25th International Congress on Modelling and Simulation, Darwin, NT, Australia, 9 to 14 July 2023
mssanz.org.au/modsim2023

Selecting validation sets based on information entropy

Dang Nguyen and Damien Watkins

Data61 CSIRO, Research Way, Clayton, VIC 3168, Australia

Email: dang-quan.nguyen@data61.csiro.au

Abstract: Considering a dataset that is used for a machine learning model, or more generally one that is used as input to an algorithm (such as a scenario data as input to a simulator). This dataset is usually split into a training set to tune the model and a test set to validate the model after it has been trained. Furthermore, part of the training set can be set aside as a validation set to avoid the bias that arises when a specific model is selected based on its high performance against one particular training set, but it then performs suboptimally with a more general input. Likewise, the validation set can also be divided into smaller sets and each model is evaluated on all small validation sets, separately, after being trained on the rest of the data. Then the performance of a model is obtained by averaging out its evaluations over the small validation sets. This cross-validation may yield a more accurate measure of a model performance, but then the training time is now multiplied by the number of validation sets.

We argue that this time-consuming validation process of a model lacks a measure to help guide the selection of a smaller subset of the validation sets that the process needs to focus on. We investigate the use of information entropy as a measure for this purpose.

We follow the notations in Géron (2019). Given a dataset $\mathbf{X}$ containing all feature vectors $\mathbf{x}^{(i)} = (x_{i1}, \ldots, x_{in})$. Each feature value x_{ij} appears with a probability p_{ij} so that $\mathbf{x}^{(i)}$ has a probability $p_i = \prod_{k=1}^{n} p_{ik}$. Let H be the entropy characterising the information contained in $\mathbf{X}$, the following theorem from Khinchin (1957) says we can approximate this dataset by a subset that conserves H up to some arbitrarily small $\eta > 0$:

Theorem 1. *Given $\varepsilon > 0$ and $\eta > 0$, arbitrarily small, for sufficiently large dataset $\mathbf{X}$ all vectors $\mathbf{x}^{(i)}$ can be devided into two groups with the following properties: 1) the probability p_i of any vector of the first group satisfies the inequality*

$$\left| \frac{\log \frac{1}{p_i}}{n} - H \right| < \eta \tag{1}$$

and 2) the sum of the probabilities of all vectors of the second group is less than ε.

Assuming the feature vectors $\mathbf{x}^{(i)}$ have large dimension n, we can rearrange them in order of decreasing probability p_i. We then select the vectors in this order until the sum of the probabilities of the selected vectors exceeds a chosen value $0 < \lambda < 1$. Suppose we have selected $N_n(\lambda)$ such vectors. The following result from Khinchin (1957) shows that this group can approximate the initial dataset:

Theorem 2. $\lim_{n\to\infty} \frac{\log N_n(\lambda)}{n} = H.$

This first group of selected vectors is used as our validation set, as it captures most of the information in our dataset. The remaining vectors (those in the second group) only add little information to our model.

References

Géron, A. (2019), *Hands-on Machine Learning with Scikit-Learn, Keras & TensorFlow*, O'Reilly.
Khinchin, A. I. (1957), *Mathematical Foundations of Information Theory*, Dover Publications.

Keywords: *Dataset selection, validation, information entropy, optimisation*