Machine Learning

Section 5: Continuous Probabilties

Stefan Harmeling

20. October 2021

Probabilities on continuous variables

From discrete to continuous variables

from Murphy 2012, p32, from Jaynes 2003, p107

- let X be real-valued variable
- define propositions $A = (X \leq a)$, $B = (X \leq b)$ and $W = (a < X \leq b)$ with $a < b$
- note that $E_B = E_{A \vee W}$
- 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)

Discrete vs. continuous probabilities (1)

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

Discrete probability theory (includes the previous case)

- 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 $\mathfrak{R}$
- sample space $\Omega = \mathfrak{R}$
- *probability density function* $f : \Omega \to \mathfrak{R}_+$, such that $\int_{\omega \in \Omega} f(\omega) d\omega = 1$

Discrete vs. continuous probabilities (2)

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. $\Omega = \Re$
- probability density function (PDF) $f : \Re \to \Re_+$, such that $\int f(x)dx = 1$
- probability $P(E) = \int_E f(x)dx$
- large letter 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.

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 PDFs $p(c)$, $p(b|c)$, $p(a|b, c)$ 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 $p(a|b, c)$, $p(b|c)$)
- ...and this is compatible with $p(a, b, c) := p(a|b, c)p(b|c)p(c)$

Using these definitions all rules hold automatically!

Definition of random variable

Definition 5.2

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

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

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)

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

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

Range of a random variable

- Consider a random variable X with values x ranging over some set $\mathcal{X}$.
- Sometimes we call $\mathcal{X}$ the *range* of X.
- Sometimes we call the subset of $\mathcal{X}$ the *range* of X where $p(x) > 0$.

Random variables and their values

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). This is just a useful convention!
- So, when we write $p(x)$ we are talking about the PDF with input x that belongs to random variable X. When we write $p(y)$ ...

Sidenote: what really is a random variable?

In mathematics:

- A random variable is a (measurable) mapping from the sample space Ω to the real numbers, i.e.

$$X : \Omega \to \mathbb{R}$$

- Assuming we have a probability measure P on Ω we get one for X as well, e.g. for $E \subset \mathbb{R}$ we choose $P(E) = P(X^{-1}(E))$ since $X^{-1}(E) \subset \Omega$.
- If the distribution of X is *absolutely continuous with respect to the Lebesgue measure* we have a density (i.e. a PDF).

Here in this lecture:

- For simplicity we always assume that we have a PDF for our continuous random variables.
- Let's view a random variable to be a variable that also knows its range and its distribution either represented by its PDF (for cont. X) or by its PMF, probability mass function (for discrete X).
- A random variable could have any type, e.g. also being a function itself like for Gaussan processes.

Expectations

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 discrete random variable X with probability mass function $p(x)$ is:*

$$\mathrm{E}\,X = \mathrm{E}(X) = \sum_x x\,p(x)$$

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

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

- The operator 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 discrete random variable X with probability mass function $p(x)$ is:*

$$\mathrm{E}\, f(X) = \mathrm{E}(f(X)) = \sum_x f(x)p(x)$$

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

$$\mathrm{E}\, f(X) = \mathrm{E}(f(X)) = \int f(x)p(x)dx$$

- $f(X)$ 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. $\mathrm{E}_x\, x$.
- "wrt" = "with respect to"

Expectations of a random variable (3)

Examples

- Mean $$\mathrm{E}\,X = \mathrm{E}(X) = \mathrm{E}_x\, x$$
- Variance

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

  is the average squared distance to the mean $\mu = E_x x$
- Product- and sum-rule combined

$$p(y) = \int p(x,y)\,dx = \int p(y|x)\,p(x)\,dx = \mathrm{E}_x\, p(y|x) = \mathrm{E}\, p(y|X)$$

- Probabilities as expectations

$$p(X \in A) = \mathrm{E}\,1_A(X) = \mathrm{E}[X \in A]$$

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

Example – infer probability of wearing glasses

Example — inferring probability of wearing glasses (1)

Question:

What's the probability that a person wears glasses?

Example — inferring probability of wearing glasses (2)

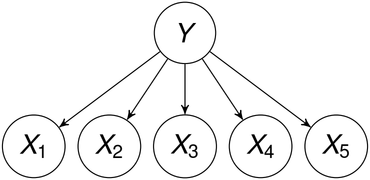
Represent all unknowns as random variables (RVs)

- probability to wear glasses is represented by RV Y
- five observations are represented by RVs X_1, X_2, X_3, X_4, X_5

Possible values of the RVs

- Y takes values $y \in [0, 1]$ (it is a probability value)
- 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

- we abbreviate $Y = y$ as y, $X_i = x_i$ as x_i
- $p(y)$ is the prior of Y, written fully $p(Y = y)$
- $p(x_i|y)$ is the likelihood of observation x_i
- note that the likelihood is a function of y

Example — inferring probability of wearing glasses (3)

Probability of wearing glasses without observations

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

Probability of wearing glasses after one observation

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

Probability of wearing glasses after two observations

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

...

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:

- 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:

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

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

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

called "Beta distribution with parameter a and b"

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.

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 ...?)

$$\begin{aligned} B(x,y) &= \int_0^1 t^{x-1}(1-t)^{y-1} \, dt \\ &= \frac{\Gamma(x)\Gamma(y)}{\Gamma(x+y)} && \text{for } x, y \in \mathbb{C} \text{ with } x+\bar{x}, y+\bar{y} > 0 \\ B(m,n) &= \frac{(m-1)!\,(n-1)!}{(m+n-1)!} && \text{for } m, n \in \mathbb{N} \\ &= \binom{m+n}{n}^{-1} \frac{m+n}{m\,n} && \text{binomial coefficient} \end{aligned}$$

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`.

Summary

From discrete to continuous random variables:

- based on probabilities for discrete variables, we can introduce probabilities for continuous variables (construction due to E.T. Jaynes (see his book “Logic of Science”)
- sum- and product-rule hold for PDFs
- definition of random variables
- expectations of random variables

Example: wearing glasses

- viewing a parameter as a random variable
- Beta distribution
- conjugate prior