CS4487
Course Project

Kaggle Competition: What's Cooking?

Debarun Dhar
52909946

Introduction

Presentation Agenda

1. Preprocessing

2. Feature Representations

3. Classification Methodology

4. More Work / Possibilities

Preprocessing

Cleaning

- Removal of all symbols and numbers using Regular Expressions.
  - Done using the ‘re’ package.

Lemmatization

- Part-of-Speech Tagging
  - Done using ‘wordnet’ from ‘nltk.corpus’
- Token Lemmatization
  - Done using ‘WordNetLemmatizer’ from ‘nltk.stem’

Convert Lists to Strings

- Self-explanatory

Feature Representations

Bag of Ingredients (BoI)

- Wrote a custom tokenizer to split the ingredient strings at the commas.
- Used the tokenizer with CountVectorizer (binary=True) to model ingredient occurrences.

Bag of Words (BoW)

- TF-IDF
  - Default tokenization
  - 1-grams

- Performs better than Bag of Ingredients.

Classification Methodology

Local Testing

- StratifiedShuffleSplit
  - 80% Training set
  - 20% Testing set
  - Preserves class distribution

- High correlation with Kaggle Leaderboard scores.

Approach 1

- Single Estimator

- Best base classifier : LinearSVC + BoW
- Kaggle Score: 0.78932

Approach 2

- Ensemble: VotingClassifier
  - Using BoW + 11 classifiers

- Kaggle Score: 0.80088

Approach 3

- Ensemble: Stacking / Blending
  - Using BoW + 12 base classifiers + MinMax Scaling

- Kaggle Score: 0.80712
- This put me at the 32nd position.

More Work / Possibilities

Dealing with imbalanced classes

- Current approach:
  - class_weight=‘balanced’

- Oversampling and/or undersampling
- SMOTE

Feature Engineering

- Number of ingredients
- Cosine Distance to nearest neighbours
- Distance to nearest neighbours in manifold

Feature Selection

- SelectKBest / SelectPercentile
- SelectFromModel
    - RandomForests / ExtraTrees
    - Logistic Regression

Diversify estimators

- Can yield better results from stacked generalization
- More non-linear models Eg. RBF SVM

- Specialists: For each type of cuisine

Hyper-parameter Tuning

- Tune classifiers at every level
- In a stacked ensemble, almost everything is a hyper-parameter!

Deep Ensemble

- BoW + BoI
- Add more levels
- Back-propagation?!

Thank You!

Any Questions?