2017 Data Mining Cup

Lingfei Cui, Weixiao Huang, Shuhao Jiao,
Haoran Li, Weitong Lin, Hugo Mailhot,
Nick Ulle, Jiaping Zhang, Jingyi Zheng

University of California, Davis

April 24, 2017

Overview

Section 1

Introduction

2017 Data Mining Cup

Task released April 5th

- Use historical data to predict revenue for an online pharmacy
- Train on 90 days of user actions
- Predict revenue for each user action over subsequent 30 days
- Model with smallest squared error $\sum_i (r_i - \hat{r}_i)^2$ wins
- Predictions due May 17th

Training Data

	lineID	day	pid	adFlag	availability	competitorPrice	click	basket	order	price	revenue
1	1.00	1.00	6570.00	0.00	2.00	14.60	1.00	0.00	0.00	16.89	0.00
2	2.00	1.00	14922.00	1.00	1.00	8.57	0.00	1.00	0.00	8.75	0.00
3	3.00	1.00	16382.00	0.00	1.00	14.77	0.00	1.00	0.00	16.06	0.00
4	4.00	1.00	1145.00	1.00	1.00	6.59	0.00	0.00	1.00	6.55	6.55
5	5.00	1.00	3394.00	0.00	1.00	4.39	0.00	0.00	1.00	4.14	4.14
6	6.00	1.00	3661.00	0.00	1.00	13.66	0.00	0.00	1.00	10.03	10.03
7	7.00	1.00	3856.00	1.00	1.00	3.03	0.00	0.00	1.00	3.58	3.58
8	8.00	1.00	16963.00	0.00	1.00	8.78	1.00	0.00	0.00	8.75	0.00
9	9.00	1.00	14560.00	0.00	1.00	10.84	1.00	0.00	0.00	12.04	0.00
10	10.00	1.00	4853.00	1.00	1.00	9.12	1.00	0.00	0.00	8.75	0.00

What do the data look like?

Training Data – train.csv

- Each of 2,756,003 rows is one user action for one product
  - click, basket, or order
- `revenue`, a multiple of `price`
- Other features:
  - day, adFlag, availability, price, competitorPrice
- No feature to identify distinct users

Test Data – class.csv

- Same structure as above, excluding user action and `revenue`
- 1,210,767 rows

What do the data look like?

Items Data – items.csv

- Each of 22,035 rows is one item
- Information that doesn't change over time
- Linked to other data sets by product ID
- Other features:
  - manufacturer
  - group ("product group")
  - content, unit
  - pharmForm, genericProduct
  - salesIndex ("dispensing regulation code")
  - category, campaignIndex
  - rrp

Section 2

Exploration Results

Initial Results

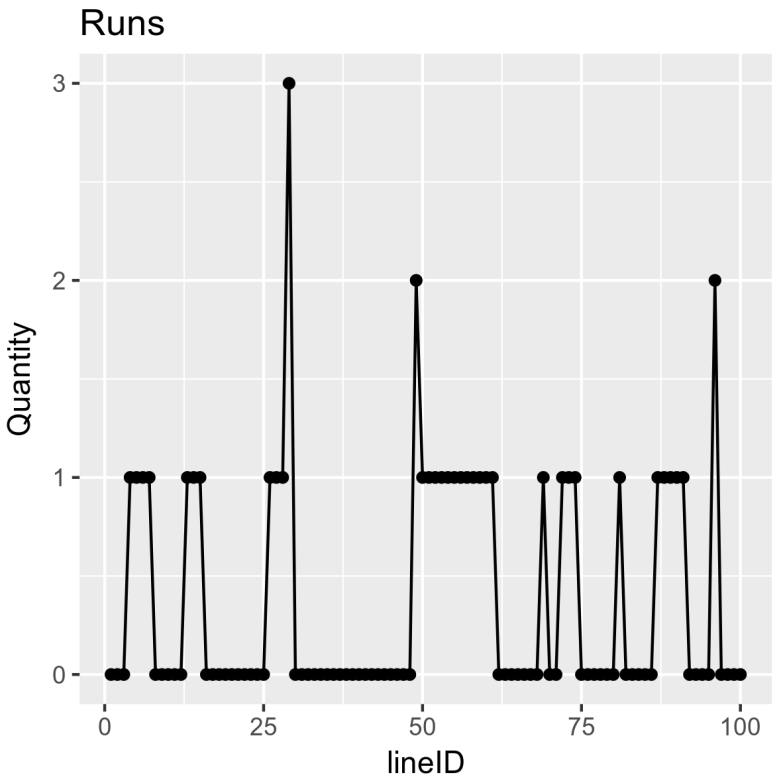
User Actions

- For each unique user and item, only the final action is recorded
- `competitorPrice` is missing for 3.7% of training data

Items

- Items with `availability` 4 rarely ordered—"out of stock"?
- Only 5.1% of items in training data have `salesIndex` 44 or 52
- 3,814 items are identical to another item, excluding ID

A Suspicious Pattern



A Suspicious Pattern

Curiouser and curiouser!

- Runs of "no order" and "order"
- "No order" run-lengths appear to have geometric distribution
- "Order" run-lengths usually less than 10
- Items rarely appear more than once within an "order" run

A Suspicious Pattern

Curiouser and curiouser!

- Runs of "no order" and "order"
- "No order" run-lengths appear to have geometric distribution
- "Order" run-lengths usually less than 10
- Items rarely appear more than once within an "order" run

Each "order" run might be a single shopping basket!

Additional Results

3-character codes in `pharmForm`

- Example: TAB, CRE, KAP, GLO, TRO
- Identifies form of medicine (tablets, syrup, salve, ...)
- German abbreviations, as listed on:
  - DocMorris
  - KohlPharma
- Closely related forms can have distinct codes

Section 3

Feature Engineering

Feature Engineering

Winning teams from many data mining competitions—including the UC Davis 2016 DMC team—say feature engineering was the most important part of their strategy.

Planned Features

- Day of week, day of month, week of month
- Windowed statistics
- Unit type (weight, volume, or pieces)
- Total units (from `content`)
- Price per unit, `competitorPrice` per unit, `rrp` per unit
- Grouped forms (from `pharmForm`)
- ...

Encoding Categorical Features

One-hot encoding

- Each category becomes a separate binary feature
- Models can eliminate unimportant categories
- Unhelpful when novel categories appear in the training data

Likelihood encoding

- For each category, each observation is assigned a likelihood
  - Likelihood is leave-one-out estimate of order probability
  - Likelihood is 0.5 for all observations in test data
- Loses information and doesn't account for order quantity

Section 4

Potential Models

Our Plan

- Generate lots of features
- Use initial model to select important features
  - Importance rankings—from random forest, for example
  - Lasso or other regularization methods
- Build an ensemble of models and refine initial model

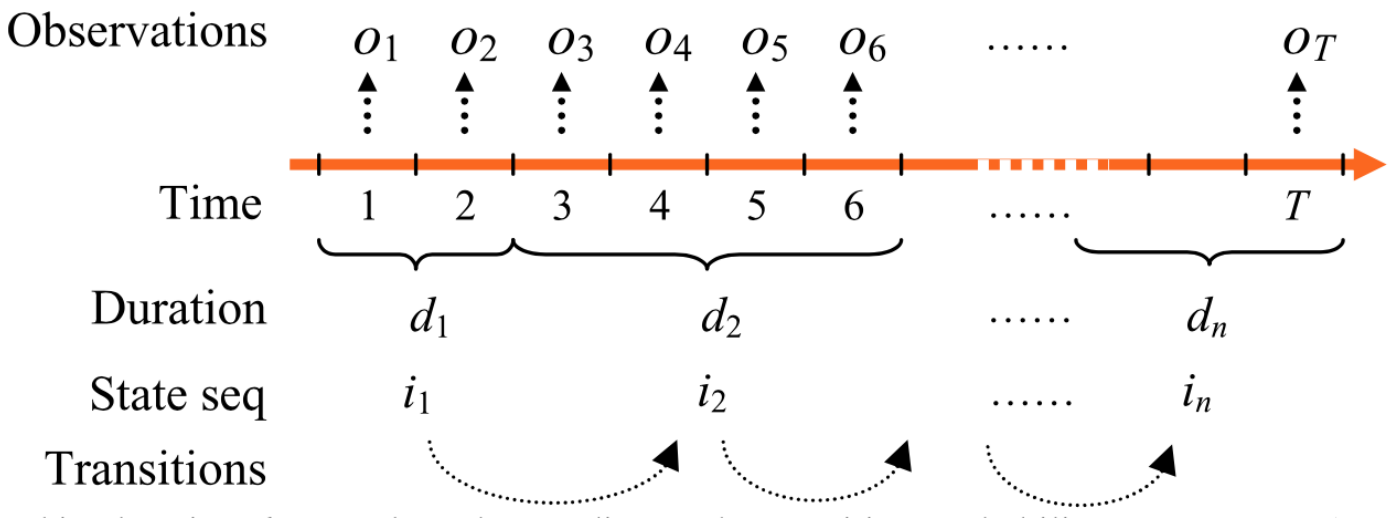
Choosing Models

We need advice on which models to use!

Proposed Models

- Quantity
  - Hidden (semi-)Markov model
  - Linear-chain conditional random field
  - Generalized linear models
  - Boosted random forest
- Revenue

Hidden Markov And Semi-Markov Models



Hidden Markov And Semi-Markov Models

Hidden Markov Model (HMM)

- Each response y_i comes from one of several subpopulations
- Subpopulations may have different distributions
- An unobserved Markov chain determines which subpopulation

Hidden Semi-Markov Model

- Generalization of HMMs
- Time in a state can affect transition probabilities

Advice? Suggestions?