AN INTRODUCTION TO PROBABILITY THEORY

MORITZ BLUME

Contents

1. Introduction

This document is a draft. It will be extended step by step whenever I feel that I learned something new about probability theory.

2. Basics

Probability theory is concerned with describing random phenomena mathematically. A basic concept is the *probabilistic experiment*. It is a repeatable experiment with the property that it is not possible to predict the outcome. Accordingly, we refer to a random quantity as the outcome of a probabilistic experiment of a complexity that makes it impossible to predict the outcome. Whether randomness really exists or not is a more philosophical than mathematical question.

The result of a probabilistic experiment ω is called an *elementary event*. All possible elementary events form a set which is called the *sample space* Ω. Any subset of Ω is called an *event*.

Even though we cannot make exact predictions about the outcome of a probabilistic experiment, we can at least define a probability for a certain event. On the way to define probability we first have to look at the term *frequency*.

We repeat an experiment for N times and write down the resulting elementary events $\omega_1, \dots, \omega_N$. The frequency of an event is defined as the number of times that an event has occurred relative to the total number of repetitions:

$$h_N(A; \omega_1, \dots, \omega_N) = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}_A(\omega_i) \quad . \tag{1}$$

An event A occurs if $\omega_i \in A$. $\mathbb{I}_A$ is an indicator function that takes value 1 if A occurred and 0 else.

The more repetitions N we have, the more likely we will receive similar values for the frequency. All those values will lie around the probability of the event. Accordingly, the probability can be defined as

$$P(A) = \lim_{N \to \infty} h_N(A) \quad . \tag{2}$$

3. Random Variables

Often it is convenient to describe elementary events by numeric values (natural numbers, real numbers, etc.). So, instead of writing $P(\omega_1)$ we can define a random variable X that takes the value 1 if ω_1 occurs and write $P(X = 1)$.

We can think of X as a mapping from the sample space to a number.

E. g., be X a random variable that takes values from 1 to 6 which represent the number of a dice. X is uniformly distributed and we write

$$P(X = x) = \frac{1}{6} \quad . \tag{3}$$

In words: the probability that X will have a value x is $\frac{1}{6}$.

It is important to note the difference between X and x: x is a concrete outcome of a probabilistic experiment, while X is a variable that stands for any possible outcome. We also say that x is a realization of a random variable.

4. Conditional Probability and Law of Bayes

We are interested in the probability of A given that we know that B has already occurred.

For our intuitive understanding, we think of elementary events as areas of a size proportional to their probability. No blank spaces are left in between events. The probability of A is then the size of the area made up by A's elementary events divided by the total size of Ω.

The elementary event that occurred lies within the area defined by B (since we know that B has already occurred). The question that remains is: how big is the probability, that it also lies in A?

This depends on the area that A is occupying in B. If A is filling B completely, then A occurs for sure, and therefore the probability of A given B is equal to one. If A fills half of B, then the probability is one half.

In general, the probability of A given that B occurred is simply the fraction of the part of A that lies in B and B:

$$P(A|B) = \frac{P(A \wedge B)}{P(B)} \quad . \tag{4}$$

This formula is a special case of the more general Law of Bayes (not discussed here).

5. Conditional Probability and Conditional Expectation

Expectation of random variable A:

$$\mathbb{E}[A] = \sum_i iP(A = i) \quad . \tag{5}$$

Conditional Expectation of random A given B:

$$\mathbb{E}[A|B] = \sum_i iP(A = i|B) \quad . \tag{6}$$

(which is a function of B). In words: the expectation of A is the mean value of the probability density function of A. The expectation of A given B is the mean value of the probability density function of A given B.

6. Properties of Poisson Distributed Random Variables

In this chapter we will look at some for us important properties of Poisson distributed random variables.

6.1. **Sum of Poisson Distributed Random Variables.** Given two Poisson distributed random variables X_1 and X_2 and its corresponding distribution parameters λ_1 and λ_2. Be $Y = X_1 + X_2$. What is the probability density function of Y?

$$P(Y = n) \overset{(1)}{=} P(X_1 + X_2 = n) \tag{7}$$

$$\overset{(2)}{=} \sum_{k=0}^{n} P(X_1 = k)P(X_2 = n - k) \tag{8}$$

$$\overset{(3)}{=} \sum_{k=0}^{n} \mathrm{e}^{-\lambda_1} \frac{\lambda_1^k}{k!} \mathrm{e}^{-\lambda_2} \frac{\lambda_2^{n-k}}{(n-k)!} \tag{9}$$

$$\overset{(4)}{=} \mathrm{e}^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{1}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \tag{10}$$

$$\overset{(5)}{=} \frac{1}{n!} \mathrm{e}^{-(\lambda_1+\lambda_2)} \sum_{k=0}^{n} \frac{n!}{k!(n-k)!} \lambda_1^k \lambda_2^{n-k} \tag{11}$$

$$\overset{(6)}{=} \mathrm{e}^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1 + \lambda_2)^n}{n!} \tag{12}$$

From (4) to (5) we multiply the whole equation by $\frac{n!}{n!}$ in order to get the binomial coefficient inside the sum. From (5) to (6) we use the well known Binomial formula in order to get rid of the sum.

Finally the result is that the sum of two Poisson distributed random variables is again Poisson distributed, with the parameter being the sum of the initial parameters.

This result can be extended to an unlimited number of random variables (the derivation is not shown here). So, given $X_1, \ldots, X_n$ being independent and Poisson distributed with parameters $\lambda_1, \ldots, \lambda_n$, the sum $\sum_i X_i$ is also Poisson distributed with parameter $\sum_i \lambda_i$.

6.2. **Conditional Expectation.** Given that X_1 and X_2 are independent Poisson distributed random variables with means λ_1 and λ_2. We are now interested in the expected value of X_1 given that the sum of X_1 and X_2 is already known (in the case of ET reconstruction the sum y can be measured):

$$\mathbb{E}\left[X_1 = x | X_1 + X_2 = y\right] \quad . \tag{13}$$

We start with looking at conditional probability $P(X_1 = x|X_1 + X_2 = y)$. If we knew about the conditional probability density function $P(X_1 = x|X_1 + X_2 = y)$, then the conditional expectation would just be the expectation of this conditional probability density.

From Bayes Theorem we know that

$$P(X_1 = x|X_1 + X_2 = y) = \frac{P(X_1 = x \wedge X_1 + X_2 = y)}{P(X_1 + X_2 = y)} \quad . \tag{14}$$

We will first develop the *nominator*:

$$P(X_1 = x \wedge X_1 + X_2 = y) \tag{15}$$
$$= P(X_1 = x \wedge x + X_2 = y) \tag{16}$$
$$= P(X_1 = x \wedge X_2 = y - x) \quad . \tag{17}$$

Now, we have the joint probability of two independent Poisson random variables.

$$P(X_1 = x \wedge X_2 = y - x) \tag{18}$$
$$= P(X_1 = x)P(X_2 = y - x) \tag{19}$$
$$= \mathrm{e}^{-\lambda_1} \frac{\lambda_1^x}{x!} \, \mathrm{e}^{-\lambda_2} \frac{\lambda_2^{y-x}}{(y-x)!} \tag{20}$$

The *denominator* $P(X_1 + X_2 = y)$ is the sum of two independently distributed random variables. We already know that this sum is then also Poisson distributed with mean $\lambda_1 + \lambda_2$. So, (14) denotes as

$$\frac{P(X_1 = x \wedge X_1 + X_2 = y)}{P(X_1 + X_2 = y)} \tag{21}$$

$$= \frac{\mathrm{e}^{-\lambda_1} \frac{\lambda_1^x}{x!} \; \mathrm{e}^{-\lambda_2} \frac{\lambda_2^{y-x}}{(y-x)!}}{\mathrm{e}^{-(\lambda_1+\lambda_2)} \frac{(\lambda_1+\lambda_2)^y}{y!}} \tag{22}$$

$$= \frac{y!}{x!(y-x)!} \frac{\lambda_1^x \lambda_2^{y-x}}{(\lambda_1+\lambda_2)^y} \tag{23}$$

$$= \binom{y}{x} \frac{\lambda_1^x \lambda_2^{y-x}}{(\lambda_1+\lambda_2)^y} \tag{24}$$

$$= \binom{y}{x} \frac{\lambda_1^x}{(\lambda_1+\lambda_2)^x} \frac{(\lambda_1+\lambda_2)^x \lambda_2^{y-x}}{(\lambda_1+\lambda_2)^y} \tag{25}$$

$$= \binom{y}{x} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x \left(\frac{\lambda_2}{\lambda_1+\lambda_2}\right)^{y-x} \tag{26}$$

$$= \binom{y}{x} \left(\frac{\lambda_1}{\lambda_1+\lambda_2}\right)^x \left(1 - \frac{\lambda_1}{\lambda_1+\lambda_2}\right)^{y-x} \quad . \tag{27}$$

So, $P(X_1 = x | X_1 + X_2 = y)$ is a binomial distribution with parameters $\left(y, \frac{\lambda_1}{\lambda_1+\lambda_2}\right)$, and it is well known that its expectation is $y\frac{\lambda_1}{\lambda_1+\lambda_2}$.

7. Maximum Likelihood Parameter Estimation

The goal of Maximum Likelihood (ML) estimation is to find parameters that maximize the probability of having received certain measurements of a random variable distributed by some probability density function (p.d.f.). For example, we look at a random variable Y and a measurement vector $\boldsymbol{y} = (y_1, ..., y_N)^\mathsf{T}$. The probability of receiving some measurement y_i is given by the p.d.f.

$$p(y_i|\boldsymbol{\Theta}) \quad , \tag{28}$$

where the p.d.f. is governed by the parameter $\boldsymbol{\Theta}$. The probability of having received the whole series of measurements is then

$$p(\boldsymbol{y}|\boldsymbol{\Theta}) = \prod_{i=1}^{N} p(y_i|\boldsymbol{\Theta}) \quad , \tag{29}$$

if the measurements are independent. The likelihood function is defined as a function of $\boldsymbol{\Theta}$:

$$L(\boldsymbol{\Theta}) = p(\boldsymbol{y}|\boldsymbol{\Theta}) \quad . \tag{30}$$

The ML estimate of $\boldsymbol{\Theta}$ is found by maximizing L. Often, it is easier to maximize the log-likelihood

$$\log L(\boldsymbol{\Theta}) = \log p(\boldsymbol{y}|\boldsymbol{\Theta}) \tag{31}$$

$$= \log \prod_{i=1}^{N} p(y_i|\boldsymbol{\Theta}) \tag{32}$$

$$= \sum_{i=1}^{N} \log p(y_i|\boldsymbol{\Theta}) \quad . \tag{33}$$

Since the logarithm is a strictly increasing function, the maximum of L and $\log(L)$ is the same.

It is important to note that we do not include any *a priori* knowledge of the parameter by calculating the ML estimate. Instead, we assume that each choice of the parameter vector is equally likely and therefore that the p.d.f. of the parameter vector is a uniform distribution. If such a prior p.d.f. for the parameter vector is available then methods of Bayesian parameter estimation should be preferred. In this tutorial we will only look at ML estimates.

In some cases a closed form can be derived by just setting the derivative with respect to $\boldsymbol{\Theta}$ to zero.

Moritz Blume, Technische Universität München, Institut für Informatik / I16, Boltzmannstrasse 3, 85748 Garching b. München, Germany

E-mail address: blume@cs.tum.edu

URL: http://wwwnavab.cs.tum.edu/Main/MoritzBlume