Probability refresher

Herman Kamper

2023-02

Overview

In many machine learning problems it is useful to have a way to deal with uncertainty.

Probability theory gives us a principled way to do this.

A probabilistic perspective is also often useful for defining and combining loss functions. For regression this was relatively easy, but for other tasks like classification it isn't always obvious which loss to use.

This note recaps the basics of probability theory.

If anything in this note looks unfamiliar, please revise these concepts from your previous probability theory or statistics course.

Notation

Some text books make a clear distinction between a random variable X and a specific value x that this random variable can take on.

Some (many/most) machine learning text books take a massive shortcut, which I will also use:

- x can refer either to the random variable itself or the value which it takes on—this is inferred from the context.
- Densities are simply denoted as $p(x)$, without a subscript indicating the random variable.
- For discrete distributions, $P(x)$ denotes the probability mass function.
- Sometimes $\Pr(\cdot)$ is used to explicitly denote the probability of some event.

For expected values, different symbols are used: $\mathbb{E}[\cdot]$, $E[\cdot]$, $\mathcal{E}[\cdot]$, $< \cdot >$

I will normally use $\mathbb{E}[\cdot]$, but sometimes I could also write $\mathbb{E}_{p(x)}[\cdot]$ to explicitly denote the density (and the random variable) over which the expectation is taken. Some people might use $\mathbb{E}_X[\cdot]$ for this purpose.

We can use the following shorthand to indicate the distribution which a variable takes on:

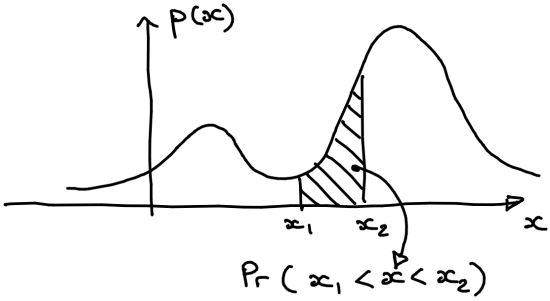
$$z \sim \mathcal{N}(\mu, \sigma^2)$$

This means that z is a sample from a Gaussian distribution with density

$$p(z) = \mathcal{N}(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\}$$

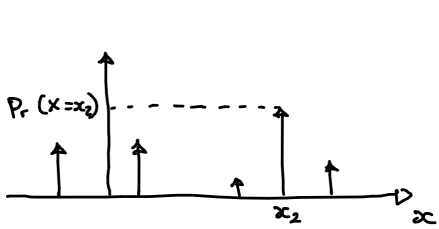
Densities for continuous and discrete random variables

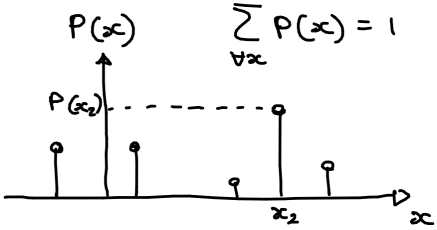
Continuous random variables:



$$\int_{-\infty}^{\infty} p(x).dx = 1$$

Discrete random variables:





Expected values

Continuous:

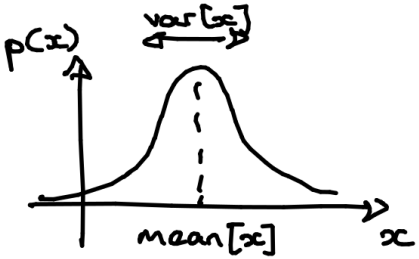
$$\mathbb{E}[g(x)] = \int_{-\infty}^{\infty} g(x)p(x)\,\mathrm{d}x$$

Discrete:

$$\mathbb{E}[g(x)] = \sum_{\forall x} g(x)P(x)$$

Mean and variance:

$$\begin{aligned}\mathrm{mean}\,[x] &= \mathbb{E}[x] \\ \mathrm{var}\,[x] &= \mathbb{E}\left[(x - \mathbb{E}[x])^2\right]\end{aligned}$$



Multiple random variables

Density:

$$\int_{x_1}^{x_2} \int_{y_1}^{y_2} p(x, y) \, \mathrm{d}x \, \mathrm{d}y = \Pr\left(x_1 < x < x_2, y_1 < y < y_2\right)$$

Expectations:

$$\mathbb{E}_{p(x,y)}\left[g(x, y)\right] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) p(x, y) \, \mathrm{d}x \, \mathrm{d}y$$

Conditional probability:

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

Bayes' rule:

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

Marginal probability:

$$p(x) = \int_{-\infty}^{\infty} p(x, y) \, \mathrm{d}y$$
$$P(x) = \sum_{\forall y} P(x, y)$$

When x and y are statistically independent:

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

Reading

- Systems and Signals 344 notes