Survey of Machine Learning models, cont’d

Chris Cornwell

April 7, 2025

Outline

Decision Trees

Intro

Decision trees are another example of a powerful model function for machine learning. They are often used for classification (binary or multi-label); however, they can be used for regression tasks as well. Similar to how neural networks are capable of producing any function, if the ”size” of decision trees is unrestrained then it can, in theory, produce an arbitrarily close approximation to any desired classification of points in the input space.

Construction of a decision tree

To build a decision tree, begin with a **decision stump** on $\mathbb{R}^d$.
Let $\omega = (b, \theta, j)$ be a triple consisting of $b \in \mathbb{R}$, $\theta \in \{1, -1\}$, and j an integer with $1 \leq j \leq d$. Then, define $f_\omega(\mathbf{x}) = \operatorname{sign}(\theta(b - \mathbf{x}_j))$, where $\operatorname{sign}(z)$ is 1 if $z \geq 0$ and is -1 otherwise. Such functions are called decision stumps.
Note that a decision stump is a special type of half-space model that has normal vector with a single non-zero component (i.e., $\mathbf{w} = (0, \ldots, 0, -\theta, 0, \ldots, 0)$).
Given sample data, $\mathcal{S} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ with $\mathbf{x}_i \in \mathbb{R}^d$, the decision stump which is the best fit may be found simply by minimizing error – that is, find the values of b, θ, and j that achieves the highest accuracy on $\mathcal{S}$. A naive algorithm to find the best one takes on the order of dn^2 operations. However, there is a more efficient approach, taking only $dn\log(n)$ time.

Construction of a decision tree

A decision stump is a decision tree with "depth" 1. One might think of a decision tree as a collection of nested decision stumps – on smaller and smaller subsets of the data.

You begin by splitting the data along some coordinate. In other words, you find a hyperplane $x_j = b$, where $1 \leq j \leq d$ and b is some number and each data point is on one side: for each $\mathbf{x}_i \in \mathcal{S}$, either $x_{i,j} < b$ or $x_{i,j} > b$. This partitions the data into two subsets.

How do you choose where to make a split? You could use the highest accuracy approach described for stumps, but this has disadvantages when the proportion of y_i that are positive, compared to negative, is not about the same.

Often, something called **Information Gain** is used, which is defined via an entropy function e. That is, set $e(r) = -r\log(r) - (1-r)\log(1-r)$. Now, before making a split, set r to be the proportion of y_i that are 1 and m the number of points.

Next, define r_+ (resp. r_-) to be the proportion of points on the positive (resp. negative) side of the split that will have label 1, and let m_+ (resp. m_-) be the numbers of points on the positive (resp. negative) side. The information gain of the split is

$$e(r) - (\frac{m_+}{m}e(r_+) + \frac{m_-}{m}e(r_-)).$$

Multiple branches

The goal of the process is to recursively partition each side. On each step, one chooses the split with maximum Information Gain, at each step restricting to the two subset of data points on one side. The process ends when points in the same part have the same label, or until some predetermined depth is reached. An innermost region (where points have the same label) corresponds to a **leaf** of the tree.

Decision tree model

Start by determining a decision tree on training data, as above. Then, given test data (not yet ”seen” by the model), the decision tree model will check in which partition the test point resides. Then, it labels the test point with the label of the corresponding leaf. Often, in an effort to avoid overfitting, you decide beforehand on the depth of the tree. Since this will result in having more than one label in some of the partitions (and possibly all), the label given to a test point in each leaf is some function of the training labels in the corresponding partition — if the goal is classification, with some categorical labels, the mode could be used; if the goal is regression, with numerical labels, the mean, or the median, could be used.