Week 13 Lecture: Applied Machine Learning

L. Jason Anastasopoulos `ljanastas@princeton.edu`

April 2, 2018

Dimensionality Reduction Methods

- All of the models discussed used the original predictors in some form.
- Dimensionality reduction methods transform the predictors into variable clusters and then use these transformed variables to fit a model.

Dimensionality Reduction Methods

Consider a linear combination $Z_1, \cdots, Z_M$ of the features $X_1, \cdots, X_1$ such that $M < p$ where:

$$Z_m = \sum_{j=1}^{p} \phi_{jm} X_j$$

For some constants $\phi_1, \cdots, \phi_M; m \in [1, M]$. We can then fit the linear regression model:

$$y_i = \Theta_0 + \sum_{m=1}^{M} \Theta_m z_{im}$$

Dimensionality Reduction Methods

The model

$$y_i = \Theta_0 + \sum_{m=1}^{M} \Theta_m z_{im}$$

now has $M + 1 < p + 1$ predictors and, if chosen well, can result in a better fit through estimating fewer parameters than the original regression model.

Dimensionality Reduction Methods

To be clear take a simple linear regression model with three features:

$$Y = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \theta_3 X_3 + \epsilon$$

Define $z_1 = \phi_1 X_1 + \phi_3 X_3$ and $z_2 = \phi_2 X_2$. We can now estimate the reduced model:

$$\begin{aligned} Y &= \Theta_0 + \Theta_1 z_1 + \Theta_2 z_2 + \epsilon \\ &= \Theta_0 + \Theta_1(\phi_1 X_1 + \phi_3 X_3) + \Theta_2(\phi_2 X_2) + \epsilon \end{aligned}$$

Dimensionality Reduction Methods

- Again the key here is that we are estimating a model with fewer predictors, thus reducing the *dimensionality* of the model.
- This is especially useful in problems where p is large relative to n. Variance will be significantly reduced in this case and this is not uncommon in machine learning problems (ie text analysis)

All dimensionality reduction methods involves two steps

1. Transformed predictors $Z_1, \cdots, Z_M$ are first obtained.

All dimensionality reduction methods involves two steps

1. Transformed predictors $Z_1, \cdots, Z_M$ are first obtained.

2. A model is fit using the M predictors.

All dimensionality reduction methods involves two steps

1. Transformed predictors $Z_1, \cdots, Z_M$ are first obtained.

2. A model is fit using the M predictors.

- There are several methods for accomplishing this but we will focus on principal components analysis.

Principal Components Analysis (PCA)

$$f : \mathcal{X} \rightarrow \mathcal{F}$$

$$\mathcal{X} \in \mathbb{R}^{nxp}, \mathcal{F} \in \mathbb{R}^{nxm}; p << m$$

- PCA is often discussed in the context of *unsupervised learning* and we'll discuss it in that context later on in the semester.
- It's a popular means of transforming a high dimensional feature space $\mathcal{X}$ into a very low-dimensional space $\mathcal{F}$

Principal Components Analysis (PCA)

- **First principal component** is the dimension along which the data vary the most and would be the most useful for a regression approach.

```
# Predicting political party with votes
library(mlbench)
data(HouseVotes84)
head(HouseVotes84)
```

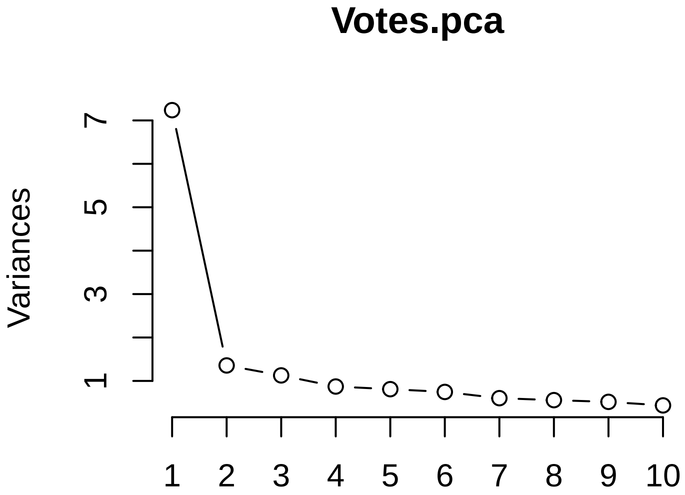
```
##         Class   V1 V2 V3   V4   V5 V6 V7 V8 V9 V10  V11  V12 V13 V14 
## 1 republican    n  y  n    y    y  y  n  n  n   y <NA>    y   y   y
## 2 republican    n  y  n    y    y  y  n  n  n   n    n    y   y   y
## 3   democrat <NA>  y  y <NA>    y  y  n  n  n   n    y    n   y   y
## 4   democrat    n  y  y    n <NA>  y  n  n  n   n    y    n   y   n
## 5   democrat    y  y  y    n    y  y  n  n  n   n    y <NA>   y   y
## 6   democrat    n  y  y    n    y  y  n  n  n   n    n    n   y   y
##    V16
## 1    y
## 2 <NA>
## 3    n
## 4    y
## 5    y
```

Predicting political party from votes, 1984

```
##
## Call:
## lm(formula = Party ~ ., data = data.frame(Votes))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.82054 -0.04439  0.01879  0.08784  0.70224
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.7671523  0.1423941   5.388 1.20e-07 ***
## V1          -0.0264965  0.0213872  -1.239 0.216080
## V2          -0.0282866  0.0200216  -1.413 0.158459
## V3          -0.1988778  0.0296838  -6.700 6.76e-11 ***
## V4           0.6314606  0.0322082  19.606  < 2e-16 ***
## V5           0.0768241  0.0379082   2.027 0.043339 *
## V6          -0.0481934  0.0272663  -1.768 0.077872 .
## V7           0.0642615  0.0288318   2.229 0.026355 *
## V8           0.0559900  0.0352681   1.588 0.113144
## V9          -0.0855327  0.0314344  -2.721 0.006780 **
## V10          0.0478126  0.0187176   2.554 0.010990 *
## V11         -0.1240756  0.0199903  -6.207 1.30e-09 ***
```

Predicting political party from votes, 1984

- Can the votes be explained with a single dimension?



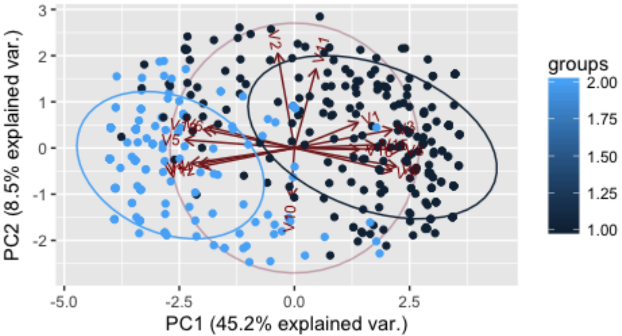
Predicting political party from votes, 1984

- Can the votes be explained with a single dimension?

```r
summary(Votes.pca)
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5     PC
## Standard deviation     2.6901 1.16470 1.06151 0.93320 0.90006 0.8638
## Proportion of Variance 0.4523 0.08478 0.07043 0.05443 0.05063 0.0466
## Cumulative Proportion  0.4523 0.53706 0.60749 0.66192 0.71255 0.7591
##                            PC7     PC8     PC9    PC10    PC11     PC
## Standard deviation     0.77631 0.74664 0.71935 0.66043 0.64191 0.576
## Proportion of Variance 0.03767 0.03484 0.03234 0.02726 0.02575 0.020
## Cumulative Proportion  0.79686 0.83170 0.86404 0.89130 0.91705 0.937
##                           PC13    PC14    PC15    PC16
## Standard deviation     0.56507 0.52220 0.48542 0.40914
## Proportion of Variance 0.01996 0.01704 0.01473 0.01046
## Cumulative Proportion  0.95777 0.97481 0.98954 1.00000
```

Predicting political party from votes, 1984



Predicting political party from votes, 1984

- Took 16 dimensions, reduced to 1 or 2 that still explain about 50% of the variance.
- Can use these dimensions in regression for comparison.
- Let's just use dimensions one and two

Predicting political party from votes, 1984

$$Party = \Theta_0 + \Theta\pi_1 + \Theta_2\pi_2$$

- Took 16 dimensions, reduced to 1 or 2 that still explain about 50% of the variance.
- Can use these dimensions in regression for comparison.
- Let's just use dimensions one and two.

Predicting political party from votes, 1984

$$Party = \Theta_0 + \Theta\pi_1 + \Theta_2\pi_2$$

```
pi1<-Votes.pca$x[,1]
pi2<-Votes.pca$x[,2]
summary(lm(Party~pi1 + pi2))
```

```
##
## Call:
## lm(formula = Party ~ pi1 + pi2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.9021 -0.1181  0.0211  0.1560  0.9177
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)  1.386207   0.012918 107.312   <2e-16 ***
## pi1         -0.144037   0.004807 -29.961   <2e-16 ***
## pi2         -0.105903   0.011104  -9.538   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Problems with PCA

- Very sensitive to scaling
- Is a good idea to standardize the predictors.