Machine Learning: Review

Contents

Overview

In this section we review the essential notation and conceptual foundations of machine learning.

A grasp of notation is needed to fully engage with both the instructional (, educational) literature on ML; library documentation; and the broader community ecosystem.

Data and Variables

In a tabular view, a column represents a variable (, covariate, random variable, feature, target). The rows of that column are its observed values, each row corresponding to a particular example or data point.

The observation index, i, is often given as a superscript or left off (ie., so we read x as an x rather than the whole dataset x). So that the subscript refers to different variables, ie., x1, x2, .. are different features of a single observation x.

The entire data set is often split into test/training subsets. The training subset is used to solve the learning problem, and the test set to evaluate the quality (, performance) of that solution.

Models, Linear Models

In general, we aim to find a model (estimate function) $\hat{f}$ of a true function f.

$f = \arg\max_f \mathbb{E}_{unseen}\, l(f(x), y)$

Estimate:

$\hat{f} = \arg\max_f \frac{1}{N} \Sigma_{training}\, l(f(x), y)$

Eg., in regression the estimate $\hat{f}$ accepts features of a observation x and provides a real number estimate for y, ie., $\hat{y}$.

Types of Learning

In supervised learning, $\hat{f}(x)$ is found by varying the model or its parameters until its estimates for y are the best according to some criterion formalized by the loss.

To determine whether an estimate is good, we require known target values y for each known observation x.

In unsupervised learning we tend to be estimating $\hat{f} = p(x_1, x_2, ...x_n)$, ie., the function to be learned is the joint probability of the dataset (roughly: how likely it is to observe x_1 *and* x_2 occuring).

A distinction is sometimes made between classification: $\hat{f} = ...p(x|y)...$, and prediction: $\hat{f} = ...p(y|x)....$ Where in the former case the probability of observing a feature x, given a known label y is the heart of the decision about what x is. In the latter case, the estimate for y is derived from how likely it is given x.

Classification is concerned with the past in the sense that the estimating function is trained taking the target y as a given (we known the complete domain of y and the distribution of y in the dataset biases the model). Prediction is concerned with the future in the sense that it is trained not assuming y (the domain is open; the distribution of y in training tends not to bias). In classification, we generalize from the known labels to the unknown features. In prediction, we generalize from known features to unknown targets.

Regression

Regression is a supervised learning problem whose target is a real number. Eg., estimating the price of a car; the age of a student; the profit from an ad campaign; the probability of reoffence for a given crime.

Classification

Classification is a supervised learning problem. The target, y, is a discrete number representing one of several labels. A binary classification problem usually has the target domain $-1, +1$ as the "negative" and "positive" outcomes. Binary classification is a model of many experiments and decision making processes (null hypothesis vs expected; yes vs no; pass vs play; etc.).

In general classification is the labelling of some observation x with an estimated label $\hat{y}$ via training a model on seen pairs (x, y).

Classification as Regression

A regression model is often the basis of solving a classification problem. In this case the real-number output of the model is interpreted as a score and mapped to a class.

This map is called a decision function, eg.,

decsiisionsdlfnlkjasfd

Loss

The loss provides a measure of the quality of a given prediction. The total loss (risk, cost, etc.) over the entire training data set therefore provides the quality of the model, ie., how well it performs on all data.

If a model has parameter w, minimizing the total loss with respect to w, finds the best model, f(x).

Ie., if we observe the total loss as we vary w, the value of w for the smallest observed total loss, provides the best parameter value. The model fixed at this value is then the best quality estimator for y (according to our choice of loss).

That is, we find $\hat{f}(x; w)$ st.

Distance

The quality of an estimate for y can often be specified as a kind of distance: how far the estimate differs from y.

distance diagram

Regularization

The best parameters to a model are not necessarily those that provide the least distance between its estimates and the observed training data. Typically, a model that accurately captures the training data, equally, fails to generalize beyond these observations.

A "best" parameter value therefore may often be found by adjusting the raw error score, so that the model captures less of the detail peculiar to the training data.

$loss_i = error_i + regularization_i$

When visualized, these regularized models have more regular structure. Unregularized models tend to be noisey and uneven, being overfit to observation.

viz boundary

Optimization

The process of finding the optimal $\hat{f}$ is called **optimization**. There are many algorithms for finding a minimum loss. Naively searching the infinite space of possible values of w will yield a global minimum, however the infinite search isn't, in principle, possible. Therefore various algorithms will adopt search strategies that will, in general, yield different minima: points where w appears to take its best value (ie., least loss) but is not the global minimum result.

Most optimization algorithms work on a principle of gradient descent: starting at a random guess for w, update w, in the direction of locally-decreasing loss.

Since we need an infinite scan to know the true direction of least-loss, a local scan provides a sub-optimal useful strategy.

Gradients

A gradient is a generalization of the notion of slope or steepness. In the linear 1D case, the gradient is the slope of the line. In a multi-dimensional case, the gradient is an arrow pointing in the direction of greatest ascent, ie., the gradient is a list of multiple numbers (a vector) whose entries describe the rate of ascent in each possible direction of motion.

There are multiple notations for the gradient.

Learning Rate

A gradient descent algorithm will fail to find a minimum if w is always updated "too much" or "too little", that is, if the value of the least loss occurs at a value for w which is never tried.

Since we cannot try an infinite number of values, a tradeoff is needed. Rather than update w by the total value of the gradient at a trial point, a percentage of the gradient is taken. Suppose this learning rate is 0.01, ie., 1%, then a gradual sweep is made of the space; moving only a little ways up or down.

Even 1% may be too small a step, or indeed, too large. Varying the learning rate during training may therefore find better loss minima.

Explanatory vs. Statistical Approaches

Explanatory Models

Explanatory models are relations of causal variables, parameterised by the strength of effect of each variable.

An explanatory model, eg. $dial \xrightarrow{causes} temperature$

May predict an observed association:

$temperature(dial) = 16 + 3dial$

A causal variable is one we may act upon and which, post-interaction, produces a change in a effect variable,

(ie., f(x) – f(do(x = 3)) -> ...)

Explanatory models describe all possible states of the world (, environment) under change of relevant variable: given complete (cause1, cause2, ... effect) variables described by the model, it should not be possible to observe a value for the effect untracked by the model.

Insofar as explanatory models fail they do so because they are incomplete, or take the wrong form (eg., temperature is not merely caused by thermostat activity, but also, window-breeze; temperature is not linearly related to thermostat position).

Explanatory models are gaurenteed to generalize from the training set (in-sample) to the unseen set (out-sample), in the sense that, if the explanatory model is complete and applicable, the same causal laws work in the both domains.

Statistical-Associative Models

Associative statistical modelling aims to find any function of any variable and parameters that tracks the variation in the target y as seen the training set (x, y). Regardless of whether acting upon x, causes, y to change.

Eg., suppose in a room all the people in coats ($x = 1$) sat closer to the door, and all people in t-shirts ($x = 2$) further away, then

$temperature(x) = 16 + 5x$

So that the temperature near the t-shirt area is $16 + 10 = 26$; and near the coats $16 + 5 = 21$.

This may perfectly predict the distribution of temperature, but taking off a coat does not cause the temperature to increase! Acting upon the variable x is not the same as observing a different value for it. An action produces discontinuous changes of association, ie., all data on the current association of a dial to a temperature lacks any predictive power when the dial is changed.

I can obtain an infinite amount of data on how one angle of a dial maps to an infinite amount of variables of interest (eg., molecule velocity). An associative model trained on this infinity is still unpredictive when the dial angle is changed!

The causal laws of a system produce abitarily-many such associations (every dial angle corresponds to a new total state of the room) – it is therefore impossible to "reverse" a single associative model to produced a causal one.

(Science finds causal models by generating many associative ones under action upon candidate causal variables, taking the pattern in the patterns, as causal variables change as a guide for theory building. Candidate explanatory models need to reporduce every such pattern and in addition predict novel ones to be found plausible).

Consider also a image classifier,

classify(image) -> cat

Supppose this classifer tracks the presence of whiskers. Clipping whiskers on cat does not turn it into a dog. The parameters of the model conspire to increase the response to whisker-like patterns in images: this produces predictive acuracy with respect to a kind of pictoral dataset, but is still merely predictive.

Mere Predictive Models and The Semantic Gap

A scientific theory comprises many explanatory models and explicit conceptual-variable connections, saying that, eg., our intuitive notion of "falling" is gravity and that gravity is tracked by a force term, etc.

A goal of scientific enquirey is therefore to provide explanatory models and causal semantics for these models: a meaning for each of their terms which map them to, in principle, pieces of the environment that we may act upon.

A perfectly predictive statistical model fails in many more ways that a complete explanatory model. A merely predictive model is in a deep way essentially coincidental: it happened to be that coats/tshirts were so-laid-out. It happened to be that all images of cats had whiskers. These aren't inexplicable, but they cannot be relied upon to hold in general.

We should therefore regard associative models as unsafe and unreliable without expert supervision, or additional systems to verify their applicability.

An associative model is likely to fail in wholey unpredictable ways: because what we rely on for predictive intuitions is explantory models. When we misclassify an obstacle, we may crash. But we can explain in what circumstances we are likely to do so, because our models of obstacles are explanatory.

Mere associative models are not necessarily parameterised by interpretable terms, and are therefore likely to misclassify, eg., road obstacles, in inexplicable scenarios.

More precisely we rely on *theories*, that is, a wide class of explanatory models. When navigating a road, we model the behaviour of other drives via empathetic and cognitive systems that assign likely behaviour, thoughts, goals, intentions, etc. These estimates are constantly updated by, eg., a indicator light (or the failure to indicate). When navigating social environments, such as roads, we employ a "Theory of Mind".

This leaves merely associative models worse than inexplicable, but essentially kinds of illusions. Even when they work we misattribute the mechanism of how they work. When we ask alexa to turn on the lights we assume there is some causal-semantic notion of "lights", "on", etc. These sets us up to confidently predict the successful operation of an associative system when in fact it will fail catestrophically.

Training and The Semantic Gap

Training associative systems to "overcome" semantic gaps quickly runs into infinites. Suppose I wish to train a classifier on 2D images, even with an infinite number, I can make a dog "look like" a cat in 2D – using, eg., lighting conditions, occlusion, fur fluffing, hair dressing, coat dying, etc.

Ie., suppose I choose a 2D image labelled "Cat" in the infinite set – I can at least photograph some dog to produce an identitical 2D image!

Suppose then I add in infinities of 3D info, of skeleton structure, etc. Still the model only tracks coincidences-we-hope-generalize. The semantics of "Cat" are not captured by image data, eg., at least, DNA is relevant.

The ability of a neural network to learn "any function" is often sold as some deeply magical ability which subtly implies an ability to overcome any semantic gap. As we see here however even learning the True Function over an infinity of 2D images, fails to learn the True Function separating Cat/Dog – 2D images do not contain the pertinent explanatory information.

Deploying and The Semantic Gap

Given the problems with associative models, we should use them when:

(1) the in-sample is guaranteed to look like the out-sample (eg., with handwriting analysis, it is unlikely letter shapes will change).

(2) the in-sample will not look like the out-sample, and we have expert supervision (eg., in fraud analysis, the model suggests fraud to an investigator, rather than, convicts).

(3) the in-sample will predictably fail to generalize from the out-sample and we have automatic control systems to monitor (eg., hard-coded thermostat rules can prevent catestrophic heating, etc.).