# Data Mining: Stage 3 Report

Yimeng Lu• 10.24.2017

# Classification

## Task

Classify papers using titles, indices and abstracts

## Steps

- Generate features
- Run different classification algorithms
- Visualize, evaluate and compare the results

## Data and tools

- Papers from ICDM, KDD and NIPS
- Sklearn, xgboost

# Data preparation and preprocessing

## Data

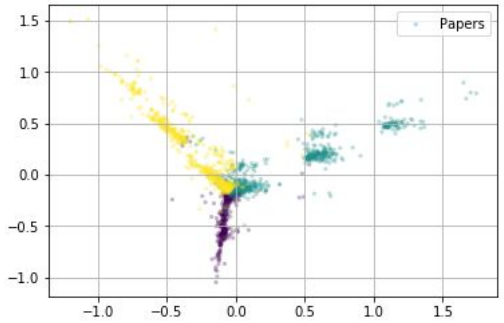
- Use conference titles as original label
- Specially chose ICDM, KDD and NIPS as their topics have less overlap
- About 600 papers for each conference

## Preprocessing

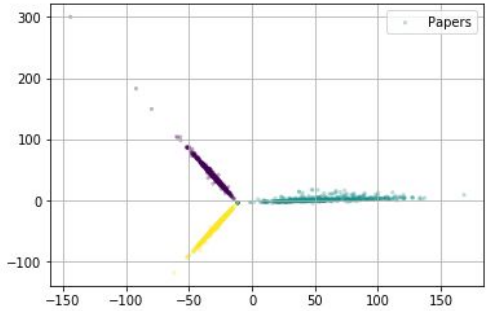
- Clean the data
- BoW->PCA selection

# Visualization for feature generation 1

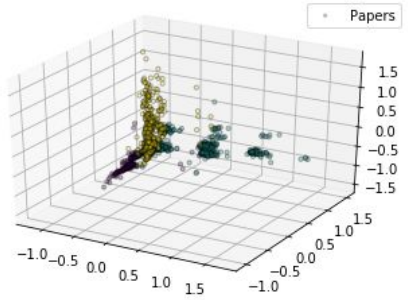
```
scatter_visualization_2D_with_labels(feature_array_title_pca, labels_attached)
```



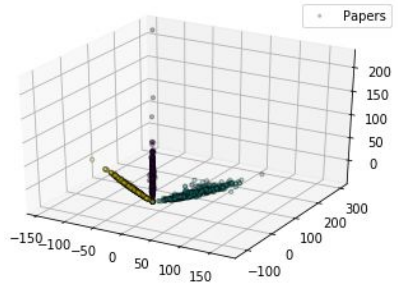

```
scatter_visualization_2D_with_labels(feature_array_abstract_pca, labels_attached)
```




```
scatter_visualization_3D_with_labels(feature_array_title_pca, labels_attached)
```




```
scatter_visualization_3D_with_labels(feature_array_abstract_pca, labels_attached
```



# Visualization for feature generation 2

## Some observations

- This time abstracts give best discrimination
- Three conferences overlap at nearly the same place

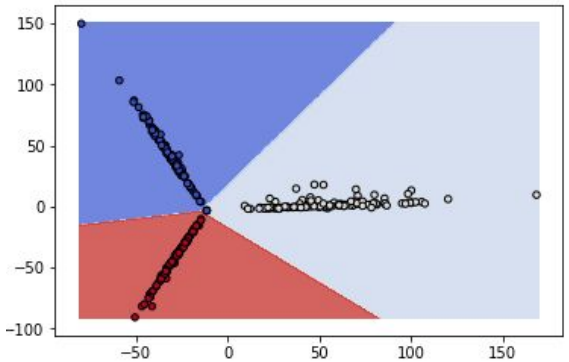
## … and interpretations

- Titles more general but abstracts more specific
- All three conferences are related to a same area(ML, etc.)

# Classification: Basic methods

## Cross validation

- Train:Test = 0.7:0.3



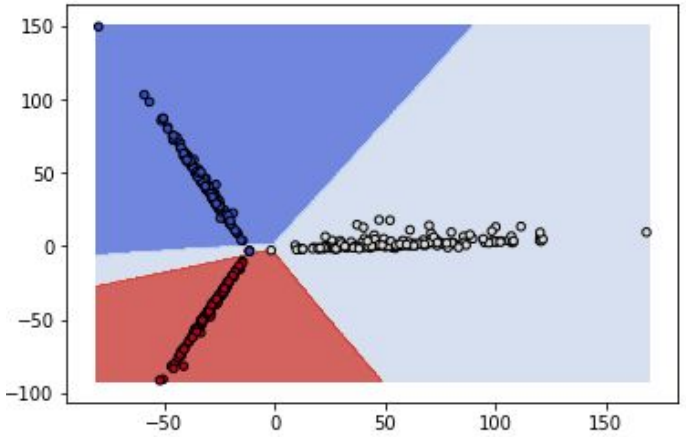

## Linear SVM

- Using only first two dimensions of PCA result
- Accuracy on Test set: 0.992

# Classification: Neural Networks

## Method

- MLPClassifier in sklearn
- /alpha = 1
- Accuracy: 0.995

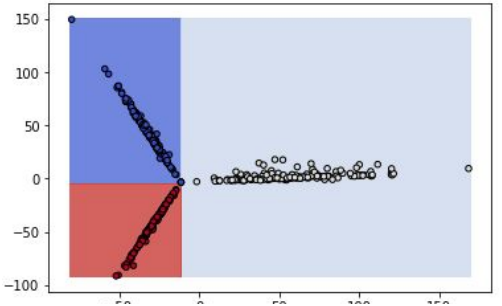


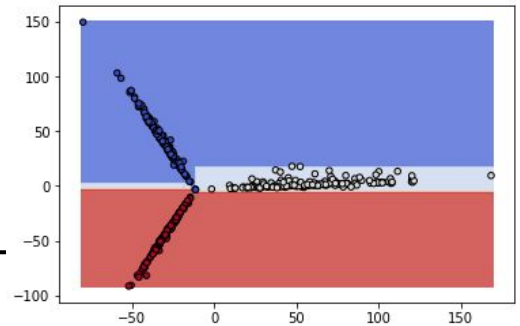
# Classification: Ensemble methods

## Methods tried(10 dim)

- **RandomForest**(0.989)
- BaggingClassifier(0.998)
- **AdaBoostClassifier**(0.996)
- XGBClassifier(0.994)

## Visualization(2 dim)



# Thank you