Lecture 1 Notes

Introduction to Data Science IS360

1 Learning Approach

1.1 Components of Learning

Given an input x and an output y where some pattern/relationship exists $f : X \to Y$, and data consisting of examples of input-output pairs $\mathcal{D} = \{(x_1, y_1), \cdots, (x_N, y_N)\}$, learning is the process of finding an approximate function $g : X \to \hat{Y}$ that maps the pattern between the input-output pairs.

Finding the 'correct' approximate/hypothesis function g involves identifying the most suitable function from a set of candidate formulas $\mathcal{H}$ called the hypothesis set (it follows from this that $g \in \mathcal{H}$) using a learning algorithm $\mathcal{A}$. Together the hypothesis set and the learning algorithm make up the learning model.

Note 1. The learning algorithm $\mathcal{A}$ determines what functions are in the hypothesis set $\mathcal{H}$. Consider a linear learning algorithm, the algorithm will only able to chose a function from the set of all linear functions, even if the actual relationship were to be non-linear.

2 Simple Learning Algorithms

2.1 Perceptron Learning Algorithm (PLA)

The perceptron learning algorithm is a linear classifier used in binary classification problems. It learns a linear decision boundary (a hyperplane) that separates two classes by iteratively updating the weights to minimize misclassification.

Given our set of training points (data) $\{(x_i, y_i)\}_{i=1}^{N}$ where x is a vector representing our features $x \in \mathbb{R}^d$ and y is the corresponding binary label $y \in \{-1, +1\}$, the PLA tried to find the weights $w \in \mathbb{R}^d$ and the bias/threshold b that separate the training points with as few miscalssified points as possible.

$$h(x) = \text{sign}(w^\top \cdot x + b) \tag{1}$$

Which can also be written as

$$h(x) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i \cdot x_i\right) + b\right) \tag{2}$$

Note 2. $w \cdot x + b$ defines a linear plane that separates the space into 2 regions (hyperplane), the region above the plane is classified as 1 and the region below is classified as -1.

The PLA starts with random (or 0) values for w and b, for each example (x_i, y_i) in the training set we calculate the predicted class $\hat{y}$

$$\hat{y} = \text{sign}(w^\top \cdot x + b) \tag{3}$$

For any misclassified points $\hat{y} \neq y$ we update the weight and threshold as follows

$$w \leftarrow w + \eta \cdot y_i \cdot x_i \tag{4}$$

$$b \leftarrow b + \eta \cdot y_i \tag{5}$$

Where η (eta) is the learning rate.

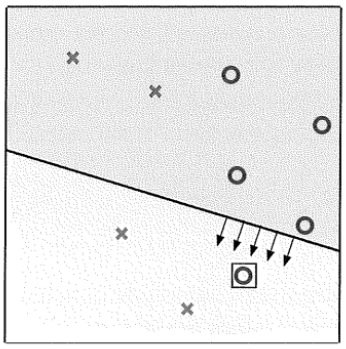

Figure 1: The update functions (4) and (5) serve to move the boundary in the direction of the misclassified point

3 Types of Learning

See Lecture 1 slides 16-18 and book pages 11-15 for the simplified definitions of supervised, unsupervised, and reinforcement learning.