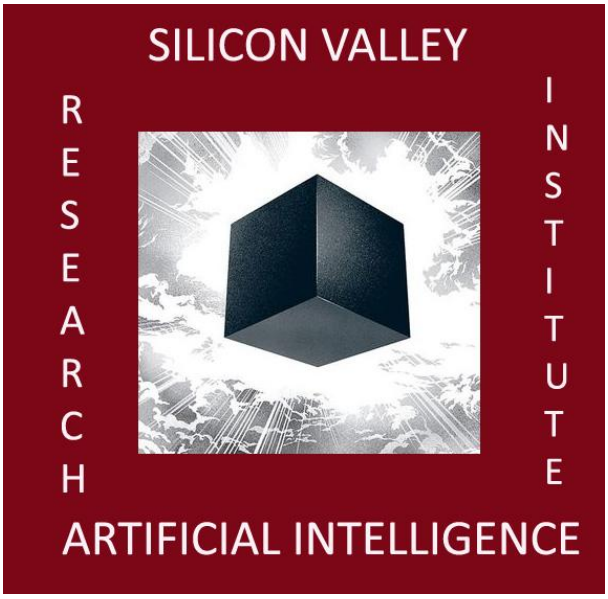

Machine Learning in Statistics

Deep learning is not the only game in town. In fact, many nonlinear algorithms have been introduced by the field of machine learning, notably during the 1990s.

The most successful method of machine learning was for a long time a method of statistical learning, the linear classifier called "support vector machine" (SVM). The original SVM algorithm was invented by the Soviet mathematician Vladimir Vapnik and Alexey Chervonenkis in 1963 (they originally called it "generalized portrait"), and improved by Tomaso Poggio at MIT (in 1975 he introduced the "polynomial kernel"), but lay dormant until in 1991 Isabelle Guyon at Bell Labs (where Vapnik had moved in 1990) adapted SVMs to pattern classification ("A Training Algorithm for Optimal Margin Classifiers", 1992) using the optimization algorithm called "minover" invented by the physicists Marc Mezard and Werner Krauth in France to improve Hopfield-style neural networks ("Learning Algorithms with Optimal Stability in Neural Networks", 1987). Guyon turned a linear algorithm into a nonlinear algorithm. Another European-born Vapnik collaborator at Bell Labs, Corinna Cortes, further improved an SVM into a "soft-margin classifier" ("Support-Vector Networks", 1995). SVM learning was, for example, used by Thorsten Joachims for classifying texts into categories ("Text Categorization with Support Vector Machines", 1998). A theorem proven by Thomas Cover at Stanford ("Geometrical and Statistical Properties of Systems of Linear Inequalities with Applications in Pattern Recognition", 1965) states that a complex pattern-classification problem (such as recognizing an image) is more likely to be linearly separable (i.e. to succeed) when cast in a high-dimensional space than in a low-dimensional space (e.g. in many dimensions than in two dimensions). A kernel is a function that transforms a two-dimensional data space into a very high dimensional space (the "feature space"). This function requires a lot of computation, but mathematicians discovered a "trick" to make it feasible (which is still called "kernel trick"). Most problems of categorization are nonlinear, a fact that makes them very hard to solve. The kernel

trick allows to transform a nonlinear problem into a linear problem, after which there are plenty of well-known linear classifiers that one can use on the feature space. This trick harkens back to a theorem proven in 1909 by the British mathematician James Mercer and adapted in 1964 by the Soviet mathematician Mark Aizerman. That's the trick that Guyon used to make Vapnik's original concept the most popular method of statistical learning.

Tin-kam Ho at Bell Labs improved decision tree analysis with the stochastic discrimination method developed in 1990 by Eugene Kleinberg (her advisor at State University of New York) and obtained "random decision forests" ("Random Decision Forests", 1995), perfected by Leo Breiman at UC Berkeley: a "forest" is a large set of decision trees.

Classification and Regression Trees or CART for short is a term introduced by Leo Breiman to refer to decision-tree algorithms that can used for classification or regression predictive modeling problems.

SVM probably remains the most used method in machine learning even in the age of deep learning. The reason is computational: you can distribute easily an SVM algorithm over dozens of machines so it will learn quickly, whereas a deep learning algorithm (which in theory is more efficient) is difficult to distribute over dozens of machines. That's because SVMs are linear models whereas neural networks are nonlinear. As of 2017, Google is de facto the only organization that has been able to perform distributed computation of deep learning algorithms. This has more to do with the big-data infrastructure (such as MapReduce, that Google published in 2005) than with the "intelligence" of the system. If you don't distribute neural networks over multiple machines, the training may take weeks or months.

The 1990s also witnessed the first applications of Bayesian inference networks: the Quick Medical Reference - Decision Theoretic or QMR-DT project (1991), by Gregory Cooper's team at Stanford, and NASA's Vista project (1992), led by Matthew Barry and Eric Horvitz.

The training of a machine-learning algorithm works best when the numbers of positive and negative instances are roughly equal. Alas, this is not the case in the real world. For example, the dataset of medical images is almost always imbalanced: healthy people rarely get a radiological scan of their chest and therefore most chest CT scans are of patients with known diseases. If the training dataset consists of 10,000 medical images of people with a heart condition and of 10 healthy people, the classifier will tend to classify healthy people as at risk of a heart attack. And viceversa: if the training dataset consists of 10,000 medical images of healthy people and of 10 people with heart conditions, the classifier will tend to classify people at risk as perfectly healthy. This problem is extremely common in many fields. This is an old problem in machine learning systems (or simply classifiers). The common remedy is to employ an "ensemble method", i.e. to combine multiple (weak) learning algorithms into a (strong) learning algorithm. A number of techniques (or meta-algorithms) have been proposed since at least the 1990s: BAgging (Bootstrap Aggregating), proposed by Leo Breiman at UC Berkeley in 1994; SMOTE (Synthetic Minority Over-Sampling Technique), developed by Nitesh Chawla and others at the University of South Florida in 2000; etc.

"Boosting" is a class of methods for improving the accuracy of both linear and nonlinear classifiers. Robert Schapire at Bell Labs proved their theoretical feasibility ("The Strength of Weak Learnability", 1990) and AdaBoost (Adaptive Boosting), developed in 1995 in collaboration with Yoav Freund, was the most popular incarnation. A booster is not actually a classifier itself, just a "booster": it turns a linear combination of weak learners into a strong learner. The weakest learner is random guessing; the next weakest learners are learners that are only slightly better than random guessing; and so forth; a strong learner is a very accurate learner. Then came: DataBoost by Hongyu Guo and Herna Viktor at the University of Ottawa in 2004; BEV (Bagging Ensemble Variation) by Cen Li at Middle Tennessee State University in 2007; RUSBoost (Random Under-Sampling Boost) by Taghi Khoshgoftaar and others at Florida Atlantic University in 2008; etc.

Michael Jordan at MIT (a former student of David Rumelhart at UC San Diego) worked on graphical models (see his book "Learning in Graphical Models", 1998); Bernhard Schoelkopf in Germany specialized in "kernels" (see his book "Learning with Kernels", 2002); Carl-Edward Rasmussen in Britain used "Gaussian processes" (see his book "Gaussian Processes for Machine Learning", 2006).

Furthermore, despite the hype around deep learning, algorithms for real-time computer vision such as David Lowe's SIFT, OpenCV (Open Source Computer Vision), the library of computer-vision functions released in 1999 by Intel, HOG (Histograms of Oriented Gradients), published in 2005 by Navneet Dalal and Bill Triggs at INRIA in Paris (the National Institute for Research in Informatics and Automation), SURF (Speeded Up Robust Features), published in 2006 by Luc Van Gool's team at ETH Zurich (the Swiss Federal Institute of Technology), and ORB, released in 2011 by Gary Bradski's team at Silicon Valley startup Willow Garage, continued to prevail over convolutional neural networks. These algorithms had several advantages over neural networks: they are easier to implement, don't require as much processing power, and can be trained with a smaller set. Therefore they are generally preferred in real-time applications. They are, however, very different from how the human brain works.

In 2001 Paul Viola and Michael Jones at the Mitsubishi research laboratories in Boston developed a face detector that achieved fast processing via cascading classifiers ("Rapid Object Detection Using a Boosted Cascade of Simple Features", 2001).

The similarities between language parsing in natural language processing and scene analysis in machine vision had been known for decades. Gabriella Csurka at the Xerox Research Centre in France developed the equivalent of the "bag-of-words" technique for machine vision: the "bag-of-visual words" or "bag-of-features" technique, which improved Perona's "constellation of parts" method for object detection and would remain the most popular technique for image classification for at least a decade ("Visual Categorization with Bags of Keypoints", 2004). For object detection, instead, the most popular method became "deformable part model" (DPM), developed by Pedro Felzenszwalb and David McAllester at the University of Chicago, that won the PASCAL VOC competition in 2007 ("Discriminatively Trained, Multiscale, Deformable Part Model", 2008).