Analysis of kiva.com Microlending Service

Hoda Eydgahi  Julia Ma  Andy Bardagjy
December 9, 2010
MAS.622j
What is Kiva?

• An organization that allows people to lend small amounts of money via the Internet to micro-finance organizations in the developing world.

• Local micro-finance organizations help local business people to post profiles and business plans on Kiva.

• Once online, lenders select profiles and business plans to fund.
Data Profile
kiva.com microlending service
REQUESTED $703
FUND $507
Borrower Gender

- Male: 22%
- Female: 78%
Loan Sectors

- Clothing
- Food
- Agriculture
- Retail
- Arts
- Services
- Construction
- Health
- Transportation
- Manufacturing
- Housing
- Other
Number of Borrowers

- 1
- [2,5]
- [6-10]
- 11+

LENDERS
Questions

• Can we predict whether a loan will be paid back?
• Which feature(s) best predict the above question?
Classes

- Defaulted (prior probability 2%)
- In repayment, delinquent (prior probability 3%)
- In repayment, not delinquent (prior probability 37%)
- Paid (prior probability 58%)
Features

- Funded amount
- Paid amount
- Sector
- Number of borrowers
- Gender
- Length of use
- Length of description
- Number of periods
- Number of commas
- Number of languages
- Partner ID
- Number of journal entries
- Number of bulk journal entries
- Country code
- Rate of payback
- Percentage paid off
- Dictionary of words
Data Snapshot

```json
{
  header: {
    total: "165452",
    page: 166,
    date: "2010-01-29T20:00:23Z",
    page_size: 1000
  },
  loans: [
    {
      id: 167286,
      name: "Francois",
      description: {
        languages: {
          "en",
        },
        texts: {
          "en": "Francois is a 31-year-old married man with 2 young children. He owns a jewelry shop where he makes and repairs jewelry. Today, Francois is asking to borrow $1,000 to buy tools and appliances needed to make a greater variety of products and thereby increase his sales."
        }
      },
      status: "in_repayment",
      funded_amount: 1000,
      paid_amount: 0,
      image: {
        id: 475617,
        template_id: 1
      },
      activity: "Jewelry",
      sector: "Retail",
      use: "to purchase tools and appliances for his business",
      location: {
        country_code: "LB",
        country: "Lebanon",
        geo: {
          level: "country",
          pairs: "33.833333 35.833333",
          type: "point"
        }
      },
      partner_id: 115,
    }
  ]
}```
Biggest Issues

• Data set is extremely large (165000+ points)
• Extracting features was computationally intensive
• Running algorithms was computationally intensive

Solutions

• Run full data set when possible
• Run code on computing cluster
• Reduce data set to a feasible size depending on the learning method
• Reduce data set size and adjust priors of classes (ie increase prior of Defaulted from 3% to 10%)
Methods
Prior Adjustment

- The dataset was too large to run most algorithms in a practical time
- Reduced the dataset from 165,000 to 30,000 data points
- Adjusted prior probabilities of classes to be more evenly distributed

<table>
<thead>
<tr>
<th></th>
<th>Defaulted</th>
<th>In Repayment, Delinquent</th>
<th>In Repayment, Not Delinquent</th>
<th>Paid</th>
</tr>
</thead>
<tbody>
<tr>
<td>165,000 point data set</td>
<td>3%</td>
<td>3%</td>
<td>36%</td>
<td>58%</td>
</tr>
<tr>
<td>40,000 point data set</td>
<td>13%</td>
<td>13%</td>
<td>37%</td>
<td>37%</td>
</tr>
</tbody>
</table>
k Nearest Neighbor

• Algorithm that classifies by majority vote of $k$-number of nearest neighbors, determined by Euclidean distance

• Tested with odd $k$-values between 1-20

• Disadvantages
  • Classes with more frequent data points could dominate majority vote
  • Sensitive to features that are bad classifiers
k Nearest Neighbor

- Feature sequential forward selection was computationally intensive
- Use leave-one-out method of testing
- Reduce data set by an order of magnitude
  - 165000 → 10000 data points
- Kept prior probability distribution the same
k Nearest Neighbor

- Best features for classification using SFS:
  - Funded amount
  - Number of commas
  - Length of description
- Classification accuracy: 49.2%
LDA

• Assumes probability density functions are normalized Gaussian distributions
• Assumes class covariances are the same
• Class means and covariances are estimated from training set
LDA

• This method was able to be run on the entire dataset.

• We also ran this method on the 30,000 point dataset with the adjusted priors.

• Used SFS to find feature subset that gave best results.

• SBS was also tried but the method didn’t eliminate very many features.
<table>
<thead>
<tr>
<th></th>
<th>Full data set</th>
<th>Modified priors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best feature subset from SFS</strong></td>
<td>Percentage paid off&lt;br&gt;Paid amount&lt;br&gt;# of borrowers&lt;br&gt;# of languages&lt;br&gt;# of journal entries&lt;br&gt;# of commas&lt;br&gt;Rate of payback&lt;br&gt;Sector&lt;br&gt;# of periods&lt;br&gt;&lt;strong&gt;Accuracy: 82.4%&lt;/strong&gt;</td>
<td><strong>Best feature subset from SFS</strong></td>
</tr>
</tbody>
</table>

**LDA**
LDA w/ diagonal covariances

<table>
<thead>
<tr>
<th></th>
<th>Full data set</th>
<th>Modified priors</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Best feature subset</strong></td>
<td>from SFS</td>
<td>from SFS</td>
</tr>
<tr>
<td>Percentage paid off</td>
<td></td>
<td>Percentage paid off</td>
</tr>
<tr>
<td># of journal entries</td>
<td></td>
<td>Partner ID</td>
</tr>
<tr>
<td>Partner ID</td>
<td></td>
<td># of languages</td>
</tr>
<tr>
<td>Paid amount</td>
<td></td>
<td>Paid amount</td>
</tr>
<tr>
<td># of languages</td>
<td></td>
<td># of borrowers</td>
</tr>
<tr>
<td>Gender</td>
<td></td>
<td>Gender</td>
</tr>
<tr>
<td># of borrowers</td>
<td></td>
<td>Rate of payback</td>
</tr>
<tr>
<td>Country code</td>
<td></td>
<td># of commas</td>
</tr>
<tr>
<td>Rate of payback</td>
<td></td>
<td># of periods</td>
</tr>
<tr>
<td><strong>Accuracy</strong>: 85.4%</td>
<td></td>
<td>Length of use</td>
</tr>
<tr>
<td></td>
<td></td>
<td># of journal entries</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Funded amount</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Length of description</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Sector</td>
</tr>
<tr>
<td></td>
<td></td>
<td><strong>Accuracy</strong>: 91.6%</td>
</tr>
</tbody>
</table>
LDA summary

• Features common to all LDA methods
  • Percentage paid off
  • Paid amount
  • # of borrowers
  • # of languages
  • # of journal entries
  • Rate of payback
• Classification accuracy improved on data set with a more even prior probability distribution
Mutual Information

- Used as a feature selection method (to compare with sequential forward selection)
- Mutual information can be tricky in that certain features may be more “informative” when combined with other features
- Calculated mutual information of single, pairs, and triplets of features
Mutual Information

Top 8 Features:
1. Rate of repayment
2. Percentage paid back
3. Paid amount
4. Partner ID
5. Country
6. Number of borrowers
7. Funded amount
8. Length of English description

Calculated:
- $I(X_i, Y)$
- $I(X_i, X_j, Y)$
- $I(X_i, X_j, X_k, Y)$

Red indicates features that were common to all LDA methods (using SFS) and mutual information.
Support Vector Machines

• Linear
  - C=1
  - Accuracy: 86.67%

• 10-fold cross validation
• 90% train, 10% test

• Radial Basis Kernel
  - C=1000
  - G=0.00001
  - Accuracy: 93.36%

• # data points >> # features
• Mapping data to a higher dimensional space via a nonlinear kernel improves classification
Support Vector Machines

\[
\text{minimize } \|\theta\|^2 / 2 + C \sum_{t=1}^{n} \xi_t \\
\text{subject to } y_t (\theta^T \phi(x_t) + \theta_0) \geq 1 - \xi_t \quad t = 1, ..., n
\]

- Support Vector Machines aim to identify the maximum margin classifier
- The value \( C \) in the equation above corresponds to the cost of allowing some points in the training data to be misclassified with the aim of achieving a wider margin
- The equations above are solved using quadratic programming algorithms
Support Vector Machines

- The identity and number of support vectors are dependent on the kernel choice
  - Linear kernel; # SVs: 5997
  - Radial basis kernel, $C = 1000$; # SVs: 5502
- … as well as the value of the cost $C$ in the regularization term
  - Radial basis kernel $C = 100$; # SVs: 5531
  - Radial basis kernel $C = 10$; # SVs: 6420
K-means

• Aims to minimize within-cluster variance:

\[
\arg\min_k \sum_{i=1}^{k} \sum_{x_j \in Z_i} ||x_j - \mu_i||^2
\]

• Estimated K by minimizing Bayesian Information Criterion (BIC)
  - Used K = 20
• Accuracy: 71.11%
K-means

- Clustering depends on initial choice of centroids
  - 5 different initial positions; chose best clustering
- K-mean uses Euclidean distance to determine within-cluster variance
  - Important to set the mean and variance to zero and one, respectively, for the features
  - Important for all algorithms; but particularly important for K-mean since it uses distance as a metric
Why BIC?

• Choosing a good $K$ essentially becomes a model selection problem:
  - Comparing “models” with different numbers of clusters
• As $K$ increases, within-cluster variance decreases
  - May be misinterpreted as a better fit to data
• BIC addresses the problem of overfitting
• BIC includes a penalty term for the number of parameters in the model (regularization)
BIC

\[ BIC_k = -2 MLL_k + d_k \log(n) \]

Where

- \( BIC_k \) is the BIC value for the model \( M_k \)
- \( MLL_k \) is its maximum log likelihood
- \( d_k \) is the number of free parameters to be estimated
- \( n \) is the number of data points

Minimize \( BIC_k \) to find the optimal value for \( K \)
BIC: Obtaining $\text{MLL}_K$

- $K$-means is a hard classifier
- It generates disjoint clusters so that each data point can only belong to one cluster only
- Therefore, it does not produce likelihoods
- To obtain likelihoods, we will interpret the clusters generated as Gaussian mixtures
BIC: Calculating $MLL_K$

$$MLL_k = \sum_{t=1}^{n} \log \sum_{z=1}^{k} q(z)N(x_t; \mu_z, \sigma I)$$

Where

- $x_t$ is the observed parameter vector
- $\mu_z$ is the vector containing the cluster centroids
- $\sigma I$ is the isotropic variance matrix for all clusters
- $q(z)$ is the weight of each cluster
K-means Pseudocode: Obtaining K

FOR $K=1:30$

FOR $iteration = 1:5$

Perform K-means

END

Accept K-means with the best clustering (evaluated based on lower within-cluster variance)

$$\text{optimal}_K = \underset{K}{\text{argmin}}(BIC)$$

END
K-means Pseudocode: Training

Run K-means to cluster data into optimal_K clusters

FOR Cluster = 1:optimal_K

FOR label = 1:4

\[ \Delta_{\text{probability}}(\text{Label}) = P(\text{Label \_in \_cluster}) - P(\text{Label \_in \_dataset}) \]

END

\[ \text{Cluster \_label} = \arg \max_{\text{label}} (\Delta_{\text{probability}}) \]

END
K-means

- The labels didn’t have an even distribution (13%, 13%, 37%, 37%)
- Therefore, we simply couldn’t choose the mode of each cluster as its label
- The performance of K-means was determined using cluster labels that were chosen by the algorithm presented in the previous slide
Gentle AdaBoost

- **Adaptive**: subsequent classifiers are tweaked in favor of points misclassified by previous classifiers
- **Boosting**: combines simple base classifiers to produce a stronger ensemble

Gentle AdaBoost can better handle outliers than classic AdaBoost
Gentle AdaBoost

• Used decision stumps as classifiers
  - 50 stumps; accuracy: 75.73%
  - 100 stumps; accuracy: 80.20%
  - 200 stumps; accuracy: 81.83%

• Accuracy tends to increase as more stumps are added into the ensemble
GOAL

Categorize loan sector using description text
METHODOLOGY - TRAINING

Form a (large) dictionary from the loans

Description:
Form company to solve pattern recognition problem sets
100 Most Common Word Cloud
Remove Unary Words

24,000 → 14,000
Mix loan applications with the dictionary
Reduce Dimensionality

In this case PCA
Support Vector Machine

Projected Vectors

Classification Hyperplanes

Classified Vectors
Support Vector Machine

Projected Vectors

Classification Hyperplanes

Classified Vectors
Mix loan applications with the dictionary
Reduce Dimensionality

Map vectors to previously computed basis vectors
Support Vector Machine

Projected Vectors

Classification Hyperplanes

Classify with previous hyperplanes

Classified Vectors
## RESULTS

<table>
<thead>
<tr>
<th>Dimensionality Reduction</th>
<th>SVM Kernel</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>PCA, 500 Components</td>
<td>Linear</td>
<td>61.3%</td>
</tr>
<tr>
<td>PCA, 500 Components</td>
<td>Epsilon SVR</td>
<td>74.9%</td>
</tr>
<tr>
<td>PCA, 1000 Components</td>
<td>Epsilon SVR</td>
<td>76.2%</td>
</tr>
</tbody>
</table>
Conclusions: Performance

Accuracy:

• KNN: 49.2%
• K-means: 71.11%
• AdaBoost: 81.83%
• Linear Kernel SVD: 86.67%
• LDA with diagonal covariance: 91.6%
• LDA: 92.3%
• Radial Basis Kernel SVM: 93.36%
Conclusions: Best Method

- Radial Basis Kernel SVM performed the best
- Since # data points >> # features, mapping data to a higher dimensional space via a nonlinear kernel improves classification
- It would be interesting to try other nonlinear methods to see if they perform as well
Conclusions: Worst Methods

• Distance-based metrics, KNN & K-means, performed the worst

• This could be due to the fact that Euclidean distance becomes a worse metric of distance as the number of features increases

• LDA and AdaBoost, which are linear methods, performed significantly better than KNN & K-means
  - Therefore, we know that the data is indeed linearly separable
  - KNN & K-means’ relatively poor performance could not be due to the data being potentially nonlinear
Conclusions

• Modern pattern recognition methods can be successfully applied to the analysis to loan microfinance data

• Hands on application of these techniques expose subtleties not apparent to the casual scholar

• With large datasets some algorithms become computationally handicapped
Future Work

• Run all methods on full data set
  • Limited by computing power and time
• Do bag-of-words methods on whole dictionary
• Compare loans from different sectors/countries
• Create UI to help borrowers solidify their business plans, possibly help their chances of repayment