Notes on prml

Hao Su

July 16, 2025

Contents

Chapter 1

Introduction

1.6 Information Theory

We begin by considering a discrete random variable x and we ask how much information is received when we observe a specific value for this variable, which can be viewed as the 'degree of surprise' on learning the value of x. Our measure of information content will therefore depend on the probability distribution $p(x)$, and we therefore look for a quantity $h(x)$ that is a monotonic function of the probability $p(x)$ and that expresses the information content. The form of h can be found by noting that if we have two events x and y that are unrelated, then the information gain from observing both of them should be the sum of the information gained from each of them separately, so that $h(x, y) = h(x) + h(y)$. Two unrelated events will be statistically independent and so $p(x, y) = p(x)p(y)$. From these two relationships, it is easily shown that $h(x)$ must be given by the logarithm of $p(x)$ and so we have

$$h(x) = -\log_2 p(x)$$

where the choice of basis is arbitrary. Now suppose that a sender wishes to transmit the value of a random variable to a receiver. The average amount of information that they transmit in the process is obtained by taking expectation with respect to the distribution $p(x)$, given by

$$\mathrm{H}[x] = -\sum_x p(x) \log_2 p(x),$$

which is called the *entropy* of x. When using 2 as the basis, $\mathrm{H}[x]$ is indeed the lower bound on the number of *bits* needed to *encode* the state of x.

Chapter 3

Linear Models for Regression

A helpful insight is that when

$$Ax = b$$

is irresolvable (A being not invertible, etc), we turn to $A^\dagger$ instead to achieve an optimal result in the sense that $\|Ax - b\|$ is minimized. This is a motivation for *Moore-Penrose pseudo-inverse* matrix.