**The few things to know about Machine Learning**

**-Domingos P**

Machine learning systems automatically learn programs from data. Machine learning is used in Web search, spam filters, recommender systems, ad placement, credit scoring, fraud detection, stock trading, drug design, and many other applications.

The focus is on the most mature and widely used machine learnings: classification. A classifier is a system that inputs (typically) a vector of discrete and/or continuous feature values and outputs a single discrete value, the class. A learner inputs a training set of examples, and outputs a classifier.

* **LEARNING = REPRESENTATION + EVALUATION + OPTIMIZATION**

Learning algorithms consists of combinations of just three components:

✦ Representation: choosing the set of classifiers that it can possibly learn. This set is called the hypothesis space of the learner. If a classifier is not in the hypothesis space, it cannot be learned

✦ Evaluation: An evaluation function/objective function/scoring function is needed to distinguish good classifiers from bad ones. The evaluation function used internally by the algorithm may differ from the external one that we want the classifier to optimize.

✦ Optimization: needing a method to search among the classifiers in the language for the highest-scoring one. The choice of optimization technique is key to the efficiency of the learner.

* **IT’S GENERALIZATION THAT COUNTS**

The fundamental goal of machine learning is to generalize beyond the examples in the training set. The most common mistake among machine learning beginners is to test on the training data and have the illusion of success. Cross-validation: randomly dividing your training data into (say) ten subsets, holding out each one while training on the rest, testing each learned classifier on the examples it did not see, and averaging the results.

* **DATA ALONE IS NOT ENOUGH**

Every learner must embody some knowledge or assumptions beyond the data it’s given. Very general assumptions—like smoothness, similar examples having similar classes, limited dependences, or limited complexity—are often enough to do very well, and this is a large part of why machine learning has been so successful. One of the key criteria for choosing a representation is which kinds of knowledge are easily expressed in it: if we have a lot of knowledge about what makes examples similar in our domain, instance- based methods may be a good choice. If we have knowledge about probabilistic dependencies, graphical models are a good fit. And if we have knowledge about what kinds of preconditions are required by each class, “IF . . . THEN . . .” rules may be the best option.

* **OVERFITTING HAS MANY FACES**

Overfitting is the bugbear of machine learning. When my learner outputs a classifier that is 100% accurate on the training data but only 50% accurate on test data, when in fact it could have output one that is 75% accurate on both, it has overfit. Way to understand overfitting: **Bias-Variance Decomposition**

* + Bias: A learner’s tendency to consistently learn the same wrong things.
  + Variance: The tendency to learn random things irrespective of the real signal.

A linear learner has high bias: because when the frontier between two classes is not a hyperplane the learner is unable to induce it. Decision trees don’t have this problem because they can represent any Boolean variance. But they can suffer from high variance! Decision trees learned on different training sets generated by the same phenomenon are often very different, when in fact they should be the same. Methods to combat overfitting: Cross-validation,Adding a regularization term to the evaluation function, statistical significance test like chi-square. But severe overfitting can occur even in the absence of noise.

* **INTUITION FAILS IN HIGH DIMENSIONS**

Curse of dimensionality: many algorithms that work fine in low dimensions become intractable when the input is high-dimensional. Similarity-based reasoning that machine learning algorithms depend on, breaks down in high dimensions. There is an effect that partly counteracts the curse, which might be called the “blessing of non-uniformity.” In some applications examples are not spread uniformly throughout the instance space, but are concentrated on or near a lower-dimensional manifold. k-nearest neighbour works quite well for handwritten digit recognition even though images of digits have one dimension per pixel, because the space of digit images is much smaller than the space of all possible images.

* **Theoretical guarantees are not what they seem**

The main role of theoretical guarantees in machine learning is not as a criterion for practical decisions, but as a source of understanding and driving force for algorithm design. The most common type of theoretical guarantees is a bound on the number of examples needed to ensure good generalization. Another common type is asymptotic: given infinite data, the learner is guaranteed to output the correct classifier. Note that because of the bias-variance trade-off, if learner A is better than learner B given infinite data, B is often better than A given finite data.

* **Featuring Engineering is a key**

some machine learning projects succeed and some fail. What makes the difference? the most important factor is the features used. Often, the raw data is not in a form that is amenable to learning, but you can construct features from it. Machine learning is not a one-shot process of building a data set and running a learner, but rather an iterative process of running the learner, analysing the results, modifying the data and/or the learner, and repeating.

* **More Date beats a cleaver algorithm**

Suppose you’ve constructed the best set of features you can, but the classifiers you’re getting are still not accurate enough. There are two main choices:

✦ design a better learning algorithm

✦ gather more data

As a rule of thumb, a dumb algorithm with lots and lots of data beats a clever one with modest amounts of it. Two main limited resources are time and memory. Enormous mountains of data are available, but there is not enough time to process it, so it goes unused. This leads to a paradox: even though in principle more data means that more complex classifiers can be learned, in practice simpler classifiers used, because complex ones take too long to learn. As a rule, it pays to try the simplest learners first.

* **Learn Many Models, Not Just One**

In the early days, most effort went into trying many variations of it and selecting the best one. Nowadays, still trying many variations of many learners and selecting just the best one. But researchers combine many variations instead of selecting the best one. Creating such **model ensembles** is now standard: Bagging, Boosting, Stacking. Model ensembles should not be confused with Bayesian model averaging(BMA).

* **SIMPLICITY DOESN’T IMPLY ACCURACY**

Occam’s razor can be applied in machine learning, but there are many counterexamples to it, and the “no free lunch” theorems imply it cannot be true. Counterexamples: Ensembles learning. A more sophisticated view: smaller hypothesis spaces allow hypotheses to be represented by shorter codes. A further complication arises from the fact that few learners search their hypothesis space exhaustively. Simpler hypotheses should be preferred because simplicity is a virtue in its own right, not because of a hypothetical connection with accuracy.

* **REPRESENTABLE DOES NOT IMPLY LEARNABLE**

Every function can be represented, or approximated arbitrarily closely, using this representation. However, just because a function can be represented does not mean it can be learned. Given finite data, time and memory, standard learners can learn only a tiny subset of all possible functions, and these subsets are different for learners with different representations. Therefore the key question is not “Can it be represented?”, to which the answer is often trivial, but “Can it be learned?

## Correlation Does Not Imply Causation

The point that correlation does not imply causation is made so often that it is perhaps not worth belabouring. Machine learning is usually applied to observational data, where the predictive variables are not under the control of the learner, as opposed to experimental data, where they are. Many researchers believe that causality is only a convenient fiction.

* **CONCLUSION**

Like any discipline, machine learning has a lot of "folk wisdom" that can be difficult to come by, but is crucial for success. Advance machine learning is needed.