

probability distribution

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Synonyms

distribution, law

Abstract

A machine learning (ML) method reads in a dataset and delivers an output, such as a hypothesis or the prediction for a given data point. Feeding it a different dataset delivers a different output. Modelling the dataset as realizations of independent and identically distributed (i.i.d.) random variables (RVs) with a common probability distribution turns that output into an RV, whose behaviour is what an analysis of the method describes. The probability distribution decides which data points are typical, and it is never observed; only the dataset is. The probability distribution of a binary RV is fully specified by a single probability, that of a continuous real-valued RV may be specified by a probability density function (pdf), and in the most general case by a probability measure.

Definition

A classifier that flags incoming emails as spam is fitted to one collection of emails, and the emails it will be shown next are a different collection. More generally, consider a machine learning (ML) method that reads in a dataset $\mathcal{D} = \{\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)}\}$ and delivers an output, such as a hypothesis $\hat{h}$ or the prediction it produces for a given data point. Feeding a different dataset into the same method delivers a different output. One way to analyze the method is therefore to study how its output behaves when the dataset consists of realizations of independent and identically distributed (i.i.d.) random variables (RVs) $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)} \sim \mathbb{P}$ with a common probability distribution $\mathbb{P}$ (see Fig. 1).

The probability distribution $\mathbb{P}$ is what such an analysis rests on. It decides which data points are typical, and thereby how the output of the method varies from one dataset to the next. Two numbers attached to the method are then RVs in their own right: the training error it achieves on the dataset it was given, and the risk it incurs on a fresh data point drawn from the same $\mathbb{P}$. Generalization is the statement that these two are close, and it is a statement about $\mathbb{P}$ before it is a statement about the method. Note that $\mathbb{P}$ is never observed; only the dataset is.

How a probability distribution is specified depends on the values the RV takes. The probability distribution of a binary RV $y \in \{0, 1\}$ is fully specified by

the single probability $\mathbb{P}(y = 0)$, since $\mathbb{P}(y = 1) = 1 - \mathbb{P}(y = 0)$; more generally, an RV with finitely many values is specified by a probability mass function (pmf). The probability distribution of a continuous real-valued RV $x \in \mathbb{R}$ might be specified by a probability density function (pdf) $p^{(x)}(\cdot)$, for which $\mathbb{P}(x \in [a, b]) \approx p^{(x)}(a) |b - a|$ holds for a short interval $[a, b]$. In the most general case, the probability distribution of an RV $\mathbf{z}$ defined on a probability space $(\Omega, \Sigma, \mathbb{P}(\cdot))$ is the measure that $\mathbf{z}$ induces on the set of values it takes: it assigns to a measurable set $\mathcal{A}$ of those values the probability $\mathbb{P}(\mathbf{z} \in \mathcal{A})$. A measure obtained this way assigns the value 1 to the set of all values, and is therefore called a probability measure (Billingsley, 1986; Gray, 2009).

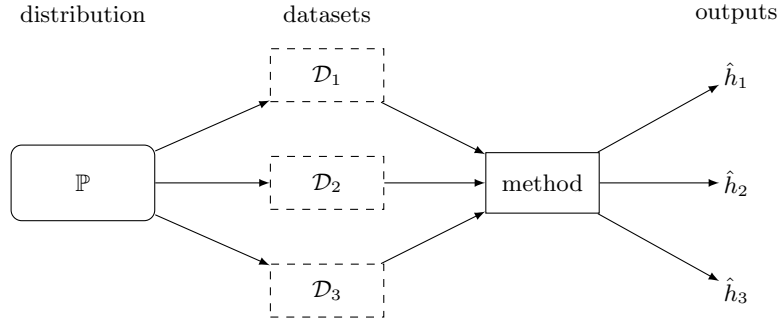


Figure 1: One probability distribution $\mathbb{P}$, three datasets of realizations of i.i.d. RVs drawn from it, and the three outputs one ML method delivers. The spread of the outputs is a property of $\mathbb{P}$

Cross References

i.i.d., realization, RV, probability, pdf, pmf, probability measure, probability space, expectation, generalization

References

1. Billingsley (1986). Probability and Measure. 2nd ed., Wiley, New York, NY, USA.
2. Gray (2009). Probability, Random Processes, and Ergodic Properties. 2nd ed., Springer Science+Business Media, New York, NY, USA.