

sample size

Author and editors

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

Editors: Konstantina Oliumtsevs, Ekkehard Schnoor, Aalto University, Espoo, Finland

*Corresponding author: alex.jung@aalto.fi

Abstract

The sample size is the number m of data points in a sample. Many machine learning (ML) methods model the data points as realizations of independent and identically distributed (i.i.d.) random variables (RVs) and judge a hypothesis by its expected loss, the risk; since the probability distribution is unknown, the risk is estimated by the average loss on the sample, for training by empirical risk minimization (ERM) and for validation. How close that average is to the risk depends on the sample size. A concentration inequality such as Hoeffding's inequality bounds the deviation of the average from the risk and turns a required accuracy ϵ and confidence $1 - \delta$ into a sample size of order $\log(1/\delta)/\epsilon^2$ for validation; for training, the sample size must also be compared with the size of the model. For data points recorded in sequence, autocorrelation lowers the effective sample size. The same reasoning sizes the sample for a statistical test, such as whether two features are correlated: the sample size needed to detect a correlation grows as the inverse square of that correlation.

Definition

A frost warning for Krems an der Donau predicts, from the minimum and the maximum temperature of a day, whether the next day brings frost. Whether a hypothesis h for this task is any good is judged by its expected loss, the risk $\bar{L}(h) = \mathbb{E}\{L(\mathbf{z}, h)\}$, where the expectation is over a data point $\mathbf{z}$ drawn from the probability distribution $\mathbb{P}(\mathbf{z})$ of the days. Many machine learning (ML) methods use the independent and identically distributed assumption (i.i.d. assumption): the data points are realizations of independent and identically distributed (i.i.d.) random variables (RVs) $\mathbf{z}^{(1)}, \dots, \mathbf{z}^{(m)} \sim \mathbb{P}(\mathbf{z})$. In practice, $\mathbb{P}(\mathbf{z})$ is unknown, and a method observes only these m data points, a sample. Their number m is the sample size.

With the probability distribution unknown, the risk is estimated by the average loss on the sample,

$$\hat{L}(h|\mathcal{D}) := \frac{1}{m} \sum_{r=1}^m L(\mathbf{z}^{(r)}, h), \quad (1)$$

which is the empirical risk. Empirical risk minimization (ERM) minimizes this average over the hypotheses of a model on a training set, and validation com-

compares the learned hypotheses by the same average on a validation set. Both uses require the average to be close to the risk, and how close it is depends on the sample size: by the law of large numbers, the average (1) converges to $\bar{L}(h)$ as m grows (Shalev-Shwartz and Ben-David, 2014, Sect. 2.2; Papoulis and Pillai, 2002). Fig. 1 shows this for the frost warning: the sample average of the 0/1 loss of one fixed hypothesis, computed on samples of $m = 10$ to 1000 days, scatters around the risk 0.118 with a spread that falls like $1/\sqrt{m}$.

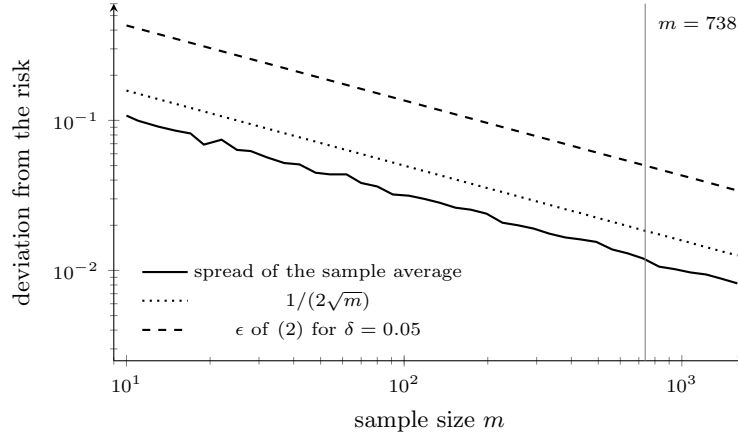


Figure 1: The sample average of the 0/1 loss of a fixed frost-warning hypothesis on m days of 2024 at Krems an der Donau, drawn uniformly with replacement, against the sample size. The solid curve is the standard deviation (spread) of the average over 400 draws, the dotted curve $1/(2\sqrt{m})$, the largest standard deviation a loss in $[0, 1]$ allows; the dashed curve is the deviation ϵ that Hoeffding’s inequality (2) guarantees with probability 0.95, and the vertical line marks the sample size 738 at which that guarantee reaches $\epsilon = 0.05$. Data generated by `pythondemos/samplesize.py`

How large should the sample be? One approach uses a concentration inequality, which bounds the probability that the average deviates from the risk by more than a given amount. For a loss function with values in $[0, 1]$, such as the 0/1 loss, Hoeffding’s inequality gives, for every $\epsilon > 0$,

$$\mathbb{P}\left(\left|\hat{L}(h|\mathcal{D}) - \bar{L}(h)\right| > \epsilon\right) \leq 2 \exp(-2m\epsilon^2), \quad (2)$$

see Hoeffding (1963) and Shalev-Shwartz and Ben-David (2014), Lemma 4.5. The bound holds for a hypothesis chosen before the sample is drawn, which is the situation of validation: the hypothesis was learned on the training set, and the validation set is a fresh sample. Setting the right-hand side to a tolerated failure probability δ and solving for m turns a required accuracy ϵ and confidence $1 - \delta$ into a sample size:

$$m \geq \frac{\log(2/\delta)}{2\epsilon^2}, \quad (3)$$

and this many data points make the validation error lie within ϵ of the risk with probability at least $1 - \delta$ (Shalev-Shwartz and Ben-David, 2014, Thm. 11.1). For the frost warning, $\epsilon = 0.05$ and $\delta = 0.05$ give $m \geq 738$ days; at that sample size all of 2000 drawn samples in Fig. 1 land within 0.05 of the risk. The bound is loose, since it holds for every probability distribution and every loss function in $[0, 1]$; the sample size (3) is therefore enough, but not always needed. Its shape is what carries over: halving ϵ quadruples the sample size, while a smaller δ costs only a logarithm.

For training, the same question has a second answer. The hypothesis is then chosen by looking at the sample, so the bound (2) for a fixed hypothesis no longer applies to the learned one. The sample size then has to be compared with the size of the model. For a model of Vapnik–Chervonenkis dimension (VC dimension) d , a sample size proportional to d/ϵ^2 makes the training error of every hypothesis in the model lie within ϵ of its risk at once (Shalev-Shwartz and Ben-David, 2014, Thm. 6.8). A sample that is small against the effective dimension of the model is the typical cause of overfitting, and the learning curve shows the validation error of the learned hypothesis as a function of the sample size (Hastie et al., 2009, Sect. 7.10).

The sample sizes above assume i.i.d. RVs. Data points recorded in sequence, such as consecutive days, often show autocorrelation: a frost day is more likely after a frost day, and so is a wrong prediction. Dependent data points carry less information than independent ones. For a stationary sequence with autocorrelation ρ_ℓ at lag ℓ , the variance of the average of m consecutive values is that of m independent values multiplied by

$$1 + 2 \sum_{\ell=1}^{m-1} \left(1 - \frac{\ell}{m}\right) \rho_\ell, \quad (4)$$

see Brockwell and Davis (1991), Sect. 7.1. The average is therefore as accurate as that of m divided by (4) independent values, the effective sample size. For the 0/1 loss of the frost warning, the sample autocorrelation of consecutive days is 0.24 at lag 1 and stays positive for a week, and the factor (4) is 4.2 for $m = 30$. The average over 30 consecutive days therefore scatters as much as that over 7 days drawn at random: a month of consecutive days is worth a week of random ones. A sample of consecutive data points must then be larger than (3) or (5) say, by the factor (4), or be drawn at random from the whole record, as the days of Fig. 1 are.

Sample size enters a third task, hypothesis testing: are two features of a data point correlated? For the days of 2024 at Krems, is the minimum temperature of a day correlated with its diurnal range, the maximum minus the minimum? Over the whole year, the Pearson correlation coefficient of the two features is $\rho = 0.16$. A test sees a sample of m days and computes the sample correlation r , a function of sample means. Under the hypothesis that the two features are uncorrelated, r scatters around 0 with a standard deviation close to $1/\sqrt{m}$, whatever their probability distribution. For features that are Gaussian random variables (Gaussian RVs), Fisher’s transformation $\text{artanh}(r)$ is approximately a

Gaussian RV with mean $\text{artanh}(\rho)$ and standard deviation $1/\sqrt{m-3}$ (Fisher, 1921). The test rejects “uncorrelated” when $|\text{artanh}(r)|\sqrt{m-3}$ exceeds the quantile $z_{1-\alpha/2}$ of the standard Gaussian RV, which bounds the probability of a false rejection by the significance level α . A correlation ρ is then detected with probability at least $1 - \beta$, the power of the test, once

$$m \geq 3 + \left(\frac{z_{1-\alpha/2} + z_{1-\beta}}{\text{artanh}(\rho)} \right)^2. \quad (5)$$

For $\rho = 0.16$, $\alpha = 0.05$ ($z_{0.975} = 1.96$) and a power of 0.8 ($z_{0.8} = 0.84$), the sample size (5) is 304 days. Fig. 2 shows the power of the test against the sample size: 0.84 at $m = 304$, and 0.10 for a sample of 30 days. The shape is that of (3) again: halving the correlation to be detected quadruples the sample size, so a weak correlation needs a large sample before “uncorrelated” can be rejected.

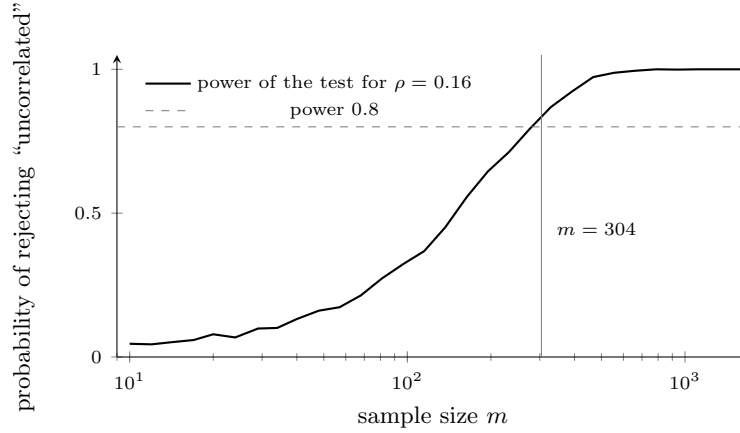


Figure 2: The power of the correlation test at significance level $\alpha = 0.05$ for the minimum temperature and the diurnal range of a day at Krems an der Donau, whose Pearson correlation coefficient over 2024 is $\rho = 0.16$, against the sample size (solid curve, the fraction of 1000 samples of m days on which the test rejects “uncorrelated”). The dashed line marks the power 0.8, the vertical line the sample size 304 of (5). Data generated by `pythondemos/samplesize.py`

Cross References

sample, data point, dataset, i.i.d. assumption, risk, empirical risk, law of large numbers, concentration inequality, Hoeffding’s inequality, validation, validation error, VC dimension, learning curve, overfitting, autocorrelation, hypothesis testing, Pearson correlation coefficient

References

1. Brockwell and Davis (1991). Time Series: Theory and Methods. 2nd ed., Springer-Verlag, New York, NY, USA.
2. Fisher (1921). On the “probable error” of a coefficient of correlation deduced from a small sample. *Metron*, 1(4), pp. 3–32.
3. Hoeffding (1963). Probability Inequalities for Sums of Bounded Random Variables. *J. Amer. Statistical Assoc.*, 58(301), pp. 13–30.
4. Shalev-Shwartz and Ben-David (2014). Understanding Machine Learning: From Theory to Algorithms. Cambridge Univ. Press, New York, NY, USA.
5. Hastie et al. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd ed., Springer Science+Business Media, New York, NY, USA.
6. Papoulis and Pillai (2002). Probability, Random Variables, and Stochastic Processes. 4th ed., McGraw-Hill Higher Education, New York, NY, USA.