Data Mining Applications: Case Study

Richard Liu

June 19, 2020

Contents

Section 1

Classification

Subsection 1

Case 1: Consumer Purchase Prediction

Consumer Purchase Prediction

- Goal: Predict the future purchase behavior of users.
  - Classification Problem
  - Imbalanced Problem (Why?)
- A simple way to construct data: Using previous purchase record. (How?)
- Of course, there are many other important features for this problem (Feature Engineering).

GBDT + LR or SVM?

Problem 1

Is SVM applicable?

Problem 2

If applicable, which model is better?

Problem 3

What about Xgboost?

Confusion Matrix

- Important tool for choosing proper threshold for imbalanced classification problem.

Predicted \ True	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

Predicted \ True	1	0
1	13185	4697
0	16795	31720

Table: Boundary = 0.3

Predicted \ True	1	0
1	7547	10335
0	1654	46861

Table: Boundary = 0.8

An Interesting Method

Problem 4

After each training, we manually mark and remove the data with the same label. The remaining data will be put into the next training epoch. Does it work for better performance? Why?

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
- See here for details.

Evaluation Process

- Online Evaluation and A/B Test
- Offline Evaluation
- Requirement: the data for offline evaluation are similar to those for online evaluation and A/B Test (Data Visualization helps).

Problem 5

Is cross-validation enough for offline validation in this case?

Section 2

Dimension Reduction

Subsection 1

Case 2: Gender Prediction

Gender Prediction

- Goal: Try to predict user's gender.
  - Classification Problem
  - Data: nickname, purchasing record, etc.
  - Model: Naive Bayes (Why?)
  - Problem: Extremely sparse data matrix, high dimensionality. (Solution?)

Problem 6

Is PCA suitable for dimension reduction in this case?

Extension: Other Dimension Reduction Methods

- Variational Auto Encoder
- Pearson Correlation Coefficient
- ...

Section 3

Clustering

Subsection 1

Case 3: User Partition

User Partition

Problem 7

Differences between Classification and Clustering?

- Goal: Divide different users into different groups.
  - Old / New users
  - Active / Inactive users
  - ...
- Similar to Price Discrimination

Problem 8

K-means or DBSCAN?

Evaluation

- Elbow Method (Not Recommended).
- Construct Partition Rules based on the results of clustering, and evaluate by further tasks.

Section 4

Other Problems

Problem 9

In Linear Regression and PCA, we want to normalize our data, why?

Problem 10

What is the difference of GD, SGD and mini-batch GD?

Problem 11

Provide an example of "Curse of Dimensionality".

Closure

- We have not covered all the topics in data mining. (e.g. Optimization) But details for different models are important for understanding them and also for interviews.
- More information could be viewed here.

Thank you!