Chapter 8: Case Study: Imbalanced Learning

Richard Liu

May 10, 2020

Contents

1 Introduction

Section 1

Introduction

Introduction

- Previously, we mainly use accuracy to measure the quality of a machine learning model, but in the case of imbalanced data, this metric is not reliable.
    - Example: Financial Fraud / Rate of Cancer Infection.
- We need some other metrics. To illustrate, we first need to introduce confusion matrix.

Confusion Matrix

True Predicted	1	0
1	True Positive	False Positive
0	False Negative	True Negative

Table: Confusion Matrix

- One gist: Precision: See the first row. Recall: See the first column.
- F1-score: The harmonic average of precision and recall.

Case Study: Consumer Purchase Prediction

- A Typical Imbalanced Learning Problem.
- See here.

Case Study: Consumer Purchase Prediction

True Predicted	1	0
1	13185	4697
0	16795	31720

Table: Boundary = 0.3

True Predicted	1	0
1	7547	10335
0	1654	46861

Table: Boundary = 0.8

Trade-Off

- Recall-Precision Trade-Off: Higher recall rate leads to lower precision, vice versa.
- Bias-Variance Trade-Off: How to behave better?
  - Add more data.
  - Add more complexity.
  - Add more random elements: Ensemble Learning, Sub-sampling, Over-sampling, Under-sampling

Thank you!