Predict User Ratings Based on Review Texts (Yelp Dataset Challenge)

Supervisor: Prof. Pietro Michiardi
Students: MAI Vinh Tuong, CAI Kehe

Outline

- ❖ Introduction
- ❖ Related Work
- ❖ Methodology
  - ➢ Data preparation
  - ➢ Feature extraction + selection
  - ➢ Prediction (Classification, Regression)
- ❖ Experiments and Result Evaluation
- ❖ Conclusion and Future Work

Introduction

Yelp introduced a dataset for research purpose:

- ❖ Now 10 cities across 4 countries.
- ❖ 1.6M reviews and 500K tips by 366K users for 61K businesses.
- ❖ 481K business attributes, e.g., hours, parking availability, ambience.
- ❖ Social network of 366K users for a total of 2.9M social edges.
- ❖ Aggregated check-ins over time for each of the 61K businesses.

Introduction

File format: JSON

```
{
    'type': 'business',
    'business_id': (encrypted business id),
    'name': (business name),
    'neighborhoods': [(hood names)],
    'full_address': (localized address),
    'city': (city),
    'state': (state),
    'latitude': latitude,
    'longitude': longitude,
    'stars': (star rating, rounded to half-stars),
    'review_count': review count,
    'categories': [(localized category names)]
    'open': True / False (corresponds to closed, not business hours),
    'hours': {
        (day_of_week): {
            'open': (HH:MM),
            'close': (HH:MM)
        },
        ...
    },
    'attributes': {
        (attribute_name): (attribute_value),
        ...
    },
}
```

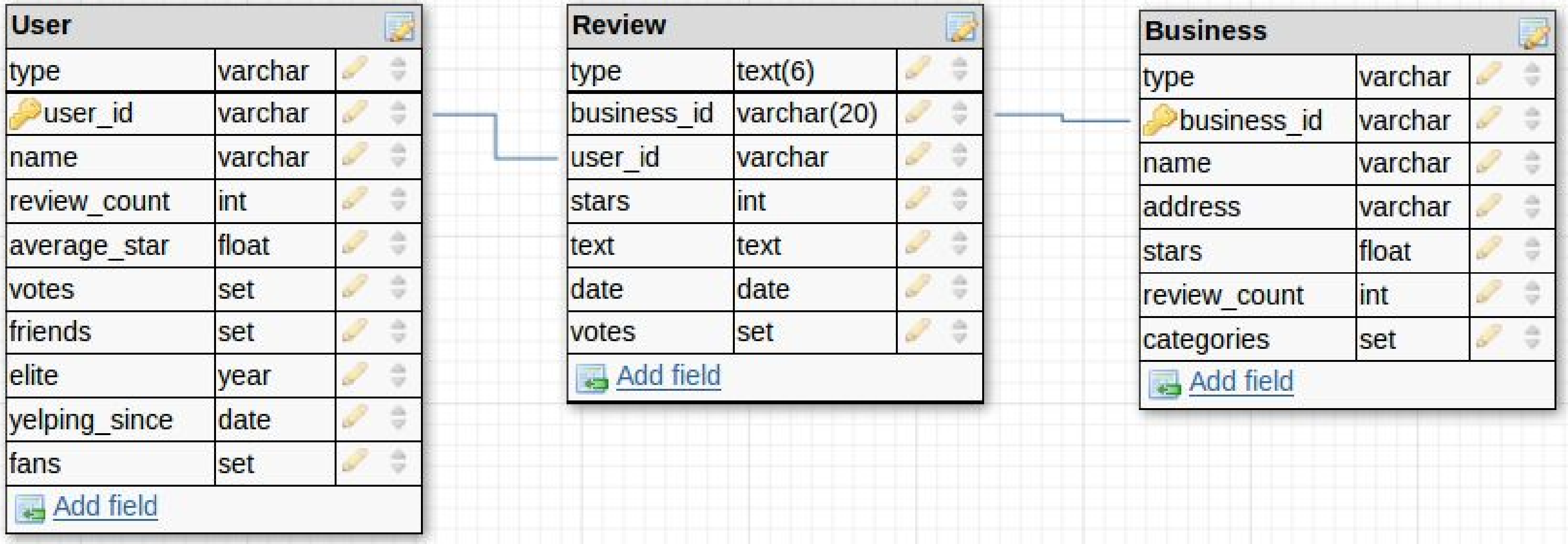
```
{
    'type': 'tip',
    'text': (tip text),
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'date': (date, formatted like '2012-03-14'),
    'likes': (count),
}
```

```
{
    'type': 'review',
    'business_id': (encrypted business id),
    'user_id': (encrypted user id),
    'stars': (star rating, rounded to half-stars),
    'text': (review text),
    'date': (date, formatted like '2012-03-14'),
    'votes': {(vote type): (count)},
}
```

```
{
    'type': 'user',
    'user_id': (encrypted user id),
    'name': (first name),
    'review_count': (review count),
    'average_stars': (floating point average, like 4.31),
    'votes': {(vote type): (count)},
    'friends': [(friend user_ids)],
    'elite': [(years_elite)],
    'yelping_since': (date, formatted like '2012-03'),
    'compliments': {
        (compliment_type): (num_compliments_of_this_type),
        ...
    },
    'fans': (num_fans),
}
```

```
{
    'type': 'checkin',
    'business_id': (encrypted business id),
    'checkin_info': {
        '0-0': (number of checkins from 00:00 to 01:00 on all Sundays),
        '1-0': (number of checkins from 01:00 to 02:00 on all Sundays),
        ...
        '14-4': (number of checkins from 14:00 to 15:00 on all Thursdays),
        ...
        '23-6': (number of checkins from 23:00 to 00:00 on all Saturdays)
    }, # if there was no checkin for a hour-day block it will not be in the dict
}
```

Introduction



Review Rating Examples

Fresh Restaurants Reviews

Joe M. reviewed The Big 4

good but not great. extremely pricey.

Joseph G. reviewed Rudys Cant Fail Cafe

Great diner food with personality. The Bacon Bleu...

Veronica C. reviewed Homeroom

I like Mac and cheese, but I don't like when you...

Veronica C. reviewed Bissap Baobab Oakland

Well I really missed having Senegalese food and...

Related Work

- ❖ Hao Wang, etc.[1] developed a system for real-time analysis of public sentiment toward presidential candidates in the 2012 U.S. election as expressed on Twitter.
- ❖ Researchers from University of California, Irvine [2] explored the problem of classifying Yelp reviews into relevant categories. They demonstrated how reviews for restaurants can be automatically classified into five relevant categories with precision and recall of 0.72 and 0.71 respectively.
- ❖ Mingming Fan, etc. [3] selected the restaurant category from the Yelp Dataset Challenge and utilized a combination of three feature generation methods as well as four machine learning models to find the best prediction result. This not only provides an overview of plentiful long review texts but also cancels out subjectivity.
- ❖ Rakesh C., etc. [4] discussed the combination of topic modeling and sentimental analysis to predict the star rating. Feature extraction methods: Latent Dirichlet Allocation (LDA), term frequency classifier and Non-negative matrix factorization (NMF) are compared and evaluated.

[1] Hao Wang, Dogan Can, Abe Kazemzadeh, Franois Bar, Shrikanth Narayanan. A System for Real-time Twitter Sentiment Analysis of 2012 U.S. Presidential Election Cycle.

[2] http://www.ics.uci.edu/ vpsaini/

[3] Mingming Fan, Maryam Khademi. Predicting a Business Star in Yelp from Its Reviewers Text alone. Eprint arXiv:1401.0864. 01,2014.

[4] Rakesh Chada, Chetan Naik. Data Mining Yelp Data - Predicting rating stars from review text. http://www3.cs.stonybrook.edu/cnaik/files/data mining report.pdf

Methodology

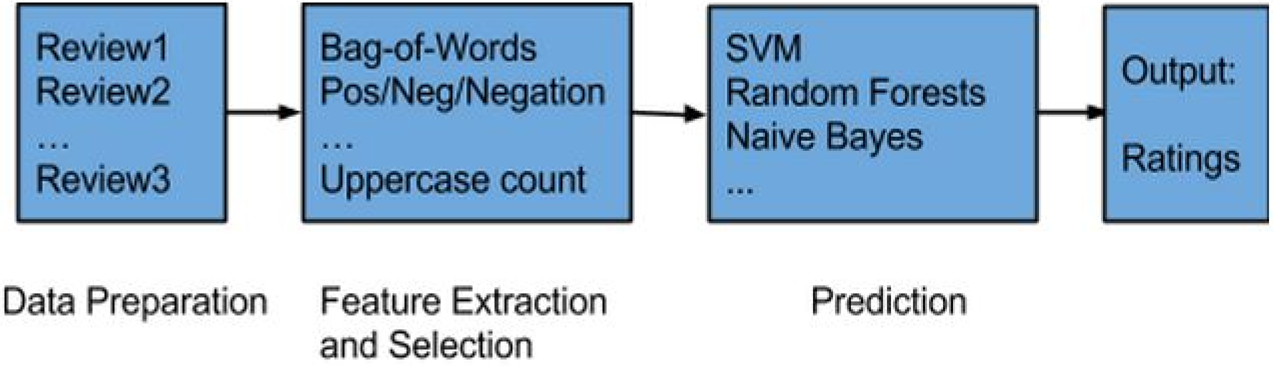
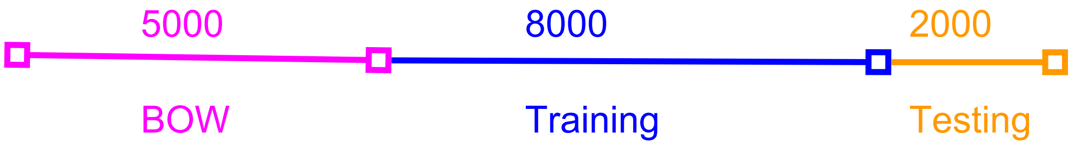


Fig.1 The main flow chart of our process

Data selection

- Randomly select subset of data (15000) from 1M samples in the reviews entity.
- Split the dataset into 3 separated parts:
  - Use 5000 samples to build Bag-of-Words (BOW) dictionary.
  - 8000 samples for training and 2000 samples for testing.



Feature extraction

- Dictionary-based
    - Subjectivity Lexicon dictionary[1]
    - WordStat dictionary[2]
    - Senti-WordNet dictionary[3]
    - **Bag-of-Words dictionary**
- Additional features
    - Day, vote, negation, character count, etc.

[1] http://mpqa.cs.pitt.edu/lexicons/subj_lexicon/

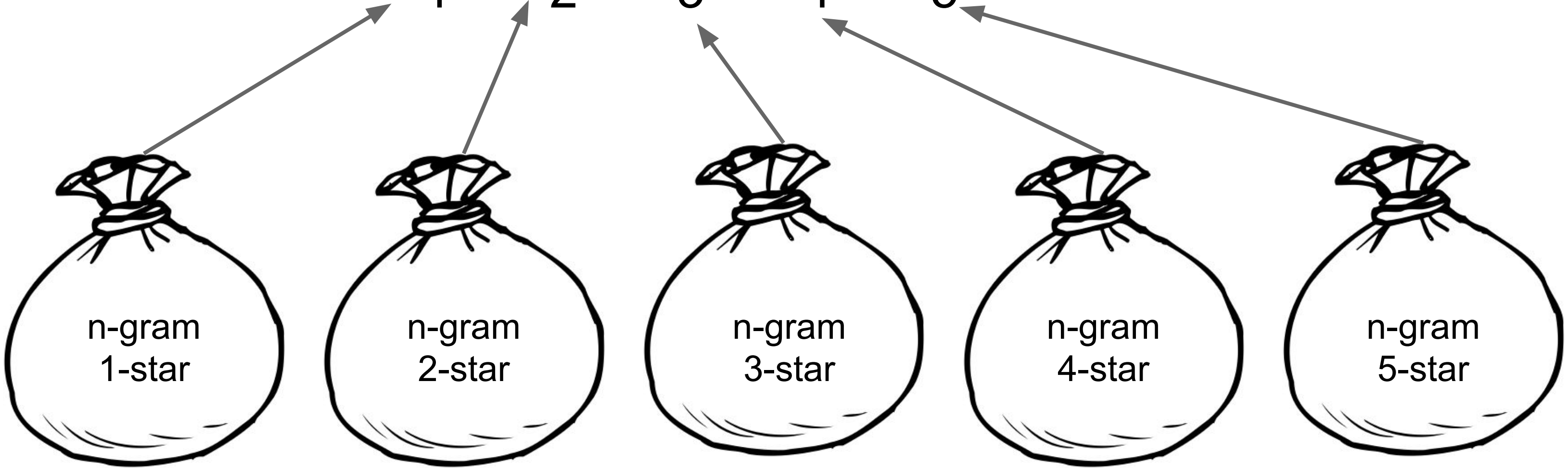
[2] http://www.provalisresearch.com/wordstat/Sentiment-Analysis.html

[3] http://sentiwordnet.isti.cnr.it/downloadFile.php

BOW dictionary

n-gram: uni-gram, bi-gram, tri-gram.

review → $[x_1 , x_2 , x_3 , x_4 , x_5]$



List of features

1	Strong Positive	6	Weak Negative	11	Day	16	Uppercase Count	21	1-Star (BoW)
2	Weak Positive	7	Positive (WordStat)	12	Vote	17	Lowercase Count	22	2-Star (BoW)
3	Strong Neutral	8	Negative (WordStat)	13	Positive (WordNet)	18	Punctuation Count	23	3-Star (BoW)
4	Weak Neutral	9	Negation	14	Negative (WordNet)	19	Alphabetic Count	24	4-Star (BoW)
5	Strong Negative	10	Length	15	Character Count	20	Numeric Count	25	5-Star (BoW)

Feature selection

Select the most important features based on:

- Mutual information (MI)
- Pearson correlation coefficient (PCC)
- Feature_importance_score (tree-based fitted model) (I-Score)

Propose some feature combinations

Feature selection

TABLE II: MI, PCC and IScore between every feature and user ratings

NO.	MI	PCC	IScore	NO.	MI	PCC	IScore
1	0.0928	0.0904	0.0460	14	1.4622	-0.2398	0.0517
2	0.0943	-0.0570	0.0303	15	1.1787	-0.1460	0.0323
3	0.0504	-0.1083	0.0214	16	0.1798	-0.1264	0.0380
4	0.0648	-0.1370	0.0255	17	1.1032	-0.1430	0.0343
5	0.0935	-0.2794	0.0352	18	0.2191	-0.1341	0.0355
6	0.0728	-0.2406	0.0276	19	1.1266	-0.1434	0.0228
7	0.1359	-0.0568	0.0352	20	0.0602	-0.1406	0.0818
8	0.1457	-0.2698	0.0373	21	0.9918	-0.0133	0.0543
9	0.0757	-0.2576	0.0319	22	0.7934	0.1204	0.0622
10	0.6385	-0.1531	0.0324	23	0.6780	0.1198	0.0874
11	0.0027	-0.0017	0.0116	24	0.9904	-0.1389	0.0632
12	0.0449	-0.0266	0.0232	25	0.8340	0.1487	0.0228
13	1.4597	-0.0405	0.0461				

#	Feature Combination
1	[1,2,3,4,5,6,9]
2	[1,2,3,4,5,6,9, 10,11,12]
3	[1,2,3,4,5,6,9, 15,16,17,18,19, 20]
5	[7,8,9]
5	[7,8,9,10,11,12]
6	[7,8,9,15,16,17, 18,19,20]
7	[1,2,3,4,5,6,9, 10,11,12,15,16, 17,18,19,,20]
8	[7,8,9,10,11,12, 15,16,17,18,19, 20]
9	[9,13,14]
10	[9,10,11,12,13, 14]
11	[9,13,14,15,16,17,18,19,20]
12	[9,10,11,12,13, 14,15,16,17,18, 19,20]
13	[21,22,23,24, 25]
14	[10,11,12,21, 22,23,24,25]
15	[15,16,17,18, 19,20,21,22,23, 24,25]

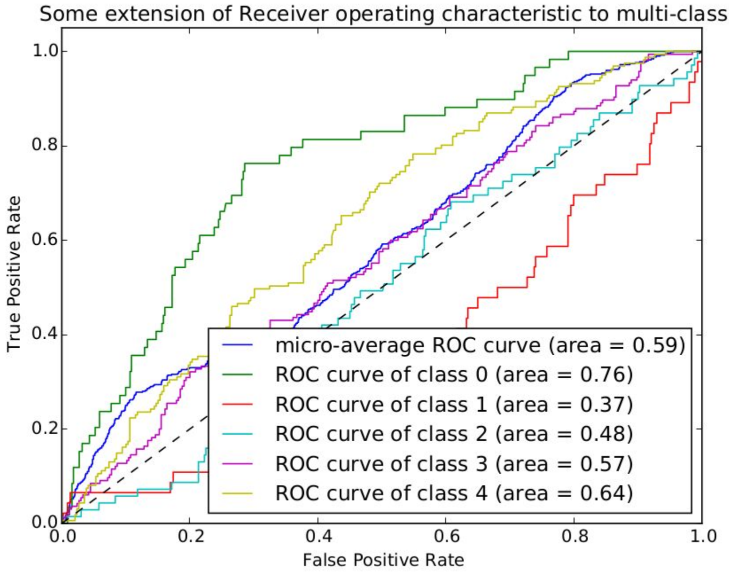
Feature evaluation

Evaluate combinations of features based on:

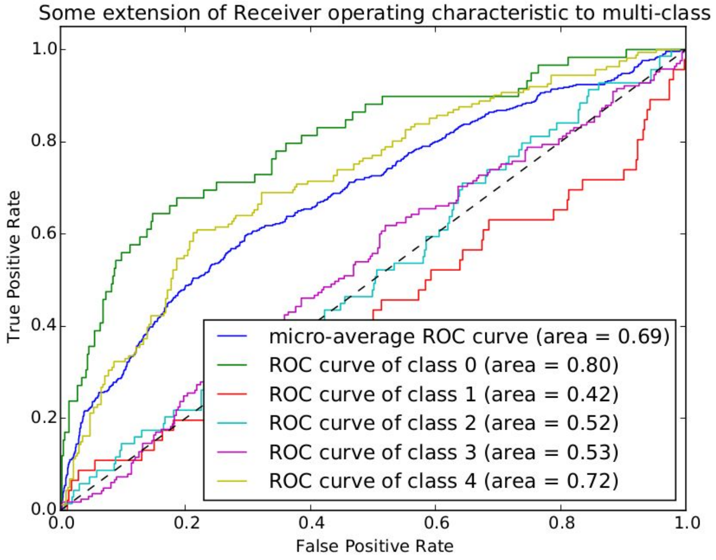
- ROC curve
- Cosine similarity distance
- Experiment

Feature evaluation

Analyze of accuracy and precision of feature combination based on ROC curve and COS distance



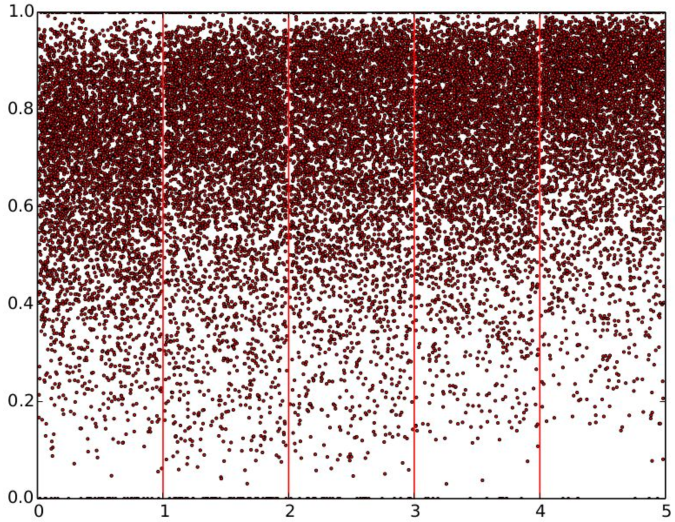
Features [7, 8, 9, 10, 11, 12]



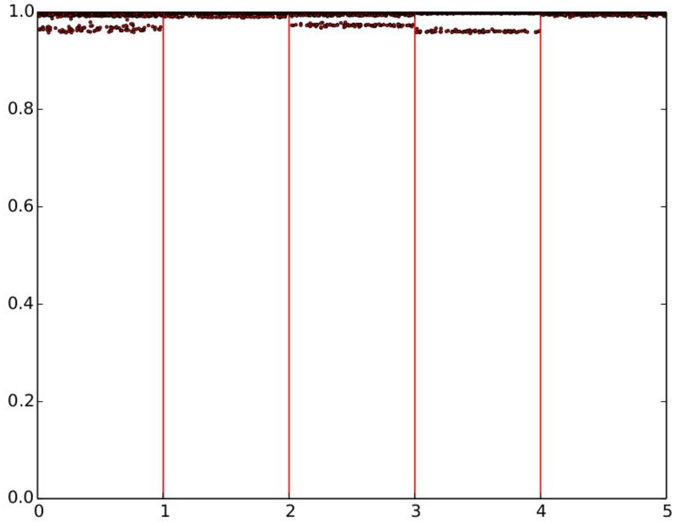
Features [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]

Feature evaluation

Analyze of accuracy and precision of feature combination based on ROC curve and COS distance



Features [1, 2, 3, 4, 5, 6, 9]



Features [15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25]

Feature evaluation

Experiment with these combinations

TABLE III: The classification accuracy by using SVM and Random Forests (RF) classifiers on several feature combinations

#	Feature Combination	SVM			RF		
		Accuracy	Wrong Rate	R2 Score	Accuracy	Wrong Rate	R2 Score
1	[1,2,3,4,5,6,9]	0.480	0.215	0.0003	0.350	0.305	$\boxed{0.2829}$
2	[1,2,3,4,5,6,9, 10,11,12]	0.430	0.205	-0.0393	0.405	0.225	-0.0393
3	[1,2,3,4,5,6,9, 15,16,17,18,19, 20]	0.470	0.225	-0.1186	0.435	0.230	0.0790
5	[7,8,9]	0.440	0.205	-0.0535	0.435	0.230	-0.0110
5	[7,8,9,10,11,12]	0.485	0.220	-0.2092	0.370	0.240	-0.3508
6	[7,8,9,15,16,17, 18,19,20]	0.400	0.210	-0.2262	0.400	0.215	-0.2574
7	[1,2,3,4,5,6,9, 10,11,12,15,16, 17,18,19,,20]	0.480	0.205	-0.1186	0.415	0.215	-0.2290
8	[7,8,9,10,11,12, 15,16,17,18,19, 20]	0.450	0.210	-0.1044	0.430	0.195	-0.0223
9	[9,13,14]	0.505	0.205	-0.0846	0.355	0.295	-0.3367
10	[9,10,11,12,13, 14]	0.425	0.205	-0.1356	0.345	0.245	-0.2574
11	[9,13,14,15,16,17,18,19,20]	0.480	0.200	-0.0563	0.405	0.230	-0.1413
12	[9,10,11,12,13, 14,15,16,17,18, 19,20]	0.430	0.195	-0.0563	0.385	0.225	-0.2205
13	[21,22,23,24, 25]	$\boxed{0.535}$	$\boxed{0.165}$	$\boxed{0.1504}$	0.490	0.185	0.0315
14	[10,11,12,21, 22,23,24,25]	0.460	0.190	-0.0790	0.445	0.175	-0.0450
15	[15,16,17,18, 19,20,21,22,23, 24,25]	0.495	0.180	0.0457	$\boxed{0.515}$	$\boxed{0.165}$	$\boxed{0.1533}$

Prediction

Classification ?

Regression ?

Review -> 4 stars

Review -> 3.6 stars

➔ ***Scikit-learn*** open source library

Prediction

Classification

- SVM
- Random Forest Classifier
- Decision Tree Classifier
- Naive Bayes

Regression

- SVR
- Random Forest Regressor
- Decision Tree Regressor
- Bayesian Ridge

Prediction - Blending

Classification

- Boosting
  - AdaBoost Classifier
- Bagging
  - Bagging Classifier
- Stacking
  - Logistic Regression

Regression

- Boosting
  - AdaBoost Regressor
- Bagging
  - Bagging Regressor
- Stacking
  - Linear Regression

Result evaluation

Classification

- Accuracy rate
    - Spot-on, off by 1, off by 2, ...
- R-squared score

Regression

- RMSE
    - Root mean square error
- R-squared score

Experiments and Result Evaluation

TABLE IV: Classification Experiment Results (1000 samples)

Classifier(s)	Absolutely correct (Spot on)(*)	Off by 1 star	Off by 2 stars	Off by 3 stars	Off by 4 stars	R2 Score	Running Time
SVM (kernel=rbf)	51%	32.5%	8%	6.5%	2%	0.1221	1s
Random Forests (n_ests=510)	51%	36%	7%	4.5%	2%	0.2269	<3s
Naive Bayes	45%	38%	8.5%	6.5%	2%	0.0796	<1s
AdaBoost Decision Tree	46.5%	38%	8.5%	5.5%	1.5%	0.1759	3s
Ensemble (RF & SVM)	53.5%	32%	9%	4%	1.5%	0.2750	3s

TABLE V: Regression Experiment Results (1000 samples)

Regressor(s)	R2 Score	RMSE
SVN	0.1485	1.226
Random Forests (RF)	0.3892	1.038
Bayesian Ridge (BR)	0.3222	1.094
AdaBoost Decision Tree (ADT)	0.4025	1.027
RF,ADT,BR+Linear Regression	0.3931	1.035
RF,ADT,BR+SVR	0.3590	1.064
RF,ADT,BR,SVR+Linear Regression	0.4196	1.012

Experiments and Result Evaluation

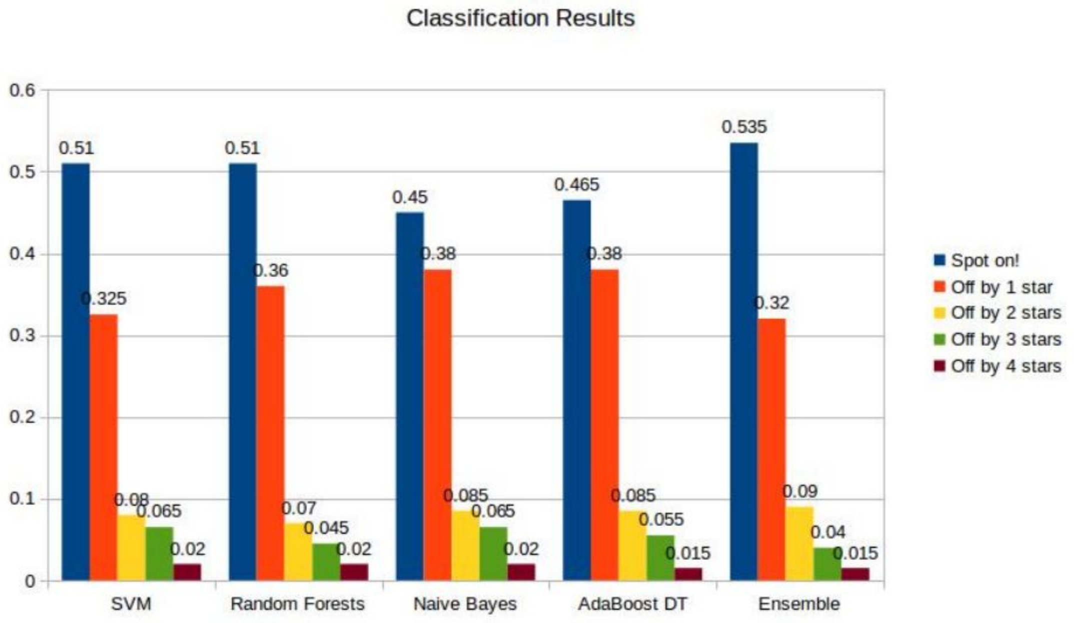


Fig. 4: Classification Results

Experiment Results with Bagging

Bagging (max_samples=0.5, max_features=1.0, n_estimators=10)	**Results** (Wrong: off by 4 stars)
Random Forest	#Right: 0.5315 #Wrong: 0.1535 #R2 score: 0.164845212074
SVM	#Right: 0.5235 #Wrong: 0.1825 #R2 score: 0.00655723653193
Decision Tree	#Right: 0.523 #Wrong: 0.1565 #R2 score: 0.139130559008
Naive Bayes	#Right: 0.4235 #Wrong: 0.19 #R2 score: 0.0217003100044

Experiment Results with Stacking

Classifiers	**Results** (Wrong: off by 4 stars)
Random Forest, SVM, Decision Tree, Naive Bayes (Labels) => Logistic Regression	#Right: 0.528 #Wrong: 0.1595 #R2 score: 0.159416563094
Random Forest, SVM, Decision Tree, Naive Bayes (Probabilities) => Logistic Regression	#Right: 0.5145 #Wrong: 0.1305 #R2 score: 0.247132101887
Random Forest, SVM (Labels) => Logistic Regression	#Right: 0.525 #Wrong: 0.156 #R2 score: 0.174559636566
Random Forest, SVM (Probabilities) => Logistic Regression	#Right: 0.515 #Wrong: 0.1345 #R2 score: 0.238274832498
Naive Bayes (Probabilities) => SVM	#Right: 0.4575 #Wrong: 0.22 #R2 score: -0.257732253318
Random Forest, Decision Tree, Naive Bayes (Probabilities) => SVM	#Right: 0.5055 #Wrong: 0.1375 #R2 score: 0.233131901884

Conclusion and Future Work

- Our first time using machine learning to solve practical problem
- Completely finished prediction task following common machine learning process (data preparation -> feature extraction -> feature selection -> prediction -> evaluation).
- Achieved initial goals:
  - Weekly reports, final report
  - Source code (https://github.com/DistributedSystemsGroup/YELP-DS)
  - Also submitted the results to Yelp challenge
- Proposed an adaptive **BOW** dictionary that works efficiently.
- Extended part: **Blending** method improves accuracy significantly.
- Lesson learned: Characteristics of machine learning models (which model is better than others in specified cases), how to **evaluate features** and how to **evaluate results.**

➔ Future works focus on: negation, smiley, parallelism for larger data.

Work needs to be improved

- **Data selection**
  - Randomly select data with replacement.
  - Try to build BOW from training set.
- **Feature extraction**
  - Explore and use more features.
- **Feature selection**
  - Select combinations automatically.
- **Prediction**
  - Focus more on parameter tuning (bagging, boosting, SVM, tree-based, etc.)
  - Try more stacking scenarios.
- **Evaluation**
  - Change the way to evaluate the results.

Thank you

Question and Answer