mean

Authors

Alexander Jung*, Department of Computer Science, Aalto University, Espoo, Finland
*Corresponding author: alex.jung@aalto.fi

Synonyms

expectation, expected value

Abstract

The mean of a random variable (RV) is its expectation, defined as the Lebesgue integral of the RV with respect to its probability distribution. The term also refers to the sample mean, i.e., the average of the data points in a dataset. The two usages are consistent: the sample mean is the mean of the RV obtained by drawing a data point uniformly at random from the dataset. For an RV with a finite second-order moment, the mean is the unique solution of a risk minimization problem under the squared error loss. For the RV associated with a dataset, this optimization problem reduces to empirical risk minimization (ERM) in the simplest regression setting: predicting a numeric label without any features, for which the learned hypothesis is the sample mean of the labels.

Definition

The mean of a random vector $\mathbf{x}$, which takes on values in a Euclidean space $\mathbb{R}^d$, is its expectation $\mathbb{E}\{\mathbf{x}\}$. It is defined as the Lebesgue integral of $\mathbf{x}$ with respect to the underlying probability distribution $\mathbb{P}(\cdot)$ (e.g., see Rudin (1976) or Billingsley (1986)), i.e.,

$$\mathbb{E}\{\mathbf{x}\} = \int_{\mathbb{R}^d} \mathbf{x} \, \mathrm{d}P(\mathbf{x}).$$

The term is also used to refer to the sample mean of a finite dataset $\mathcal{D} = \big\{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(m)} \in \mathbb{R}^d\big\}$. These two usages are consistent. Indeed, it is possible to use a dataset to construct a discrete random variable (discrete RV) $\widetilde{\mathbf{x}}^{(\mathcal{D})} = \mathbf{x}^{(I)}$ on the sample space $\{1, \ldots, m\}$. Here, the index I is chosen uniformly at random, i.e., $\mathbb{P}(I = r) = 1/m$ for all $r = 1, \ldots, m$. The mean of $\widetilde{\mathbf{x}}^{(\mathcal{D})}$ is precisely the average $(1/m)\sum_{r=1}^{m} \mathbf{x}^{(r)}$.

For a random variable (RV) with a finite second-order moment, i.e., $\mathbb{E}\{\|\mathbf{x}\|_2^2\}$ is well defined and finite, the mean is characterized as the unique solution of the following risk minimization problem, whose objective function is a strictly convex function of $\mathbf{c}$ (Bertsekas and Tsitsiklis, 2008):

$$\mathbb{E}\{\mathbf{x}\} = \operatorname*{arg\,min}_{\mathbf{c} \in \mathbb{R}^d} \mathbb{E}\big\{\|\mathbf{x} - \mathbf{c}\|_2^2\big\}.$$

This characterization turns the mean into the unique solution of the simplest machine learning (ML) problem: predicting a numeric label without any features. Consider a training set $\mathcal{D}^{(\text{train})} = \{y^{(1)}, \ldots, y^{(m)}\}$ that consists of numeric labels $y^{(r)} \in \mathbb{R}$ only. A hypothesis is then a single number $h \in \mathbb{R}$ that serves as the prediction for every data point. Empirical risk minimization (ERM) with the squared error loss learns

$$\widehat{h} = \operatorname*{arg\,min}_{h' \in \mathbb{R}} \frac{1}{m} \sum_{r=1}^{m} \left(y^{(r)} - h'\right)^2 = \frac{1}{m} \sum_{r=1}^{m} y^{(r)},$$

i.e., the learned hypothesis is the sample mean of the labels (see Fig. 1). This setting coincides with linear regression for data points that carry the constant scalar feature $x = 1$.

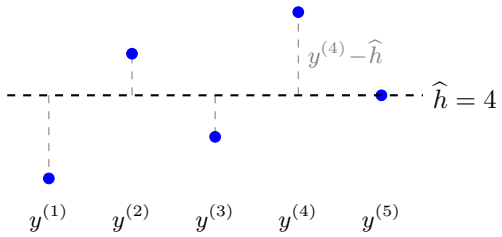

Figure 1: A training set of $m = 5$ numeric labels $y^{(r)}$ (filled circles) and the learned hypothesis $\widehat{h} = 4$ (dashed line), which is the sample mean of the labels. The gray segments are the errors $y^{(r)} - \widehat{h}$; the sample mean minimizes the average of their squares

Cross References

RV, expectation, sample mean, probability distribution, Lebesgue integral, ERM, squared error loss, linear regression

References

1. Bertsekas and Tsitsiklis (2008). Introduction to Probability. 2nd ed., Athena Scientific, Belmont, MA, USA.

2. Billingsley (1986). Probability and Measure. 2nd ed., Wiley, New York, NY, USA.

3. Rudin (1976). Principles of Mathematical Analysis. 3rd ed., McGraw-Hill, New York, NY, USA.