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.

Применение теоремы Байеса

- $p(p_h|s, H_1) = \frac{p(s|p_h, H_1)p(p_h|H_1)}{p(s|H_1)}$.
- 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 $p(p_h|H) = 1$, $p_h \in [0, 1]$ (т.е. априори мы не знаем, насколько нечестна монетка, и предполагаем это равновероятным). А $p(s|p_h, H_1)$ мы уже знаем.
- Итого получается:

$$p(p_h|s, H_1) = \frac{p_h^{n_h}(1 - p_h)^{n_t}}{p(s|H_1)}.$$

Применение теоремы Байеса

- Итого получается:

$$p(p_h|s, H_1) = \frac{p_h^{n_h}(1-p_h)^{n_t}}{p(s|H_1)}.$$

- Осталось подсчитать $p(s|H_1)$; её нужно маргинализовать из функции правдоподобия:

$$p(s|H_1) = \int_0^1 p_h^{n_h}(1-p_h)^{n_t}\,dp_a = \\ = \frac{\Gamma(n_h+1)\Gamma(n_t+1)}{\Gamma(n_h+n_t+2)} = \frac{n_h!n_t!}{(n_h+n_t+1)!}.$$

А теперь предскажем следующий исход

- Чтобы предсказать следующий исход, нужно вычислить $p(\text{heads}|s)$:

$$p(\text{heads}|s) = \int_0^1 p(\text{heads}|p_h)p(p_h|s, H_1)dp_h = \int_0^1 \frac{p_h^{n_h+1}(1-p_h)^{n_t}}{p(s)}dp_h =$$
$$= \frac{(n_h+1)!n_t!}{(n_h+n_t+2)!} \cdot \frac{(n_h+n_t+1)!}{n_h!n_t!} = \frac{n_h+1}{n_h+n_t+2}.$$

- Это называется *правило Лапласа*.

Постановка задачи

- Мы использовали предположение H_1 о равномерном априорном распределении p_h.
- Это было нашей моделью ситуации.
- Можно использовать другие модели; например, можно априори предположить, что монетка падает решкой с вероятностью $1/3$, т.е. $p_h = 1/3$. Обозначим через H_0
- Как сравнить, какая из моделей H_i лучше описывает данные?

Теорема Байеса

- Снова теорема Байеса:

$$p(H_1|s) = \frac{p(s|H_1)p(H_1)}{p(s)}, \quad p(H_0|s) = \frac{p(s|H_0)p(H_0)}{p(s)}.$$

- Нужно как-то оценить априорные вероятности $p(H_0)$ и $p(H_1)$; можно просто положить их равными $1/2$.
- Теперь можно составить отношение:

$$\frac{p(H_1|s)}{p(H_0|s)} = \frac{p(s|H_1)p(H_1)}{p(s|H_0)p(H_0)} = \frac{n_h!n_t!}{(n_h+n_t+1)!} \cdot \frac{1}{(1/3)^{n_h}(2/3)^{n_t}}.$$

- Это отношение можно использовать для сравнения моделей.

Outline

Определения

Definition

Гипотеза максимального правдоподобия (maximum likelihood hypothesis) — это набор параметров θ, который максимизирует функцию правдоподобия

$$p(D|\theta, H),$$

где D — имеющаяся информация (тестовые примеры).

- Гипотеза максимального правдоподобия — гипотеза, которая лучше всего описывает имеющиеся данные D в текущих предположениях H.

Гауссиан

- Начнём с простейшего гауссиана. Его распределение содержит два параметра:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- Функция правдоподобия данных $x_1, \ldots, x_n$:

$$p(x_1, \ldots, x_n | \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\sum_{i=1}^{n} \frac{(x_i-\mu)^2}{2\sigma^2}}.$$

Гауссиан: достаточные статистики

- Заметим, что функция эта зависит от двух параметров, а не от n:

$$p(x_1, \ldots, x_n | \mu, \sigma) = \frac{1}{(\sigma\sqrt{2\pi})^n} e^{-\frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}},$$

где

$$\bar{x} = \frac{x_1 + \ldots + x_n}{n}, \quad S = \sum_{i=1}^{n} (\bar{x} - x_n)^2.$$

- Параметры $\bar{x}$ и S называются *достаточными статистиками* (sufficient statistics).

Гауссиан: ГМП

- Какие параметры лучше всего описывают данные?
- Перейдём, как водится, к логарифму:

$$\ln p(x_1, \ldots, x_n | \mu, \sigma) = -n \ln(\sigma\sqrt{2\pi}) - \frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}.$$

- Как выяснить, при каких параметрах функция правдоподобия максимизируется?

Гауссиан: ГМП

- Какие параметры лучше всего описывают данные?
- Перейдём, как водится, к логарифму:

$$\ln p(x_1, \ldots, x_n | \mu, \sigma) = -n \ln(\sigma\sqrt{2\pi}) - \frac{S + n(\bar{x} - \mu)^2}{2\sigma^2}.$$

- Как выяснить, при каких параметрах функция правдоподобия максимизируется?
- Взять частные производные и приравнять нулю.

Гауссиан: ГМП

- По μ:
$$\frac{\partial \ln p}{\partial \mu} = -\frac{n}{\sigma^2}(\mu - \bar{x}).$$
- То есть в гипотезе максимального правдоподобия $\mu = \bar{x}$, независимо от S.
- Теперь нужно найти σ из гипотезы максимального правдоподобия.
- Для этого мы продифференцируем по $\ln \sigma$ — полезный приём на будущее. Кстати, $\frac{\mathrm{d}x^n}{\mathrm{d}(\ln x)} = nx^n$.

Гауссиан: ГМП

- $$\frac{\partial \ln p}{\partial \ln \sigma} = -n + \frac{\sum_{i=1}^{n} (x_i - \mu)}{\sigma^2}.$$
- Следовательно, в гипотезе максимального правдоподобия

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \mu)}{n}}.$$

Упражнение Оценить ошибку в полученных нами значениях (найти доверительные интервалы).

Несколько гауссианов

- Теперь то же самое для нескольких гауссианов сразу.
- Даны несколько точек $x_1, \ldots, x_n$, но они принадлежат смеси гауссианов с разными μ_k (пусть σ будет одна и та же).
- Тогда распределение будет

$$p(x_1, \ldots, x_n | \{\mu_k\}, \sigma) = \prod_{i=1}^{n} p(x_i | \{\mu_k\}, \sigma).$$

Домашнее задание

Упражнение Разработать алгоритм, который итеративно максимизировал бы функцию правдоподобия при данных тестовых примерах, т.е. находил бы μ_k. За основу можно взять метод итераций Ньютона, который для максимизации функции $f(x)$ итерирует её как

$$x \leftarrow x - \frac{f(x)}{\partial f / \partial x}.$$

- Что должно получиться?

Что должно получиться

- Те, кто может вспомнить прошлый год, вспомните.
- Где мы искали центры смеси гауссианов?

Что должно получиться

- Те, кто может вспомнить прошлый год, вспомните.
- Где мы искали центры смеси гауссианов?
- Правильно, это был алгоритм кластеризации.
- Когда вы сделаете упражнение, у вас должен получиться фактически вариант алгоритма c–средних.
- Отметим, что в данном варианте кластеры будут сферические; чтобы они были эллиптическими, нужно, чтобы дисперсии σ были разными вдоль разных осей.

Outline

Интересные распределения

- Есть некоторое количество распределений, которые появляются в разных задачах и являются наиболее полезными на практике.
- Их полезно... ну, если не знать, то хотя бы быть знакомыми.
- Сейчас по ним и пробежимся.

Биномиальное распределение

- Биномиальное распределение возникает, когда мы подбрасываем нечестную монетку (вероятность решки q) n раз и хотим найти вероятность появления r решек.

$$p(r|q,n) = \binom{n}{r} q^r (1-q)^{n-r}.$$

Пуассоново распределение

- Распределение Пуассона возникает, когда мы хотим подсчитать количество событий за фиксированный интервал, если нам дана средняя интенсивность этих событий.
- Если ожидается в среднем λ событий за этот интервал, то вероятность того, что произойдут ровно r событий, равна

$$p(r|\lambda) = e^{-\lambda}\frac{\lambda^r}{r!}.$$

Экспоненциальное распределение

- Обычно возникает в ответах на вопрос «сколько надо ждать события».
- Например, вероятность того, что выпадения орла на нечестной монетке придётся ждать ровно r шагов, равна

$$p(r|q) = q^r(1-q) = (1-q)e^{-\lambda r}, \quad \lambda = \ln\frac{1}{r}.$$

Гипергеометрическое распределение

- Обычно возникает в ситуациях выбора без замещения.
- Если есть урна, в ней n шаров, из них k чёрных, то вероятность вынуть ровно i чёрных равна

$$p(i|n,k) = \frac{\binom{k}{i}\binom{n-k}{n-i}}{\binom{n}{k}}.$$

Нормальное распределение

- Мы уже давно знаем нормальное распределение:

$$p(x|\mu,\sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

- Очень многие процессы могут моделироваться нормальным (гауссовским) распределением; обычно возникает, когда есть некое среднее значение μ и шум вокруг него.

Распределение Стьюдента

- Распределение Стьюдента применяется, когда нужно искать доверительные интервалы на параметры нормального распределения.
- Величина $T = \frac{\bar{x}_n - \mu}{S_n/\sqrt{n}}$ распределена по закону

$$f(t) = \frac{\Gamma(n/2)}{\sqrt{\pi(n-1)}\Gamma(\frac{n-1}{2})}\left(1 + \frac{t^2}{n-1}\right)^{-n/2}.$$

- Если мы выберем число A так, чтобы $p(-A < T < A) = 1 - \alpha$, то

$$\left[\bar{x}_n - A\frac{S_n}{\sqrt{n}}, \bar{x}_n + A\frac{S_n}{\sqrt{n}}\right]$$

будет доверительным интервалом для μ с вероятностью ошибки α.

Экспоненциальное распределение

- Используется для оценки времени ожидания в пуассоновском процессе.
- Если мы ждём события, которое происходит в среднем каждые $1/\lambda$ единиц времени (с интенсивностью λ), то распределение времени ожидания

$$p(x|s) = \lambda e^{-\lambda x}.$$

Гамма–распределение

- Аналог нормального распределения, но на полуоси $[0, +\infty)$.
- Плотность распределения

$$p(x|k,\theta) = x^{k-1}\frac{e^{-x/\theta}}{\theta^k \Gamma(k)}, \quad x > 0.$$

Бета–распределение

- Бета-распределение определяется над вероятностями, т.е. на интервале $[0, 1]$.
- $$p(x|\alpha, B) = \frac{1}{B(\alpha, B)} x^{\alpha-1}(1-x)^{B-1}, \quad B(\alpha, B) = \frac{\Gamma(\alpha)\Gamma(B)}{\Gamma(\alpha+B)}.$$
- $B(i, j)$ — это распределение i-й по величине случайной величины из $i + j - 1$ случайных величин, распределённых равномерно на $[0, 1]$.

Распределение Дирихле

- Обобщение бета–распределения на многомерный случай.
- На k-мерном симплексе $\{x \mid \sum_{i=1}^{k} x_i = 1\}$ плотность распределения Дирихле с параметром $\alpha = (\alpha_1, \ldots, \alpha_k)$ равна

$$p(x|\alpha) = \frac{1}{B(\alpha)} \prod_{i=1}^{k} x_i^{\alpha_i - 1}, \quad B(\alpha) = \frac{\prod_{i=1}^{k} \Gamma(\alpha_i)}{\Gamma\left(\sum_{i=1}^{k} \alpha_i\right)}.$$

Outline

Why have sex?

- Why have sex?
- К сожалению, каламбур не переводится. :)

R

- Язык программирования *R*.
- Статистический язык, который специально предназначен для решения задач статистики (ну а тем самым, может, и задачам вероятностного обучения полезен будет).
- Нужно рассказать о языке, научить основам его использования.
- http://www.r-project.org/

Кластеризация

- Иерархическая байесовская кластеризация.
- Кластеризация динамических процессов.

Приближенные методы маргинализации

- Маргинализация по Кикучи (Kikuchi).

Спасибо за внимание!

- Lecture notes, слайды и коды программ появятся на моей homepage:

  `http://logic.pdmi.ras.ru/~sergey/index.php?page=teaching`

- Присылайте любые замечания, коды программ на других языках, решения упражнений, новые численные примеры и прочее по адресам:
  `sergey@logic.pdmi.ras.ru`, `smartnik@inbox.ru`