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

XAVIER TANG

2020

Contents

Chapter 1

Distribution

1.1 Bernoulli Trials

1.2 Discrete variable: Binomial Distribution

1.3 Discrete variable: Poisson Distribution

1.4 Continuous variable: Normal/Gaussian Distribution

1.5 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

Distribution

5.1 Confidence Interval

5.2 Probability Distributed Function

5.3 Cumulative Distributed Function

5.4 Conditional Probability

5.5 Normalisation

5.6 Standardisatio

5.7 Least-squared error

5.8 R-squared error

5.9 Mean-squared error

5.10 Inferential Statistics

5.11 Bias-variance trade off