Principal Component Analysis for Semantic Classification
AMATH 582 Final Project

Benjamin Liu, Kelsey Maass, and Riley Molloy

March 13, 2016

Abstract

Principal component analysis (PCA) and classification by supervised learning are two popular topics in data science today. In this article, we combine techniques from both areas in order to classify news articles based on their word frequency content. We find that we can accurately classify the data by projecting onto a small subset of principal components, reducing the feature space from nearly 10 000 elements to just four. We also compare results from the traditional and robust PCA formulations, and discuss what additional semantic information can be inferred from our results.

1 Introduction and Overview

PCA and Robust PCA

2 Theoretical Background

2.1 Principal Component Analysis

$$A = USV^* \tag{1}$$

2.2 Robust PCA

$$\min_{L,S} \|A - L - S\|_F^2 + \lambda_1 \|L\|_* + \lambda_2 \|S\|_1, \tag{2}$$

where $\|\cdot\|_*$ is the nuclear norm, defined by

$$\|\cdot\|_* = \|\sigma(\cdot)\|_1$$

where σ maps a matrix to a vector of its singular values. Therefore the nuclear norm of a matrix is the sum of its singular values. In addition, $\|\cdot\|_1$ is the vectorized ℓ_1 norm, treating $S \in \mathbb{R}^{m\times n}$ as a vector in $S \in \mathbb{R}^{mn}$.

2.3 Classification Methods

3 Algorithm Implementation and Development

3.1 Article Data

Algorithm for turning articles into wordcount / wordfrequency data. Stemming, low wordcount removal, stopwords.

3.2 Robust PCA

3.3 Classification Schemes

3.3.1 Nearest Mean

3.3.2 K-Nearest Neighbors (KNN)

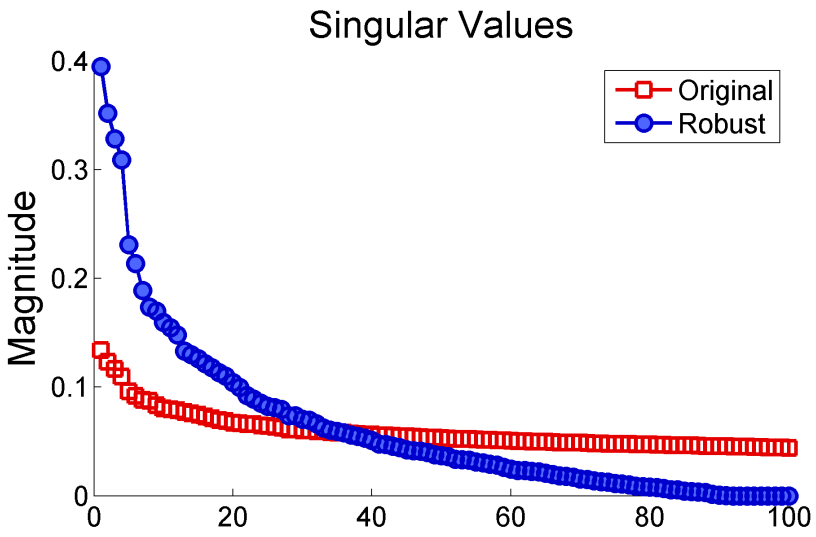
3.3.3 Logistic Regression

3.3.4 Neural Network

3.3.5 Support Vector Machine

4 Computational Results

4.1 PCA and Robust PCA



4.2 Classification

5 Summary and Conclusions

Appendix A: MATLAB Functions

ind = knnsearch(Xtrain,Xtest,k) Finds the **k** elements of **Xtrain** that are nearest to **Xtest**, and returns the indices of these elements in **ind**.

svmtrain

svmdecision

Appendix B: MATLAB Codes