PROJECT PROPOSALS

Due in Two Weeks

Review of projects

- Is the problem well defined?
- Do you have access to the data?
- What type of data mining problems will be addressed?
- What techniques do you need?

Problems

- Loan status
  - Accuracy vs. interpretability
  - Anomalous features/associations
- Sentiment of company
  - Predict stock price from sentiment
  - Predict sentiment from text
- Detect fraudulent transactions
  - Cluster transactions to find anomalies
  - Classification – fraud vs. legit
- Personality type prediction
  - Predict Myers-Briggs (MBTI) type from 50 short posts
- Amazon products
  - Predict category of product
  - Predict helpfulness of review
- Customer churn (retention)
  - Predict customer retention
  - Recommend products to customers
- Google customer revenue
  - Predict log sum revenue per user

Anomaly Detection

Exploration

- Supervised vs. unsupervised
  - Classification vs. clustering
  - Unsupervised: predict label based on cluster
- Is any of the data mislabeled?
  - After model building, look at false positives

Data Mining

- Metrics:
  - Top K Precision
    - Rank by probability of anomaly
    - Test different thresholds
  - Area under ROC curve
    - Plot true positive rate vs. false positive rate
    - Calculate area under this curve

Exploration for Problems

Sensitivity analysis

- How sensitive is the classifier to different parameters?
  - Less data
  - Subset of labels
  - Subset of features (e.g. correlation)

Sentiment

- NLTK (natural language toolkit, python library)
- LIWC (requires subscription)

Data collection

Do you have experience with:

- Using an API (e.g. Twitter API)?
- Web scraping?

Avoid data annotation, if possible

- Use existing methods to label data
- Find existing labeled data (https://toolbox.google.com/datasetsearch)

Lending Club's Loan Data

- Correlation risk
  - Given enough variables and limited data, two features may be correlated
  - Do associations make sense?

Tesla Sentiment

- Data collection
  - Web scraping / social media API
  - Existing datasets?
- Sentiment Analysis
  - NLTK (Python library)

Characterize Fraudulent Transactions

- Cluster transactions, do anomalous transactions appear in same cluster?
- *COMPA: Detecting Compromised Accounts on Social Networks*
  - Egele et al. 2013

Classification of MBTI Personality Types

- Sensitivity analysis
  - How many posts do you need to predict personality?
  - Each aspect of MBTI score
    - Are the predictors of extroverts also predictors of thinkers?

- Role of non-term features
  - URL domain (association analysis) – do extroverts share more youtube videos?

Amazon Product Data

- Reducing data
  - Pick a subset of categories
    - Category 1 vs Category 2
    - Are two categories more similar to each other than to other categories (domain knowledge)?
- Business questions
  - Do users only review products in a specific category?
  - If a user reviews products in 2 or more categories:
    - Are they more helpful in one category?
    - Are they writing fake reviews?
- Graph mining
  - Vertices are users and products
  - Edges are reviews (user -> products) and bought together/viewed together
  - Random walk to generate features (node representation)

Structured Datasets

- The **Credit defaulter graph** has 30,000 nodes representing individuals that are connected based on the similarity of their spending and payment patterns
- The **German credit graph** has 1,000 nodes representing clients in a German bank that are connected based on the similarity of their credit accounts. The task is to classify clients into good vs. bad credit risks considering clients' gender as the sensitive attribute
- Link to the datasets: Graph-Datasets
- Papers for references:
  https://arxiv.org/abs/2108.05233
  https://arxiv.org/abs/2201.03681
  https://arxiv.org/abs/2102.13186

Credits to the TA Francisco

Telco Customer Churn

- Techniques
  - Recommendation systems – is sufficient data provided?
- Business questions
  - Associations between customers – who shops at Telco?

News Classification

- Labeling is cumbersome, use third party websites to help
  - Snopes
  - *https://mediabiasfactcheck.com*
  - Apply existing methods
- or, use existing pre-labeled data
  - SemEval 2019 Task 4 - Hyperpartisan News Detection
  - https://toolbox.google.com/datasetsearch (news article bias)
- Possible data mining questions:
  - Association between author and bias

Effects of Travel on Baseball Performance

- Generalizability
  - Build a model on 2010, how well does it predict 2011 wins?
- Additional data source
  - Climate (https://toolbox.google.com/datasetsearch U.S. Hourly Climate Normals (1981-2010))
- Additional data mining problem:
  - Correlation between time in season and performance

Google store Revenue

- Where on the web do users with higher transaction rates come from?
  - Can be treated as classification problem
  - Pick a threshold, k, where users with revenue > k are positive class, and the rest are negative class
  - Are user selected features (traffic source, channel, geographic features) better than model selected features? (Sensitivity analysis)