Machine Learning Experimental Design Theory Cheat Sheet

Basics

Intelligence: The ability to adapt [1].

Machine Learner: A machine learner is a mechanism that is able to adapt to a target function in two ways: 1) through training and 2) through generalization. A mechanism that is trainable without generalization is called *memory*. A mechanism that generalizes without training is a *quantizer*.

Artificial Intelligence: A system that performs human-like tasks, often by relying on machine learning.

Classification: Quantization of an input into a set of predefined classes (see also: generalization).

Regression: Classification with a large (theoretically infinite) number of classes described in functional form.

Detection: Binary classification.

Clustering: Quantization.

Memory View of Machine Learning

Tabularization: Data is organized in N rows and D columns. Training data contains a $D + 1$th column with the prediction variable to be learned (*labels*). All *inputs* to the machine learner are *points* with D coordinates. All data in digital computers can be tabularized this way.

Target Function: A mapping from all input points to the labels, assumed to be a mathematical function.

Representation Function: The parameterized function a machine learner uses either standalone or in composition to adapt to the target function, e.g. the *activation function* of an artificial neuron.

Bit (Binary digit): Unit of measurement for memory capacity [5]. One bit corresponds to one boolean parameter of the identity function whereby each parameter can be in one of the two states with equal chance.

Intellectual Capacity: The number of unique target functions a machine learner is able to represent (as a function of the number of model parameters) [4].

Memory-equivalent Capacity: With the identity function as representation function, N bits of memory are able to adapt to 2^N target functions. A machine learner's intellectual capacity is memory-equivalent to N bits when the machine learner is able to represent all 2^N binary labeling functions of N uniformly random inputs.

Memory Equivalent Capacity: VC Dimension for uniformly random data points.

Generalization

Generalization ($\forall$): The concept of handling different objects by a common property. The set of objects of the common property are called a *class*, the elements are called *instances*. Neuron: all instances of input signals below an energy threshold are ignored.

Generalization in Machine Learning: All inputs *close enough* to each other result in the same output. This is for an assumed or trained definition of *close enough* (*generalization distance*). Memory: Only identical is *close enough* $\implies$ no generalization.

Adversarial Example: An input that contradicts the generalization assumption of a machine learner [3].

Overfitting: A machine learner at memory-equivalent capacity or higher with regards to the number of inputs in the training data could as well just use the identity function as representation function and still adapt to the training data perfectly. The machine learner is said to *overfit*.

Accuracy: A necessary but not a sufficient condition for generalization success. At the same accuracy, overfitting potentially maximizes the number of adversarial examples, capacity reduction minimizes it.

Measuring Generalization: Generalization of a class-balanced binary classifier is $G = \frac{\#correctly\ predicted\ instances}{Memory\ Equivalent\ Capacity}$, only $G > 1$ implies successful generalization.

Training Processes

Training for Accuracy: The process of adjusting the parameters of the representation function(s) of the machine learner to approximate a target function with maximum accuracy.

Training for Generalization: The process of successively reducing the capacity of a machine learner while training for accuracy. The model with the highest accuracy and the smallest capacity is the one that uses the representation function(s) most effectively. Therefore, it has the lowest chance of failing (this is requiring to increase capacity) when applied to unseen data from the same experimental setup [2].

Regularization: Reducing capacity during training by restricting the freedom of the parameters, thereby potentially improving generalization. Techniques include *drop out*, *early stopping*, *data augmentation*, or imperfect training.

Capacity Estimation

Capacity Requirement: Build a static-parameter machine learner to memorize the training data. Assume exponential improvement through training. This is, the memory-equivalent capacity can be minimally logarithmic of the size of the static-parameter machine learner [2].

Memory-equivalent Capacity for Neural Networks [2][7]:

1. The output of a single neuron is maximally one bit.
2. The memory capacity of a single neuron is the number of parameters in bits.
3. The memory capacity of neurons is additive.
4. The memory capacity of neurons in a subsequent layer is limited by the output of the layer it depends on.

Generalization Estimation

Generalization Progression: Estimate the capacity needed to memorize 10%, 20%, ..., 100% of training table. If the capacity does not stabilize at the higher percentages either there is not enough training data to generalize or the representation function(s) of the machine learner is/are not right for the task.

Find Best Machine Learner for Data: Measure/estimate Generalization Progression for different types of machine learners. Pick the one with convergence to the smallest capacity.

Testing Generalization Performance: Measuring accuracy against an independent data set after training is the only way to guarantee generalization performance. Testing against hold-out or cross validation data in training, practically makes the data part of the training set.

Occam's Razor (modern): When you have two competing theories that make exactly the same predictions, the simpler one is the better [6].

Occam's Razor for Machine Learning: Among equally accurate models, choose the one that requires the lowest memory-equivalent capacity (see also: Training for Generalization).

Created by G. Friedland, v1.01 Aug 28th, 2020
fractor@eecs.berkeley.edu

References

[1] A. Binet and T. Simon. Méthodes nouvelles pour le diagnostic du niveau intellectuel des anormaux. *L'année Psychologique*, 11(1):191–244, 1904.

[2] G. Friedland, A. Metere, and M. Krell. A Practical Approach to Sizing Neural Networks. *arXiv preprint arXiv:1810.02328*, October 2018 `https://arxiv.org/abs/1810.02328`.

[3] L. Huang, A. D. Joseph, B. Nelson, B. I. Rubinstein, and J. Tygar. Adversarial machine learning. In *Proceedings of the 4th ACM workshop on Security and Artificial Intelligence*, pages 43–58. ACM, 2011.

[4] D. J. C. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, New York, NY, USA, 2003.

[5] C. E. Shannon. The Bell System Technical Journal. *A mathematical theory of communication*, 27:379–423, 1948.

[6] W. M. Thorburn. Occam's razor. 1915.

[7] `http://tfmeter.icsi.berkeley.edu`