Machine Learning
Section 5: Continuous Probabilties
第5节：连续概率

Stefan Harmeling

13. October 2021

某一不确定量p的先验概率（Prior probability）分布是在考虑"观测数据"前，能表达p不确定性的概率分布。它旨在描述这个不确定量的不确定程度，而不是这个不确定量的随机性。

在贝叶斯统计中，一个随机事件或者一个不确定事件的后验概率（Posterior probability）是在考虑和给出相关证据或数据后所得到的条件概率。同样，后验概率分布是一个未知量（视为随机变量）基于试验和调查后得到的概率分布。"后验"在本文中代表考虑了被测试事件的相关证据。

Probabilities on continuous variables

连续变量的概率

From discrete to continuous variables

from Murphy 2012, p32, from Jaynes 2003, p107   从离散变量到连续变量

- let X be real-valued variable   设X为实值变量
- define propositions $A = (X \leq a)$, $B = (X \leq b)$ and $W = (a < X \leq b)$ with $a < b$
- note that $E_B = E_{A \lor W}$
- note that A and W are mutually exclusive, thus sum rule:

$$p(W) = p(B) - p(A)$$

- with $F(x) := p(X \leq x)$ and $f(x) = \frac{d}{dx} F(x)$ we get

$$p(a < X \leq b) = F(b) - F(a) = \int_a^b f(x) dx$$

- F is called *cumulative distribution function* (CDF) and f is called *probability density function* (PDF)

概率密度函数(PDF)

Discrete vs. continuous probabilities (1)

概率论作为命题逻辑的延伸

Probability theory as an extension of propositional logic

- finite set of propositional variables $A, B, \ldots, Z \in \{0, 1\}$ jointly ranging over all boolean assignments

  样本空间Ω={所有布尔赋值}。

- sample space $\Omega = \{$all boolean assignments$\}$
- probability mass function $f: \Omega \to [0, 1]$, such that $\sum_{\omega \in \Omega} f(\omega) = 1$

Discrete probability theory (includes the previous case)离散概率

- random variable ranging in a discrete set, e.g. $\{0, 1, 2, \ldots, \}$
- sample space $\Omega = \{0, 1, 2, \ldots\}$
- *probability mass function $f : \Omega \to [0, 1]$, such that $\sum_{\omega \in \Omega} f(\omega) = 1$*

Continuous probability theory

- random variable ranging in a continuous set, e.g. real numbers $\Re$
- sample space $\Omega = \Re$
- *probability density function $f : \Omega \to \Re_+$, such that $\int_{\omega \in \Omega} f(\omega) d\omega = 1$*

Discrete vs. continuous probabilities (2)

概率衡量样本空间的子集E`Ω的质量

Probabilities measure the mass of subsets $E \subset \Omega$ of the sample space.

Discrete probabilities

- e.g. $\Omega = \mathbb{N}$
- probability mass function $f : \mathbb{N} \to [0, 1]$, such that $\sum_n f(n) = 1$
- probability $p(E) = \sum_{n \in E} f(n)$
- small letter p

Continuous probabilities

- e.g. $\Omega = \mathfrak{R}$
- probability density function (PDF) $f : \mathfrak{R} \to \mathfrak{R}_+$, such that $\int f(x) dx = 1$
- probability $P(E) = \int_E f(x) dx$
- large letter P

小字母p

Rules for continuous variables

Theorem 5.1 (rules for PDFs)

The standard rules of probability theory do hold for PDFs,

$$f(x, y) = f(x|y)\, f(y) \qquad \textit{product rule}$$
$$f(x) = \int f(x, y)\, dy \qquad \textit{sum rule}$$

For that reason we often write p for PDFs. Sums turn into integrals. Note that the product rule implies Bayes' rule.

由于这个原因，我们经常
把p写成PDF。和变成了
积分。请注意，乘积规则
意味着贝叶斯规则。

Different way to think about the probability rules

▸ consider three random variables A, B, C with values a, b, c

▸ for probabilistic inference we need the joint PDF $p(a, b, c)$

▸ assume we get $p(c)$, $p(b|c)$, $p(a|b, c)$ from domain expert

▸ define the joint PDF $p(a, b, c) := p(a|b, c)p(b|c)p(c)$

▸ define all partial joints by integration, e.g.

$$p(a, b) := \int p(a, b, c)dc$$

▸ now all sum rules hold! (also for $p(c)$ used in the definition)

▸ define the conditional PDF as quotients, e.g.

$$p(c|a) := p(a, c)/p(a)$$

▸ now all product rules hold (also for $p(a|b, c)$, $p(b|c)$)

All rules hold by definition!

Random variable

Definition 5.2 离散随机变量X

A discrete random variable X is a variable with values x ranging over some discrete set $\mathcal{X}$ and a probability mass function (PMF)

$$f : \mathcal{X} \to [0, 1]$$
$$x \mapsto f(x)$$

概率质量函数（PMF

that sums up to one, $\sum_{x \in \mathcal{X}} f(x) = 1$.

Definition 5.3

A continuous random variable X is a variable with values x ranging over some continuous set $\mathcal{X}$ and a probability density function (PDF)

$$f : \mathcal{X} \to \mathbb{R}_{\geq 0}$$
$$x \mapsto f(x)$$

that integrates to one, $\int_{x \in \mathcal{X}} f(x)\, dx = 1$.

Random variable and its values

RV用大写字母X表示，它的值用小写字母x
更确切地说（也很复杂）：小写的x也是一个变量。

Note:

- RV are denoted by capital letters X, its values by small letters x
- More precisely (and complicated): the small x is a variable as well, that ranges over the values of the *random variable X*.
- Using the same letter, we know which small letter variable corresponds to which random variable (denoted by a capital letter).
- So, when we write $p(x)$ we are talking about the PDF with input x that belongs to random variable X. When we write $p(y)$ ...

Example – infer probability of wearing glasses

例子 - 推断戴眼镜的概率

Example — inferring probability of wearing glasses (1)

Question:

What's the probability that a person wears glasses?

一个人戴眼镜的概率是多少?

# Example — inferring probability of wearing glasses (2)

Represent all unknowns as random variables (RVs)将所有未知数表示为随机变量

- probability to wear glasses is represented by RV *Y*戴眼镜的概率用RV *Y*表示
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5 Possible values of the RVs 五个观察值用-----表示 RV的可能值

- *Y* takes values $y \in [0, 1]$ (it is a probability value)Y的取值y>[0，1]（它是一个概率值）。
  - X_1, X_2, X_3, X_4, X_5 are binary, i.e. values 0 and 1

Graphical representation
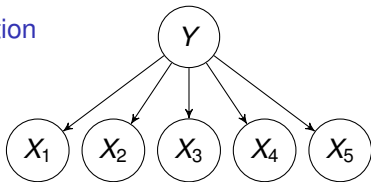
Generative model and joint probability

似然反映的是：已知结果，反推原因

- we abbreviate $Y = y$ as y, $X_i = x_i$ as x_i
- $p(y)$ is the prior of Y, written fully $p(Y = y)$ $p(y)$是Y的先验，完全写成$p(Y=y)$
- $p(x_i|y)$ is the likelihood of observation x_i $p(xi |y)$是观察点xi的似然性
- note that the likelihood is a function of y 注意，这个似然是Y的一个函数

Example — inferring probability of wearing glasses (3)

Probability of wearing glasses without observations

在没有观察的情况下，佩戴眼镜的概率

$$p(y|\text{"nothing"}) = p(y)$$

Probability of wearing glasses after one observation

观察一次后戴眼镜的概率

$$p(y|x_1) = Z_1^{-1} p(x_1|y)p(y)$$

Probability of wearing glasses after two observations

$$p(y|x_1, x_2) = Z_2^{-1} p(x_2|x_1, y)p(x_1|y)p(y) = Z_2^{-1} p(x_2|y)p(x_1|y)p(y)$$
$\cdots$

Probability of wearing glasses after five observations

$$p(y|x_1, x_2, x_3, x_4, x_5) = Z_5^{-1} \left(\prod_{i=1}^{5} p(x_i|y) \right) p(y)$$

Example — inferring probability of wearing glasses (4)

What is the likelihood?

$$p(x_1|y) = \begin{cases} y & \text{for } x_1 = 1 \\ 1 - y & \text{for } x_1 = 0 \end{cases}$$

More helpful RVs:

N表示观察值为1的数量

- ▸ RV N for the number of observations being 1 (with values n)
- ▸ RV M for the number of observations being 0 (with values m)

Probability of wearing glasses after five observations

$$\begin{aligned} p(y|x_1, x_2, x_3, x_4, x_5) &= Z_5^{-1} \left(\prod_{i=1}^{5} p(x_i|y) \right) p(y) \\ &= Z_5^{-1} y^n (1 - y)^m p(y) \\ &= p(y|n, m) \end{aligned}$$

Example — inferring probability of wearing glasses (5)

Posterior after seeing five observations:在看到五次观察后的后方。

$$p(y|n, m) = Z_5^{-1} y^n (1 - y)^m p(y)$$

What prior $p(y)$ would make the calculations easy?

什么样的先验p(y)会使计算变得容易？

$$p(y) = Z^{-1} y^{a-1} (1 - y)^{b-1} \qquad \text{with parameters } a > 0, b > 0$$

called "Beta distribution with parameter a and b" 称为 "参数为a和b的贝塔分布"

Let's give the normalization factor Z of the beta distribution a name!

$$B(a, b) = \int_0^1 y^{a-1} (1 - y)^{b-1} dy$$

called "Beta function with parameters a and b"

Note: for $a = 1, b = 1$, the Beta distribution is the uniform distribution $p(y) = 1$ for $y \in [0, 1]$, zero elsewhere. 对于a=1，b=1，Beta分布是均匀分布p(y)=1对于y[0，1]，其他地方为零。

Gamma function, Beta function, and all that

from http://en.wikipedia.org/wiki/Gamma_function
and http://en.wikipedia.org/wiki/Beta_function

Gamma function (extension of factorial function)

$$\Gamma(z) = \int_0^\infty e^{-t} t^{z-1} \, dt \qquad \text{for } z \in \mathbb{C}$$

$$\Gamma(n) = (n-1)! = n!/n \qquad \text{for } n \in \mathbb{N}$$

Beta function (extension of ... ?)

$$B(x, y) = \int_0^1 t^{x-1} (1-t)^{y-1} \, dt$$

$$= \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} \qquad \text{for } x, y \in \mathbb{C} \text{ with } x + \bar{x}, y + \bar{y} > 0$$

$$B(m, n) = \frac{(m-1)! \, (n-1)!}{(m+n-1)!} \qquad \text{for } m, n \in \mathbb{N}$$

$$= \left(\begin{array}{c} m+n \\ n \end{array} \right)^{-1} \frac{m+n}{m \, n} \qquad \text{binomial coefficient}$$

Example — inferring probability of wearing glasses (6)

The prior of the probability with parameters *a* and *b*:

$$p(y) = \frac{y^{a-1}(1-y)^{b-1}}{B(a,b)}$$

The likelihood of the observations:

$$p(n, m \mid y) = y^n(1-y)^m$$

The posterior with the beta prior:

$$p(y|n,m) = \frac{y^{n+a-1}(1-y)^{m+b-1}}{B(a+n, b+m)}$$

Note:

▶ If the prior and the posterior have the same form (here: both are Beta distribution), we call prior the *conjugate* prior for likelihood. See https://en.wikipedia.org/wiki/Conjugate_prior.

共轭先验

Expectations 期望

Integration with densities

Weighted average/sum

$$\sum_i w_i x_i$$

with weights w_i

Weighted integral

$$\int w(x) x \, dx$$

with weights/density $w(x)$

Expectations of a random variable (1)

Definition 5.4 (expected value of *x*)

1. *The expected value of a continuous random variable X with PDF p(x) is:*

$$E X = E(X) = \int x \, p(x) dx$$

2. *The expected value of a discrete random variable X with probability mass function p(x) is:*

$$E X = E(X) = \sum_x x \, p(x)$$

▸ The E turns a random variable *X* into a single number (usually into a single value from its range).

Expectations of a random variable (2)

Definition 5.5 (expected value of a function of x)

1. *The expected value of a function f wrt the continuous random variable X with PDF $p(x)$ is:*

$$E f(X) = E(f(X)) = \int f(x)p(x)dx$$

2. *The expected value of a function f wrt the discrete random variable X with probability mass function $p(x)$ is:*

$$E f(X) = E(f(X)) = \sum_x f(x)p(x)$$

- $f(X)$ can be also understood as a new random variable $Y = f(X)$.
- Sometimes we write E_x to specify which variable is summed up. Note that we use a small x in those cases since the subindex to E binds the variable and makes it local, e.g. $E_x x$.
- "wrt" = "with respect to"

Expectations of a random variable (3)

Examples

- Mean $\qquad\qquad E\,X = E(X) = E_x\,x$

- Variance

$$\mathrm{Var}\,X = \mathrm{Var}(X) = E(X - \mu)^2 = E_x(x - \mu)^2$$

  is the average squared distance to the mean $\mu = E_x\,x$

- Product- and sum-rule combined

$$p(y) = \int p(x, y)\,dx = \int p(y|x)\,p(x)\,dx = E_x\,p(y|x) = E\,p(y|X)$$

- Probabilities as expectations

$$p(X \in A) = E(1_A(X)) = E[X \in A]$$

  with Iverson bracket $[F] = 1$ if formula F is true, and zero otherwise.