STATISTICS CONCEPTS USED IN THE DOMAIN OF DATA SCIENCE AND MACHINE LEARNING

XAVIER TANG

2020

Contents

Chapter 1

Distribution

This chapter deals with concepts mainly related to various probability distribution.

1.1 Bernoulli Trials

This is a random experiment with exactly 2 possible outcomes. The probability of success is the same every time experiment is conducted. A similar analogy: Flipping a (possibly) biased coin, each coin has probability p of landing heads (success) and probability $1 - p$ of landing tails (failure).

Closely related to a Bernoulli trial is a binomial experiment, which consists of a fixed number n of statistically independent Bernoulli trails, each with a probability of success p, and counts the number of success. The number k of success in n Bernoulli trials is Binomially distributed. The probability of exactly k success (out of n) is given by the probability mass function:

$$P(k) = \binom{n}{k} p^k (1-p)^{n-k} \tag{1.1}$$

where $\binom{n}{k}$ is the binomial coefficient:

$$\binom{n}{k} = \frac{n!}{k!(n-k)!} \tag{1.2}$$

Example:

Let's say a bank made 100 mortgage loans. It is possible that anywhere between 0 and 100 of the loans will be defaulted upon. You would like to know the probability of getting a given number of defaults, given that the probability of a default is $p = 0.05$. To investigate this, you will do a simulation. You will perform 100 Bernoulli trials. Here, a success is a default. (Remember that the word 'success' just means that the Bernoulli trial evaluates to be True, i.e., did the loan recipient default?) You will do this for another 100 Bernoulli trials. And again and again until we have tried it 1000 times. Then, you will plot a histogram describing the probability of the number of defaults. So we have performed 1000 times of 100 mortgage trials, the histogram should show a maximum at about 5 (because 0.05 × 100) and probability is given in Eqn. (1.1).

Example:

Consider the simple experiment where a fair coin is tossed four times. Find the probability that exactly two of the tosses result in heads.

$$\begin{aligned} P(2) &= \binom{4}{2} p^2 (1-p)^{4-2} \\ &= 6 \times (0.5)^2 \times (0.5)^2 \\ &= \frac{3}{8} \end{aligned}$$

When multiple Bernoulli trials are performed, each with its own probability of success, these are sometimes referred to as Poisson trials.

1.2 Probability Distributed Function

1.3 Cumulative Distributed Function

1.4 Probability Mass Function

1.5 Discrete variable: Binomial Distribution

1.6 Discrete variable: Poisson Distribution

1.7 Continuous variable: Normal/Gaussian Distribution

1.8 Continuous variable: Exponential Distribution

Chapter 2

Metrics

2.1 ROC

2.2 Standard Deviation

2.3 Variance

2.4 Confusion Matrix

2.4.1 Recall

2.4.2 Precision

2.4.3 F1

2.5 P value

2.6 T value

2.7 Z value

2.8 Correlation/Pearson

2.9 Covariance

Chapter 3

Testing

3.1 T test

3.2 Chi-square test

3.3 Z test

3.4 A/B testing

3.4.1 Null hypothesis

3.4.2 Alternative hypothesis

3.4.3 Type I/II error

3.4.4 Statistical Power

3.5 Test of Significance

3.6 Hypothesis testing (one-way and two-way)

3.7 ANOVA

3.8 ANCOVA

3.9 One-sample/Two-sample bootstrap hypothesis test

3.10 Time series: p, d, q parameters, unit root and box test

Chapter 4

Thoerem

4.1 Central Limit Theorem

4.2 Law of the large number

4.3 Naive Bayes Algorithm

4.4 Bayesian Statistics/Bayes Theorem

4.5 Sampling Theory

Chapter 5

General

5.1 Confidence Interval

5.2 Conditional Probability

5.3 Normalisation

5.4 Standardisatio

5.5 Least-squared error

5.6 R-squared error

5.7 Mean-squared error

5.8 Inferential Statistics

5.9 Bias-variance trade off