Notes for Machine Learning: A Probabilistic Perspective by Kevin P. Murphy

Noah Peters

April 23, 2023

Abstract

Notes for Kevin Murphy's Machine Learning: A Probabilistic Perspective. Note template by Pingbang Hu.

Contents

Chapter 1

Introduction

1.1 Machine Learning: An Overview

Machine learning is a set of methods that can automatically detect patterns in data. There are two types: **supervised** and **unsupervised** learning.

Definition 1.1.1 (Supervised learning). **Predictive/Supervised learning**'s goal is to learn a mapping from inputs x to outputs y, given a labeled set of input-output pairs $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^N$.

Definition 1.1.2 (Unsupervised learning). **Descriptive/unsupervised learning** only consists of inputs, $\mathcal{D} = \{x_i\}_{i=1}^N$ and has the goal of finding "interesting patterns" in the data. This is sometimes called **knowledge discovery**.

Here, $\mathcal{D}$ is the **training set**, and N is the number of training examples. In the simplest setting, each x_i is a D-dimensional vector of numbers, which are called **features**. However, in general, x_i could be a complex structured object such as an image, email, etc.

The **response variable**, each y_i, can be anything, but is usually a categorical or nominal variable from some finite set, $y_i \in \{1, \dots, C\}$. When y_i is **categorical**, the problem is known as **classification** or **pattern recognition**, and when it is real-valued, the problem is known as a **regression**.

1.2 Supervised Learning

Here, the goal is to learn a mapping from inputs x to outputs y, where $y \in \{1, \dots, C\}$ with C being the number of classes. One way to formalize the problem is as a **function approximation**: we assume $y = f(x)$ for some unknown function f, and the goal of learning is to estimate the function f given a labeled training set. Then we can make predictions using $\hat{y} = \hat{f}(x)$ (where the hat symbol is used to denote an estimate).

Definition 1.2.1 (Classification). The process of learning a mapping from inputs x to outputs y, where $y \in \{1, \dots, C\}$, with C being the number of classes.

We can formalize the classificaiton problem of as a *function approximation* problem. We assume $y = f(x)$ for some unknown function f, and the goal of learning is to estimate f given a training set, and then to make predictions using $\hat{y} = f(x)$. Given a probabilistic output, we can always compute our *best guess* as to the "true label" using the **mode**.

Definition 1.2.2 (Mode). The most probable class label for a set of labels $\{c_1, c_2, \dots, c_C\}$ is known as the **mode** of the distribution. It is defined as:

$$\hat{y} = \hat{f}(x) = \arg\max_{c=1}^{C} p(y = c|x, \mathcal{D})$$

1.3 Unsupervised Learning

In unsupervised learning, we are given just output data without any inputs. Here, the goal is simply to discover "interesting structure" in the data. We formalize our task as a problem of **density estimation**, where we aim to build models of the form

$$p(x_i|\Theta).$$

The canonical example of unsupervised learning is the problem of **clustering** data into groups. The first goal is to estimate the distribution over the number of clusters, $p(\mathcal{K}|\mathcal{D})$; this tells us if there are subpopulations within the data. Our second goal is to estimate which cluster each point belongs to. Let there be **latent variables** $z_i \in \{1, \ldots, \mathcal{K}\}$ that represent the cluster to which data point i is assigned.

1.4 Basic Concepts in Machine Learning

We will focus on probabilistic models of the form $p(y|x)$ or $p(x)$, depending on whether we are interested in supervised or unsupervised learning, respectively. We first look at the distinction between models with fixed versus a growing number of parameters.

Definition 1.4.1 (Parametric Model). A **parametric model** consists of a fixed number of parameters.

Definition 1.4.2 (Non-parametric Model). A **non-parametric model** consists of a number of parameters that grows with the amount of training data.

Parametric models are often faster to use, but make stronger assumptions about the nature of the data distributions. Non-parametric models are more flexible, but computationally intractable for large datasets. We often use parametric models and make assumptions about the data distribution (either $p(y|x)$ or $p(x)$), which are known as **inductive bias**.

One of the most common models for regression is **linear regression**.

Definition 1.4.3 (Linear regression). Linear regression is defined as a method for regression that asserts that the response is a linear function of the inputs; written as:

$$y(x) = w^T x + \epsilon = \sum_{j=1}^{D} w_j x_j + \epsilon$$

where $w^T x$ represents the inner **scalar product** between the input vector x and the model's **weight vector** w, and ϵ is the **residual error** between our linear predictions and the true response.

Further, we assume that ϵ has a **Gaussian** or normal distribution, denoted by $\epsilon \approx \mathcal{N}(\mu, \sigma^2)$, where μ is the mean and σ^2 is the variance. We now write

$$p(y|x, \Theta) = \mathcal{N}(y|\mu(x), \sigma^2(x)).$$

We assume μ is a linear function of x, so $\mu = w^T x$; and that the noise is fixed: $\sigma^2(x) = \sigma^2$.

Chapter 2

Probability

2.1 Probability Theory

The expression $p(A)$ denotes the probability that the event A is true. We can define a **discrete random variable** X, which can take on any value from the countably infinite set $\mathcal{X}$. We denote the probability that $X = x$ as $p(X = x) = p(x)$, where p is a probability mass function or pmf.

Definition 2.1.1 (Probability Mass Function). A **pmf** defines the probability of the event that $X = x$ as $p(X = x)$ where p satisfies the properties $0 \leq p(x) \leq 1$ and $\sum_{x \in X} p(x) = 1$.

2.2 Discrete Distributions

2.3 Continuous Distributions

2.4 Joint Probability Distributions

2.5 Transformations of Random Variables

2.6 Monte Carlo Approximation

2.7 Information Theory

Chapter 3

Generative Models for Discrete Data

Appendix

Appendix A

Additional Proofs

A.1 Proof of ??

We can now prove **??**.

Proof of ??. See here. ■