Семинар 6: применение математической статистики в машинном обучении. Наивный байесовский классификатор

Надежда Чиркова

17 февраля 2016 г.

1 Базовые формулы теории вероятностей

Пусть x и y — дискретные случайные величины на конечном носителе[1], то есть, к примеру, y задается конечным множеством значений Y и их вероятностями $p(y), y \in Y$. Мы будем обозначать x и саму случайную величину, и ее носитель[2]. Пусть $p(x, y)$ — их совместная вероятность[3]. *Условной вероятностью* x при условии y называется величина

$$p(x|y) = \frac{p(x, y)}{p(y)} \tag{1}$$

Формула полной вероятности:

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

В соответствии с этой формулой (2) можно переписать так:

$$p(x|y) = \frac{p(x, y)}{\sum_{x'} p(x', y)}. \tag{2}$$

Из определения вероятности и формулы полной вероятности следует *формула Байеса*:

$$p(x, y) = p(x|y)p(y)dy = p(y|x)p(x) \Rightarrow$$

[1] *Носителем* дискретной случайной величины называет множество ее допустимых значений с ненулевой вероятностью; непрерывной случайной величины — множество точек, в которых плотность положительна. Конечный носитель подразумевает, что случайная величина принимает значения из конечного множества.

[2] То есть обозначение $p(x)$ подразумевает «вероятность того, что случайная величина x приняла значение x». Обычно в теории вероятностей случайную величину обозначают ξ и пишут $P(\xi = x)$. Мы отождествим эти два понятии для краткости обозначения и упрощения понимания.

[3] $p(x, y)$ — это вероятность того, что x и y приняли соответственно значения x и y. Она так и задается таблицей на всевозможных парах значений (x, y)., причем все значения в таблице должны суммироваться к единице

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

Аналогично, если x и y — непрерывные случайные величины (то есть имеющие непрерывный носитель), а $p(x, y)$ — их совместная плотность, то все формулы можно переписать с заменой суммирования на интеграл:

$$p(x) = \int p(x|y)p(y)dy = \int p(x, y)dy.$$

$$p(x|y) = \frac{p(x, y)}{\int p(x', y)dx'}.$$

$$p(x|y) = \frac{p(y|x)p(x)}{p(y)} = \frac{p(y|x)p(x)}{\int p(y|x')p(x')dx'}.$$

Интеграл здесь берется по всей числовой прямой. Однако если случайная величина имеет конечный непрерывный носитель, к примеру, отрезок $[a, b]$, то плотность вне этого отрезка равна нулю, и интеграл берется по этому отрезку.

Математическое ожидание вычисляется как

$$\mathrm{E}x = \sum_x xp(x)$$

для дискретной случайной величины и

$$\mathrm{E}x = \int xp(x)dx$$

для непрерывной случайной величины.

2 Метод максимального правдоподобия

Пусть нам дана выборка $X = \{x_1, \ldots, x_n\}$, сгенерированная из распределения $p(x|\theta)$, имеющего параметры θ: $x_i \sim p(x|\theta), i = 1, \ldots, n$. Например, если $p(x|\theta) = \mathcal{N}(x|\mu, \sigma)$, то $\theta = \{\mu, \sigma\}$. Требуется по этой выборке оценить параметры θ. Для этого часто используют *метод максимального правдоподобия*. Правдоподобие выборки (вероятность получить именно такую выборку) записывается как

$$p(X) = \prod_{i=1}^{n} p(x_i|\theta).$$

При фиксированном распределении $p(x|\theta$ это конкретная функция, зависящая от θ. Значит, ее можно максимизировать:

$$\prod_{i=1}^{n} p(x_i|\theta) \to \max_{\theta}.$$

В сложных вероятностных моделях это приходится делать с помощью численных методов, например метода градиентного подъема, который разбирался на прошлом семинаре[4]. В наших задачах эту максимизацию удается провести аналитически.

Для упрощения дифференцирования (которое, как мы знаем, в задачах оптимизации практически неизбежно) всегда рекомендуется переходить к логарифму правдоподобия. Это не меняет точки его оптимума.

Пример (одномерное нормальное распределение). Пусть $x_i \sim \mathcal{N}(x|\mu,\sigma), i = 1,\dots,n$. Оценить μ и σ методом максимального правдоподобия.

Решение:

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

Тогда правдоподобие равно

$$L = p(X|\mu,\sigma) = \prod_i \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n \prod_i e^{-\frac{(x_i-\mu)^2}{2\sigma^2}} = \left(\frac{1}{\sqrt{2\pi}\sigma}\right)^n e^{-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}} \to \max_{\mu,\sigma}$$

Дифференцируем логарифм правдоподобия выборки и находим решение:

$$\frac{\partial \log L}{\partial \mu} = \frac{\partial}{\partial \mu}\left(-\frac{\sum_i (x_i-\mu)^2}{2\sigma^2}\right) = -\frac{2\sum_i (x_i-\mu)}{2\sigma^2} = 0 \Rightarrow \mu = \frac{1}{n}\sum_i x_i.$$

$$\frac{\partial \log L}{\partial \sigma} = -n\frac{1}{\sigma} + 2\frac{(x-\mu)^2}{2\sigma^3} = 0 \Rightarrow \sigma^2 = \frac{1}{n}\sum_i (x_i-\mu)^2.$$

Равенство нулю производной (градиента) $\log L$ в точке x_0 — необходимое условие экстремума (вспоминаем прошлый семинар). Если продифференцировать $\log L$ второй раз по обеим переменным, можно убедиться, что эти значения действительно максимизируют значение правдоподобия.

В машинном обучении выборки обычно многомерные, поэтому вместо производной приходится считать градиент и приравнивать его к нулю.

Пример (многомерное нормальное распределение). Пусть $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu},\Sigma), i = 1,\dots,n$ — многомерное нормальное распределение[5]. Оценить $\boldsymbol{\mu}$ и Σ методом максимального правдоподобия.

[4]Когда градиентный метод применяется к задаче минимизации функции, он называется методом градиентного спуска, а когда к задаче максимизации — методом градиентного подъема.

[5]Основную информацию о многомерном нормальном распределении можно прочитать в Википедии. Самое важное: оно унимодальное (один пик), его математическое ожидание равно $\boldsymbol{\mu}$, а матрица ковариаций компонент равна Σ. Матрица ковариаций Σ симметрична.

Решение: Выпишем формулу правдоподобия выборки (как мы выяснили на прошлом семинаре, $\boldsymbol{x}^T A \boldsymbol{x} = \sum_i \sum_j a_{ij} x_i x_j$ — это число):

$$L = p(X|\boldsymbol{\mu}, \Sigma) = \prod_i \frac{1}{\sqrt{2\pi}^n \sqrt{|\Sigma|}} e^{-(\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})} = \frac{1}{\sqrt{2\pi}^n \sqrt{|\Sigma|}} e^{-\sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu})} \to \max_{\mu, \Sigma}.$$

Дифференцируем логарифм правдоподобия выборки и находим решение:

$$\frac{\partial \log L}{\partial \boldsymbol{\mu}} = \frac{\partial}{\partial \boldsymbol{\mu}} \left(-\sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})^T \Sigma^{-1} (\boldsymbol{x}_i - \boldsymbol{\mu}) \right) = \sum_i \frac{\partial}{\partial \boldsymbol{\mu}} \left(-\boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{x}_i - \boldsymbol{x}_i^T \Sigma^{-1} \boldsymbol{\mu} - \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{x}_i + \boldsymbol{\mu}^T \Sigma^{-1} \boldsymbol{\mu} \right) =$$

$$= \sum_i \left(-2 \boldsymbol{x}_i^T \Sigma^{-1} + \left(\Sigma^{-1} + (\Sigma^{-1})^T \right) \boldsymbol{\mu} \right) = -2 \sum_i \boldsymbol{x}_i^T \Sigma^{-1} + 2n \Sigma^{-1} \boldsymbol{\mu} = 0 \Rightarrow \boldsymbol{\mu} = \frac{1}{n} \sum_i \boldsymbol{x}_i.$$

Последний переход сделан домножением всего равенства на матрицу Σ справа. Также применены факт о том, что $\boldsymbol{q}^T \boldsymbol{p} = \boldsymbol{p}^T \boldsymbol{q}$ для любых двух векторов, и симметричность матрицы Σ; $\boldsymbol{x}^T \Sigma^{-1}$ — вектор!

Аналогично можно выполнить дифференцирование по матрице Σ и получить, что $\Sigma = \frac{1}{n} \sum_i (\boldsymbol{x}_i - \boldsymbol{\mu})^T (\boldsymbol{x}_i -_{\boldsymbol{m}} \boldsymbol{u})$, однако это выходит за границы программы семинаров.

3 Байесовский классификатор

Вопрос для самостоятельного ответа: в чем заключается задача классификации?

Пусть $X \in R^{n \times d}$ — выборка, $y \in R^n$ — вектор правильных ответов в задаче классификации (y_i принадлежат конечному множеству меток классов $1, \ldots, K$). Будем предполагать, что выборка сгенерирована из распределения $p(\boldsymbol{x}|\theta)$, но параметры θ свои у каждого класса:

$$\boldsymbol{x}_i \sim p(\boldsymbol{x}_i|y_i) = p(\boldsymbol{x}_i|\theta_{y_i}).$$

Мы сможем оценить параметры распределения методом максимального правдоподобия отдельно для каждого класса, составив K соответствующих подвыборок.

Теперь, чтобы предсказать класс для нового объекта $\boldsymbol{x}$, мы найдем максимум $p(y|x)$:

$$y = \text{argmax}_y p(y|x) = \text{argmax}_y \frac{p(x|y)p(y)}{p(x)}. \tag{3}$$

Знаменатель не зависит от y, поэтому мы можем его не учитывать:

$$y = \text{argmax}_y p(x|y)p(y).$$

Априорные вероятности классов $p(y)$ выбираются исходя из задачи. Их можно брать равномерными, пропорциональными долям классов выборке (нехороший способ, так как соотношение классов в выборке не всегда соответствует их соотношению в реальной жизни) или из знаний

о рассматриваемой области. Например, мы можем знать, что женщин на планете чуть больше, чем мужчин, и установить вероятности как $p(women) = 0.55$, $p(men) = 0.45$.

Полученный классификатор называется *байесовским*. Его предположения кроются в выбранном распределении $p(x|\theta)$, а модель представлена формулой (3). Интересно, что его оптимизационная задача решается аналитически, поэтому процедура обучения выполняется очень быстро. Недостаток же состоит в том, что нужно уметь грамотно выбирать $p(x|\theta)$.

Многомерных распределений меньше, чем одномерных, и находить для них параметры, как мы увидели в задаче, несколько сложнее. Поэтому часто применяют *наивный* байесовский классификатор, в котором предполагается, что $p(\boldsymbol{x}|y) = \prod_{i=1}^{d} p(x_i|y)$, то есть вместо сложной многомерной плотности, учитывающей взаимодействия между признаками, используют одномерные плотности.

Например, если оценить среднее μ и дисперсию σ нормального распределения отдельно для каждого признака в каждом классе (как в первой задаче), то мы получим нормальный наивный байесовский классификатор. Он также реализован в sklearn.