Machine Learning: A Probabilistic Perspective

immediate

December 31, 2018

1 Introduction

1.1 Types of machine learning

Predictive or Supervised Learning The goal is to learn a mapping from inputs to outputs given a labeled set of input-output pairs.

Descriptive or Unsupervised Learning The goal is to find interesting patterns in a given dataset.

1.2 Supervised Learning

Classification Learning a mapping from inputs x to outputs y where $y \in \{1, ..., C\}$, with C being the number of classes. Classification can be formalized as **function approximation**. Assume $y = f(x)$ for some unknown function f. We estimate f given a labeled training set, and make predictions from the estimated function, which we denote $\hat{f}(x)$.

Binary Classification Where the number of classes is 2.

Multiclass Classification Where the number of classes is greater than 2.

Multi-label classification Where each instance may belong to multiple classes.

Generalization Making predictions on inputs not in the training set.

Probabilistic Predictions Given a model that outputs a probability for each class, we can make our "best guess" for the class of the instance by returning the most probable class label; the mode of the distribution. This is called the **MAP Prediction** or **Maximum A Posteriori**, meaning that the class has the highest likelihood given the instance.

$$\hat{y} = \hat{f}(x) = argmax_c \, p(y = c|x, D) \tag{1}$$

where D is the training dataset.

Real World Applications of Classification

- Document Classification
- Spam filtering
- Image Classification
- Handwriting Recognition
- Face detection and Recognition

Regression Like classification except that the response variable you are trying to predict is continuous.

1.3 Unsupervised Learning

The goal is to discover *structure* in the data; this is sometimes called *knowledge discovery*. Can be formalized as **density estimation**, meaning building models that return a probability density for each input: $p(x_i|\theta)$. **Supervised and unsupervised learning can be distinguished as conditional and unconditional density estimation, respectively.**

1.3.1 Clustering

The problem of separating data into groups. Let K denote the number of clusters. The first goal is to estimate a probability distribution over the number of clusters, $p(K|D)$. The second goal is to estimate which cluster each point belongs to. The cluster each point belongs to would be known as a *latent* or *hidden* variable, because it isn't observed directly in the data.

Discovering Latent Factors It's often useful to reduce the dimensionality of data to a lower dimensional subspace that caputres the "essence" of the data. This is called **dimensionality reduction**.

The motivation is that the raw data may be high dimensional but there may be only a small number of degrees of variability, corresponding to latent factors.

Discovering Graph Structure Sometimes we measure a set of correlated variables, and we would like to discover which ones are most correlated with others. This can be represented by a graph G, in which nodes represent variables, and edges represent direct dependence between variables

Matrix Completion Sometimes we have missing information, and inferring plausible values for missing values from existing values is called **imputation**. This is sometimes called *matrix completion*. An example application is *image inpainting*, in which holes in an image are filled in with realistic textures and shapes.

Inferring values in an extremely large, extremely sparse matrix is sometimes called *collaborative filtering*.