Machine Learning and Data Mining - Notes

Philip Warton

September 29, 2021

1 9/29

1.1 Lecture 1.2: Statistical Learning - MLE / MAP

Jhomas.

1.1.1 Probability

Definition 1.1. A sample space Ω is a set of all possible outcomes.

Definition 1.2. An event A is a subset of Ω. That is, $A \subset \Omega$.

A probability must be non-negative for any event. Must be 1 for the entire sample space, 0 for the empty set, and must not be double-counting.

Marginalization:

$$P(A) = \sum_{b \in \text{Val}(B)} P(A, B = b) \qquad \text{(discrete)} \tag{1}$$

$$P(A) = \int_{b \in \text{Val}(B)} P(A, B = b) \qquad \text{(continuous)} \tag{2}$$

Conditional Distribution:

$$P(A \mid B) = \frac{P(A, B)}{P(B)} \tag{3}$$

Chain Rule:

$$P(A, B) = P(A \mid B)P(B) = P(B \mid A)P(A) \tag{4}$$

Bayes Rule:

$$P(A \mid B) = \frac{P(B \mid A)P(A)}{P(B)} \tag{5}$$

Definition 1.3. A random variable X is a mapping between events in Ω to numbers. They can be discrete or continuous.

A probability density describes the mapping from values of a random variable X to probabilities. Some common discrete distributions are the following:

$$\text{Bernoulli:} \quad p_X(x) = \theta^x (1-\theta)^{(1-x)} \tag{6}$$

$$\text{Categorical:} \quad p_X(x) = \theta_x \tag{7}$$

Common continuous distributions:

$$\text{Gaussian:} \quad f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \tag{8}$$

Definition 1.4. The expectation of a random variable is given by

$$E_X[g(x)] = \int_{x \in \text{Val(X)}} f_x(x) g(x) dx$$

1.1.2 MLE Algorithm

There are two steps to maximum likelihood estimation:

(i) Assume a probabilistic model of how the data was generated

(ii) Find $\hat{\theta}_{MLE}$ that maximizes the probability (or likelihood) of generating the training data under the probabilistic model.