Machine Learning: A Probabilistic Perspective

Summarized from K. Murphy's book, Michaelmas 2021

Oussama Chaib

October 2021

Contents

1 Introduction

1.1 Machine learning: what and why?

- Approach of the book: the best way to make machines that can learn from data is to use the tools of probabilistic theory.

- Probability theory: applied to anything involving uncertainties.

  - What is the best prediction?
  - What is the best model?
  - What measurement should I perform next?

- This systematic approach of using probability theory is often referred to as the **Bayesian approach**.

- To avoid upsetting some audiences, we use the more ”neutral” term **probabilistic approach** (some of the methods we use like *maximum likelihood* estimation are not Bayesian, but certainly fall under probabilistic methods).

- **Machine learning:** A set of methods that can automatically detect patterns in data, and then use them to predict future data or to perform other kinds of decision making using that data.

1.2 Types of machine learning

1. **Supervised (predictive):** Learn a mapping from inputs x to outputs y given a training set $D = (x_i, y_i)_{i=1}^{N}$ containing N samples.

   (a) **Classification:** When the output y_i is nominal (categorical) variable of a finite set (i.e: gender).

   (b) **Regression:** When the output y_i is real-valued.

2. **Unsupervised (descriptive):** No specified pattern, no obvious error metric to use (i.e: neural networks).

3. **Reinforcement learning:** Learning how to behave when given occasional reward or punishment signals.

1.3 Supervised learning

1.3.1 Classification

- Typical example: $y = f(x)$ with y is a **finite** number of points (x can be continuous, discrete, or a combination of both).

- We use the hat symbol to denote an estimate (i.e: $\hat{y}$ is an estimate of y).

- We would like to predict the result on novel input x_*, meaning ones that weren't seen before.

- **Probability notation:** Probability of output y given the input x, the training dataset D, and the model M.

- If the model is known and we do not wish to compare models, we drop the M so that: $p(y|x, D, M) \equiv p(y|x, D)$.
- Our best guess (most probable class label, mode of the distribution, MAP: maximum a posteriori estimate) will maximize this probability.

$$\hat{y} = argmax_{c=1}^{C} \; (p(y|x, D, M)$$

1.3.2 Regression

- Just like classification but the response variable y is **continuous**.
- Will be explored further.

1.4 Models for supervised learning

1. **Parametric models:** $p(y|\mathbf{x})$ Fixed number of parameters. Usually faster but require *stronger assumptions* about the nature of the data distribution.
2. **Non-parametric models:** $p(\mathbf{x})$ More flexible but computationally hungry for large datasets.

1.4.1 Linear regression

Can be written as follow:

$$y(\mathbf{x}) = \mathbf{w^T}.\mathbf{x} + \epsilon$$

where $\mathbf{w^T}$ is the vector containing **weights**, and ϵ the **residual error** (or noise) between our linear predictions and the input data.
We often assume that the error vector follows a **Gaussian** or **normal** distribution:

$$\epsilon \sim \mathcal{N}(\mu, \sigma^2)$$

where μ is the **mean** and σ^2 the variance, and $\mathcal{N}$ represents the normal distribution. The parameters of the model can then be defined such that:

$$\theta = (\mathbf{w}, \sigma^2)$$

Linear regression can be used to model **non-linear** relationships by introducing a **basis function $\mathbf{\Phi}(x)$**:

$$p(y|x, \theta) = \mathcal{N}(y|w^T\mathbf{\Phi}(x), \sigma^2)$$

2 A brief review of probability theory

- Probability of a union of two events:

$$p(A \wedge B) = p(A) + p(B) - p(A \vee B)$$

- Joint probability:

$$p(A, B) = p(A \vee B) = p(A|B)p(B)$$

- **Marginal distribution**:

$$p(A, B) = \sum_b p(A|B = b)p(B = b)$$

- **Conditional probability**

$$p(A|B) = \frac{p(A, B)}{p(B)}$$

- **Bayes rule**

$$p(A|B) = \frac{p(B|A)p(A)}{p(B)}$$

For practical cases:

$$\underbrace{p(\theta|data)}_{posterior} = \frac{\overbrace{p(data|\theta)}^{\propto\, likelihood}\ \overbrace{p(\theta)}^{prior}}{p(data)}$$

- **Mean** (expected value) and **variance**

$$\mu_X = E(X) = \sum_\chi xp(x) = \int_\chi xp(x)dx$$

$$\sigma^2 = var[X] = E[(X - \mu)^2] = \int (x - \mu)^2 p(x)dx$$

- **Gaussian (normal) distribution**

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}}\ e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$$

- **Covariance:** Measures the degree of correlation between two random variables X and Y:

$$cov[X, Y] = E((X - E(X)).(Y - E(Y))) = E(XY) - E(X)E(Y)$$

Covariance can be between 0 and infinity. Correlation is normalized between -1 and 1.

- **Monte Carlo approximation**

$$z = \int f(\chi)p(\chi)d\chi = \frac{1}{T}\sum_{s=1}^{S} f(x_s)$$