Data Mining for Customized Critical Fractile Solutions

Shawn O'Neil

University of Notre Dame

The Newsvendor's Dilemma

- On each of t days, there is a demand for $d \in [m, M]$ newspapers.
- How many papers x should he order on day i?
  - Each paper costs \$$c$.
  - Can resell each for \$$r$.
  - (Have to decide x before seeing d.)
- Too many or too few ordered leads to lost profits.

Newspapers (And Other Short Lifecycle Products)

- Limited selling season.

  (Who wants Old Newspapers?)

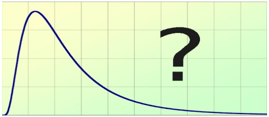
- Strong demand uncertainty.

  (When will the hot story happen?)

- Order decisions must be finalized early.

  (Long printing/manufacturing times.)

Traditional Solutions



- Assume each d drawn from static distribution f
- Maximize expected profit:
  - Known CDF — Critical Fractile Solution:

$$x \text{ s.t. } CDF_f(x) = (r-c)/r$$

  - Unknown CDF, Known μ, σ — Scarf's Solution:

$$x = \mu + \frac{\sigma}{2}\left(\sqrt{(r-c)/c} - \sqrt{c/(r-c)}\right)$$

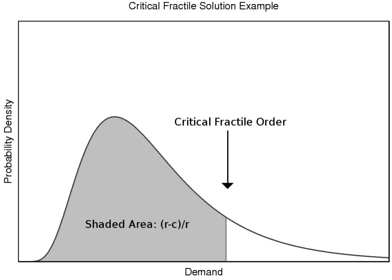
- Can we do better?

Predicting Demand Based on Features

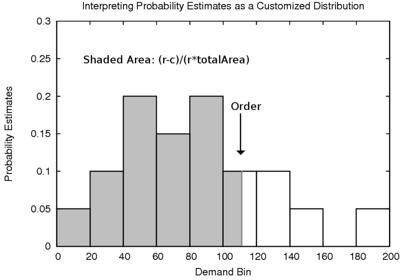
- Idea: Use product features to predict demand
- Example: Predicting CD Sales
    - (genre=hiphop, previousAlbums=4, songs=12)
      Demand: 125,000
    - (genre=grunge, previousAlbums=3, songs=10)
      Demand: 200,000
    - (genre=country, previousAlbums=2, songs=6)
      Demand: 10,000
    - (genre=rock, previousAlbums=6, songs=14)
      Demand: 400,000
    - ...

Customized Demand Distributions

- Don't directly predict on demand:
  - Discretize Demand Space.
  - Predict probabilities for classes.
  - Use critical fractile on customized distribution.



$\longrightarrow$



Test Data: Slashdot Comments

- Slashdot as Newsvendor Items
  - Products: story summaries.
  - Demand: Number of comments gotten.
  - Features: Words in the story.
  - Item cost: \$1, Item resale value: \$4
- Text Classification
  - Need to use feature selection.
  - Ranked stemmed terms by information gain.
  - How many of the top terms, X, should we pick?

Methods Tried: KNN and Naive Bayes

- Use Cost Sensitive Techniques!
- Naive Bayes: Not So Hot
    - Course discretization required for accuracy.
    - Too much information lost.
- K Nearest Neighbors: Better
    - Discretized demand to 250 bins.
    - Best results with k=100, X=200.

Results

- Profits Over Critical Fractile (Validation Set, 2006 Data)
  - 6% Increase over “realistic” critical fractile.
  - 2.2% Increase over “perfect” critical fractile.
- Profits Over Critical Fractile (Test Set, 2007 Data)
  - 4.21% Increase over “realistic” critical fractile.
  - 0.77% Increase over “perfect” critical fractile.
- Respectable?