УДК 004

НУРЕДДИН БЕННИС АДЕЛЬ, студент группы BIT15
Университет Кади Айяд, КФУ
Марокко, г. Марракеш

НЕГАУССОВСКИЕ ПРИЗНАКИ В ДАТАСЕТАХ МАШИННОГО ОБУЧЕНИЯ

Рассмотрим один из наиболее общих случаев, когда признаки в наборе данных не будут являться гауссовскими (то есть распределенными по нормальному закону).

Для рассмотрения общего случая необходимо обратиться к экспоненциальным плотностям распределений, поскольку векторы ответов являются классами, а описывать необходимо признаки. Предположим, что каждый признак j имеет некоторую одномерную плотность распределения и эта плотность принадлежит экспоненциальному семейству.

Сформируем формулу (1), которая позволит нам применить принцип максимума правдоподобия.

$$p\left(x^j \middle| y;\ \theta_{yi}, \varphi_{yi}\right) = \exp\left(\frac{x^j \theta_{yi} - c\left(\theta_{yi}\right)}{\varphi_{yi}} + h\left(x^j, \varphi_{yi}\right)\right), \quad (1)$$

где $\theta_{yi}, \varphi_{yi} -$ параметры, а $c(\theta),\ h(y, \varphi) -$ параметры функции.

Плотность определена с точностью до произвольной функции $c(\theta)$ и произвольной функции $h(y, \varphi)$, которые не зависят от x и не зависят от θ. Далее мы будем фиксировать функциональные параметры $c(\theta),\ h(y, \varphi)$ и по принципу максимума правдоподобия настраивать числовые параметры $\theta_{yi}, \varphi_{yi}$. Запишем принцип максимума правдоподобия (2). Сразу переставим знаки алгебраической суммы, вынеся за скобку сумму по признакам и сумму по классу. Благодаря этому внутри скобок останется сумма по всем объектам одного класса.

$$L(\theta, \varphi) = \sum_{j=1}^{n} \sum_{y\ \epsilon\ Y} \left(\sum_{x_i \epsilon X_y} \ln \mathrm{p}\left(x^j \middle| y;\ \theta_{yi}, \varphi_{yi}\right) \right) \rightarrow \max_{\theta, \varphi} \quad (2)$$

Две внешние суммы будут отражать математическую операцию декомпозирования и факторизации: для каждого сочетания признака и класса возникнет отдельная задача на параметры θ_{yi} и φ_{yi}, которые зависят только от этого признака и только от этого класса. В этой концепции и состоит основная решающая система предположения о независимости: мы можем отдельно оценить все параметры всех распределений по каждому объекту в каждом признаке отдельно.

Задача распадается на независимые подзадачи для каждого (y, i) (формула (3)).

$$\sum_{x_i \epsilon X_y} \left(\frac{x^j \theta_{yi} - c(\theta_{yi})}{\varphi_{yi}} + h(x^j, \varphi_{yi}) \right) \rightarrow \max_{\theta_{yi}, \varphi_{yi}} \quad (3)$$

Рассмотрим наиболее частные случаи распределений. Для удобства распределения представлены в виде приведенной обобщенной формулы (3) к частному виду для каждого отдельного вида (рисунок 1).

$$\frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) = \exp\left(\frac{x\mu - \frac{1}{2}\mu^2}{\sigma^2} - \frac{x^2}{2\sigma^2} - \frac{1}{2}\ln(2\pi\sigma^2)\right)$$

$$\mu^x (1-\mu)^{1-x} = \exp\left(x \ln \frac{\mu}{1-\mu} + \ln(1-\mu)\right)$$

$$C_k^x \mu^x (1-\mu)^{k-x} = \exp\left(x \ln \frac{\mu}{1-\mu} + k \ln(1-\mu) + \ln C_k^x\right)$$

$$\frac{1}{x!} e^{-\mu} \mu^x = \exp\left(x \ln(\mu) - \mu - \ln x!\right)$$

распределение	значения	$c(\theta)$	$c'(\theta)$	$[c']^{-1}(z)$	φ	$h(x, \varphi)$
нормальное	$\mathbb{R}$	$\frac{1}{2}\theta^2$	θ	z	σ^2	$-\frac{x^2}{2\varphi} - \frac{\ln(2\pi\varphi)}{2}$
Бернулли	$\{0,1\}$	$\ln(1+e^\theta)$	$\frac{1}{1+e^{-\theta}}$	$\ln\frac{1-z}{z}$	1	0
биномиальное	$\{0,\ldots,k\}$	$k\ln(1+e^\theta)$	$\frac{k}{1+e^{-\theta}}$	$\ln\frac{k-z}{z}$	1	$\ln C_k^x$
Пуассона	$\{0,1,\ldots\}$	e^θ	e^θ	$\ln z$	1	$-\ln x!$

Рисунок 1. Приведение распределений к экспоненциальной форме

Если признак числовой, то очень часто (с некоторым приближением) можно считать, что он распределён нормально. Если признак не распределён по закону Гаусса, то необходимо использовать какие-то другие распределения. Так, распределение Бернулли относится в большинстве случаев к бинарным (дихотомическим) признакам. Биномиальное распределение относится к целочисленным признакам, у которых есть границы сверху (в частности, номинальные признаки). Распределение Пуассона – это распределение, которое предназначается для целочисленных признаков, у которых нет границы сверху (например, это число кликов или число посещений). То есть к этой категории относятся все признаки, для которых нельзя указать четкий порог верхней границы значений, поэтому их можно описывать пуассоновским распределением.

На рисунке 1 красным цветом выделены составные множители в функции, которые являются параметром θ, а также то, каким образом этот параметр связан с математическим ожиданием распределения. Математическое ожидание присутствует здесь в обязательном порядке и всегда, так как исходно во всех семействах параметрических распределений одним из аргументов функции является математическое ожидание. Кроме этого, параметр φ оценивать очень просто во всех случаях: в распределении Гаусса это дисперсия, а во всех остальных случаях φ принимается равным 1. Теперь уже существует основа для формирования аналитического решения задачи максимума правдоподобия в случае, когда все признаки независимы и описываются экспоненциальным законом. Решение в виде

строгой математической формулы записывается в одну строчку. Для этого нам необходимо взять функционал логарифма правдоподобия (5) и его производную по θ, приравниваем выражение к нулю. Из этого сразу получается, что производная $c'(\theta_{yj})$ равна среднему значению признака j в классе y.

Список литературы:

1. Акилина, М. В. Использование технологии data cleaning для очистки большого объёма данных / М. В. Акилина, П. А. Пылов, А. В. Протодьяконов // Россия молодая : СБОРНИК МАТЕРИАЛОВ XII ВСЕРОССИЙСКОЙ, НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ МОЛОДЫХ УЧЕНЫХ С МЕЖДУНАРОДНЫМ УЧАСТИЕМ, Кемерово, 21–24 апреля 2020 года. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2020. – С. 21118.1-21118.5. – EDN ICXURQ.
2. Пылов, П. А. Сопоставление оценщиков и алгоритмов искусственного интеллекта для обнаружения аномалий в наборах данных / П. А. Пылов // Россия молодая : СБОРНИК МАТЕРИАЛОВ XII ВСЕРОССИЙСКОЙ, НАУЧНО-ПРАКТИЧЕСКОЙ КОНФЕРЕНЦИИ МОЛОДЫХ УЧЕНЫХ С МЕЖДУНАРОДНЫМ УЧАСТИЕМ, Кемерово, 21–24 апреля 2020 года. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2020. – С. 21117.1-21117.4. – EDN GDIUOK.
3. Стародубов, А. Н. Модернизация архитектуры скрытой марковской модели как новая основа эффективного решения задачи интеллектуального учета энергопотребления / А. Н. Стародубов, П. А. Пылов // Инновации в топливно-энергетическом комплексе и машиностроении (ТЭК-2022) : сборник трудов III Международной научно-практической конференции, Кемерово, 19–21 апреля 2022 года. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2022. – С. 28-36. – EDN RBAYOK.
4. Пылов, П. А. Обработка естественного языка в прикладной задаче ранжирования сложности философских трудов по авторам произведений / П. А. Пылов, О. А. Ивина // Россия молодая : Сборник материалов XIV Всероссийской научно-практической конференции с международным участием, Кемерово, 19–21 апреля 2022 года / Редколлегия: К.С. Костиков (отв. ред.) [и др.]. – Кемерово: Кузбасский государственный технический университет имени Т.Ф. Горбачева, 2022. – С. 31524.1-31524.5. – EDN TYTSNX.
5. Стародубов, А. Н. Модель искусственного интеллекта на основе сверточной нейронной сети для цифрового распознавания рукописных цифр / А. Н. Стародубов, П. А. Пылов // Системы автоматизации (в образовании, науке и производстве) : AS'2021 : труды XIII Всероссийской научно-практической конференции

(с международным участием), Новокузнецк, 02–03 декабря 2021 года. – Новокузнецк: Сибирский государственный индустриальный университет, 2021. – С. 354-358. – EDN VHMHRJ.

6. Свидетельство о государственной регистрации программы для ЭВМ № 2021666164 Российская Федерация. The White Mind : № 2021665235 : заявл. 30.09.2021 : опубл. 08.10.2021 / П. А. Пылов, Р. В. Майтак, А. В. Протодьяконов. – EDN VYTFQC.