Modern Information Retrieval

Dimensionality reduction and feature selection[1]

Hamid Beigy

Sharif university of technology

December 8, 2023

[1] Some slides have been adapted from slides of Manning, Yannakoudakis, and Schütze.

Table of contents

Introduction

Introduction

The complexity of any classifier depends on the number of input variables or features. These complexities include

1. Time complexity: In most learning algorithms, the time complexity depends on the number of input dimensions(D) as well as on the size of training set (N). Decreasing D decreases the time complexity of algorithm for both training and testing phases.
2. Space complexity: Decreasing D also decreases the memory amount needed for training and testing phases.
3. Samples complexity: Usually the number of training examples (N) is a function of length of feature vectors (D). Hence, decreasing the number of features also decreases the number of training examples.
   Usually the number of training pattern must be 10 to 20 times of the number of features.

Introduction

1. In text classification, we usually represent documents in a high-dimensional space, with each dimension corresponding to a term.
2. In this lecture: axis = dimension = word = term = feature
3. Many dimensions correspond to rare words.
4. Rare words can mislead the classifier.
5. Rare misleading features are called noise features.
6. Eliminating noise features from the representation increases efficiency and effectiveness of text classification.
7. Eliminating features is called feature selection.

Introduction(example)

1. Let's say we're doing text classification for the class *China*.
2. Suppose a rare term, say ARACHNOCENTRIC, has no information about *China*.
3. But all instances of ARACHNOCENTRIC happen to occur in *China* documents in our training set.
4. Then we may learn a classifier that incorrectly interprets ARACHNOCENTRIC as evidence for the class *China*.
5. Such an incorrect generalization from an accidental property of the training set is called over-fitting.
6. Feature selection reduces over-fitting and improves the accuracy of the classifier.

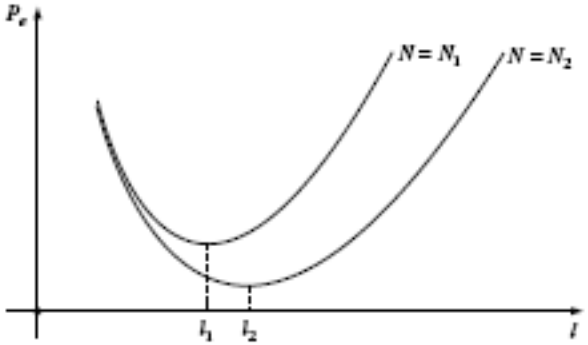
Introduction

There are several reasons why we are interested in reducing dimensionality as a separate preprocessing step.

1. Decreasing the time complexity of classifiers or regressors.
2. Decreasing the cost of extracting/producing unnecessary features.
3. Simpler models are more robust on small data sets. Simpler models have less variance and thus are less depending on noise and outliers.
4. Description of classifier is simpler / shorter.
5. Visualization of data is simpler.

Peaking phenomenon

1. In practice, for a finite N, by increasing the number of features we obtain an initial improvement in performance, but after a critical value further increase of the number of features results in an increase of the probability of error.
2. This phenomenon is also known as the peaking phenomenon.

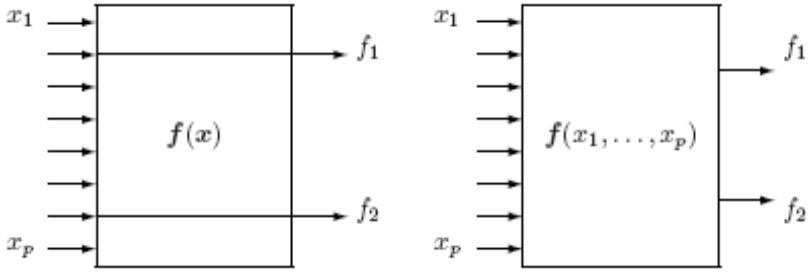


3. If the number of samples increases ($N_2 \gg N_1$), the peaking phenomenon occurs for larger number of features ($l_2 > l_1$).

Dimensionality reduction methods

Dimensionality reduction methods

1. There are two main methods for reducing the dimensionality of inputs
   - Feature selection: These methods select d ($d < D$) dimensions out of D dimensions and $D - d$ other dimensions are discarded.
   - Feature extraction: Find a new set of d ($d < D$) dimensions that are combinations of the original dimensions.



Feature selection methods

Feature selection methods

1. Feature selection methods can be categorized into three categories.
   - Filter methods: These methods use the statistical properties of features to filter out poorly informative features.
   - Wrapper methods: These methods evaluate the feature subset within classifier/regressor algorithms. These methods are classifier/regressors dependent and have better performance than filter methods.
   - Embedded methods:These methods use the search for the optimal subset into classifier/regression design. These methods are classifier/regressors dependent.
2. Two key steps in feature selection process.
   - Evaluation: An evaluation measure is a means of assessing a candidate feature subset.
   - Subset generation: A subset generation method is a means of generating a subset for evaluation.

Basic feature selection algorithm (filter methods)

1. The filter methods has the following structure

$\textsc{SelectFeatures}(\mathbb{D}, c, k)$
1 $V \leftarrow \textsc{ExtractVocabulary}(\mathbb{D})$
2 $L \leftarrow []$
3 **for each** $t \in V$
4 **do** $A(t, c) \leftarrow \textsc{ComputeFeatureUtility}(\mathbb{D}, t, c)$
5 $\quad \textsc{Append}(L, \langle A(t, c), t \rangle)$
6 **return** $\textsc{FeaturesWithLargestValues}(L, k)$

2. How do we compute A, the feature utility?

Different filter methods

1. A feature selection method is mainly defined by the feature utility measure it employs
2. Feature utility measures:
   - Frequency – select the most frequent terms
   - Mutual information – select the terms with the highest mutual information
   - Mutual information is also called information gain in this context.
   - Chi-square (see book)

Mutual information

1. In probability theory and information theory, the mutual information (MI) of two random variables is a measure of the mutual dependence between the two variables.
2. MI determines how similar the joint distribution $p(x, y)$ is to the products of factored marginal distribution $p(x)$ and $p(y)$.
3. Formally, the mutual information of two discrete random variables x and y can be defined as
$$MI(x, y) = \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right)$$
4. In the case of continuous random variables, the summation is replaced by a definite double integral
$$MI(x, y) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p(x, y) \log \left(\frac{p(x, y)}{p(x)p(y)} \right) dxdy$$

Mutual information

1. Compute the feature utility $A(t, c)$ as the mutual information (MI) of term t and class c.
2. MI tells us "how much information" the term contains about the class and vice versa.
3. For example, if a term's occurrence is independent of the class (same proportion of docs within/without class contain the term), then MI is 0.
4. Definition:

$$I(U; C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{P(U = e_t, C = e_c)}{P(U = e_t) P(C = e_c)}$$

How to compute MI values

1. Based on maximum likelihood estimates, formula we actually use is

$$I(U;C) = \frac{N_{11}}{N}\log_2\frac{NN_{11}}{N_{1.}N_{.1}} + \frac{N_{01}}{N}\log_2\frac{NN_{01}}{N_{0.}N_{.1}} + \frac{N_{10}}{N}\log_2\frac{NN_{10}}{N_{1.}N_{.0}} + \frac{N_{00}}{N}\log_2\frac{NN_{00}}{N_{0.}N_{.0}}$$

2. N_{10}: number of documents that contain t ($e_t = 1$) and are not in c ($e_c = 0$);
3. N_{11}: number of documents that contain t ($e_t = 1$) and are in c ($e_c = 1$);
4. N_{01}: number of documents that do not contain t ($e_t = 1$) and are in c ($e_c = 1$);
5. N_{00}: number of documents that do not contain t ($e_t = 1$) and are not in c ($e_c = 1$);
6. $N_{1.} = N_{10} + N_{11}$.
7. $N = N_{00} + N_{01} + N_{10} + N_{11}$.

How to compute MI values

1. Alternative way of computing MI:

$$I(U;C) = \sum_{e_t \in \{1,0\}} \sum_{e_c \in \{1,0\}} P(U = e_t, C = e_c) \log_2 \frac{N(U = e_t, C = e_c)}{E(U = e_t)E(C = e_c)}$$

2. $N(U = e_t, C = e_c)$ is the count of documents with values e_t and e_c .
3. $E(U = e_t, C = e_c)$ is the expected count of documents with values e_t and e_c if we assume that the two random variables are independent.

MI example for poultry/EXPORT in Reuters

	$e_c = e_{poultry} = 1$	$e_c = e_{poultry} = 0$
$e_t = e_{\text{EXPORT}} = 1$	$N_{11} = 49$	$N_{10} = 27{,}652$
$e_t = e_{\text{EXPORT}} = 0$	$N_{01} = 141$	$N_{00} = 774{,}106$

Plug these values into formula:

$$
\begin{aligned}
I(U;C) &= \frac{49}{801{,}948} \log_2 \frac{801{,}948 \cdot 49}{(49+27{,}652)(49+141)} \\
&\quad + \frac{141}{801{,}948} \log_2 \frac{801{,}948 \cdot 141}{(141+774{,}106)(49+141)} \\
&\quad + \frac{27{,}652}{801{,}948} \log_2 \frac{801{,}948 \cdot 27{,}652}{(49+27{,}652)(27{,}652+774{,}106)} \\
&\quad + \frac{774{,}106}{801{,}948} \log_2 \frac{801{,}948 \cdot 774{,}106}{(141+774{,}106)(27{,}652+774{,}106)} \\
&\approx 0.000105
\end{aligned}
$$

MI feature selection on Reuters

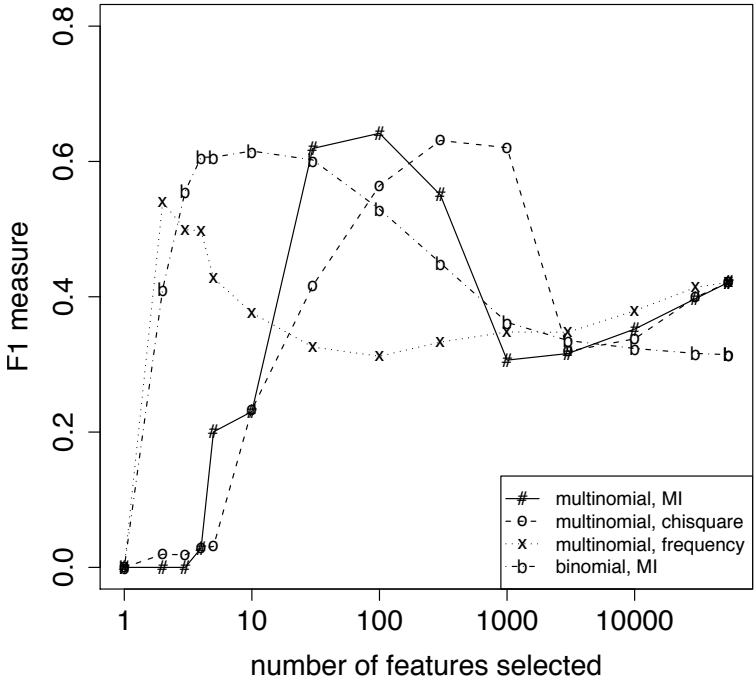
Class: *coffee*

term	MI
COFFEE	0.0111
BAGS	0.0042
GROWERS	0.0025
KG	0.0019
COLOMBIA	0.0018
BRAZIL	0.0016
EXPORT	0.0014
EXPORTERS	0.0013
EXPORTS	0.0013
CROP	0.0012

Class: *sports*

term	MI
SOCCER	0.0681
CUP	0.0515
MATCH	0.0441
MATCHES	0.0408
PLAYED	0.0388
LEAGUE	0.0386
BEAT	0.0301
GAME	0.0299
GAMES	0.0284
TEAM	0.0264

Effect of feature selection (Naive Bayes)



Feature extraction

1. Let S consist of N points over D feature, i.e. it is an $N \times D$ matrix

$$S = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1D} \\ x_{21} & x_{22} & \dots & x_{2D} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{ND} \end{pmatrix}.$$

2. Point $x_i = (x_{i1}, x_{i2}, \dots, x_{iD})^\top$ is a D-dimensional vector spanned by the D basis vectors $e_1, e_2, \dots, e_D$, e_i corresponds to i^{th} feature.
3. The standard basis is an orthonormal basis: the basis vectors are pairwise orthogonal $e_i^\top e_j = 0$, and have unit length $\|e_i\| = 1$.
4. Given any other set of D orthonormal vectors $u_1, u_2, \dots, u_D$,with $u_i^\top u_j = 0$ and $\|u_i\| = 1$ (or $u_i^\top u_i = 1$), we can re-express each point x as the linear combination

$$x = a_1 u_1 + a_2 u_2 + \ldots + a_D u_D.$$

Feature extraction

Principal component analysis

Principal component analysis (Best 1-dimensional approximation)

1. PCA projects $D-$dimensional input vectors to $k-$dimensional input vectors via a linear mapping with minimum loss of information.
2. Dimensions are combinations of the original D dimensions.
3. The problem is to find a matrix W such that the following mapping results in the minimum loss of information.
$$Z = W^\top X$$
4. PCA is unsupervised and tries to maximize the variance.
5. The principle component is w_1 such that the sample after projection onto w_1 is most spread out so that the difference between the sample points becomes most apparent.
6. For uniqueness of the solution, we require $\|w_1\| = 1$,
7. Let $\Sigma = Cov(X)$ and consider the principle component w_1, we have

$$\begin{aligned} z_1 &= w_1^\top x \\ Var(z_1) &= E[(w_1^\top x - w_1^\top \mu)^2] = E[(w_1^\top x - w_1^\top \mu)(w_1^\top x - w_1^\top \mu)^\top] \\ &= E[w_1^\top (x-\mu)(x-\mu)^\top w_1] = w_1^\top E[(x-\mu)(x-\mu)^\top] w_1 = w_1^\top \Sigma w_1 \end{aligned}$$

Principal component analysis (Best 1-dimensional approximation)

1. The mapping problem becomes

$$w_1 = \underset{w}{argmax} \quad w^\top \Sigma w \qquad \text{subject to } w_1^\top w_1 = 1.$$

2. Writing this as Lagrange problem, we have

$$\underset{w_1}{maximize} \quad w_1^\top \Sigma w_1 - \alpha(w_1^\top w_1 - 1)$$

3. Taking derivative with respect to w_1 and setting it equal to 0, we obtain

$$2\Sigma w_1 = 2\alpha w_1 \quad \Rightarrow \quad \Sigma w_1 = \alpha w_1$$

4. Hence w_1 is eigenvector of Σ and α is the corresponding eigenvalue.
5. Since we want to maximize $Var(z_1)$, we have

$$Var(z_1) = w_1^\top \Sigma w_1 = \alpha w_1^\top w_1 = \alpha$$

6. Hence, we choose the eigenvector with the largest eigenvalue, i.e. $\lambda_1 = \alpha$.

1. Let $\epsilon_i = x_i - x_i'$ denote the error vector. The MSE equals to

$$\begin{aligned} MSE(W) &= \frac{1}{N}\sum_{i=1}^{N}\|\epsilon_i\|^2 \\ &= \sum_{i=1}^{N}\frac{\|x_i\|^2}{N} - W^\top \Sigma W \\ &= Var(S) - W^\top \Sigma W. \end{aligned}$$

2. Since $var(S)$, is a constant for a given dataset S, the vector W that minimizes $MSE(W)$ is thus the same one that maximizes the second term,

$$\begin{aligned} MSE(W) &= Var(S) - W^\top \Sigma W \\ &= Var(S) - \lambda_1 \end{aligned}$$

3. Example: Let

$$\Sigma = \begin{pmatrix} 0.681 & -0.039 & 1.265 \\ -0.039 & 0.187 & -0.320 \\ 1.265 & -0.320 & 3.092 \end{pmatrix}$$

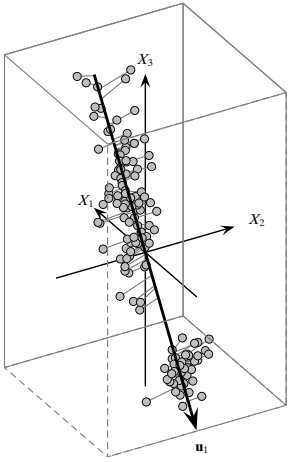
The largest eigenvalue of Σ equals to $\lambda = 3.662$ and the corresponding eigenvector equals to $w_1 = (-0.390, 0.089, -0.916)^\top$

Principal component analysis (Minimum squared error approach)

1. The variance of S equals $var(S) = 0.681 + 0.187 + 3.092 = 3.96$.
2. MSE equals to

$$\begin{aligned} MSE(W_1) &= var(S) - \lambda_1 \\ &= 3.96 - 3.662 = 0.298 \end{aligned}$$

3. Principle component



1. The second principal component, w_2, should also
   - maximize variance
   - be unit length
   - orthogonal to w_1 (z_1 and z_2 must be uncorrelated)
2. The mapping problem for the second principal component becomes

$$w_2 = \underset{w}{argmax} \quad w^\top \Sigma w \qquad \text{subject to } w_2^\top w_2 = 1 \text{ and } w_2^\top w_1 = 0.$$

3. Writing this as Lagrange problem, we have

$$\underset{w_2}{maximize} \quad w_2^\top \Sigma w_2 - \alpha(w_2^\top w_2 - 1) - \beta(w_2^\top w_1 - 0)$$

4. Taking derivative with respect to w_2 and setting it equal to 0, we obtain

$$2\Sigma w_2 - 2\alpha w_2 - \beta w_1 = 0$$

5. Pre-multiply by $w_1^\top$, we obtain

$$2w_1^\top \Sigma w_2 - 2\alpha w_1^\top w_2 - \beta w_1^\top w_1 = 0$$

6. Note that $w_1^\top w_2 = 0$ and $w_1^\top \Sigma w_2 = (w_2^\top \Sigma w_1)^\top = w_2^\top \Sigma w_1$ is a scaler.

1. Since $\Sigma w_1 = \lambda_1 w_1$, therefore we have

$$w_1^\top \Sigma w_2 = w_2^\top \Sigma w_1 = \lambda_1 w_2^\top w_1 = 0$$

2. Then $\beta = 0$ and the problem reduces to

$$\Sigma w_2 = \alpha w_2$$

3. This implies that w_2 should be the eigenvector of Σ with the second largest eigenvalue $\lambda_2 = \alpha$.
4. Let the projected dataset be denoted by A.
5. The total variance for A is given as

$$var(A) = \lambda_1 + \lambda_2$$

Principal component analysis (Best k-dimensional approximation)

1. We are now interested in the best k-dimensional ($k \ll D$) approximation to S.
2. Assume that we have already computed the first $j-1$ principal components or eigenvectors, $w_1, w_2, \ldots, w_{j-1}$, corresponding to the $j-1$ largest eigenvalues of Σ
3. To compute the j^{th} new basis vector w_j, we have to ensure that it is normalized to unit length, that is, $w_j^\top w_j = 1$, and is orthogonal to all previous components w_i (for $i \in [1, j)$).
4. The projected variance along w_j is given as $w_j^\top \Sigma w_j$
5. Combined with the constraints on w_j, this leads to the following maximization problem with Lagrange multipliers:

$$\underset{w_j}{\textit{maximize}} \quad w_j^\top \Sigma w_j - \alpha(w_j^\top w_j - 1) - \sum_{i=1}^{j-1} \beta_i (w_i^\top w_j - 0)$$

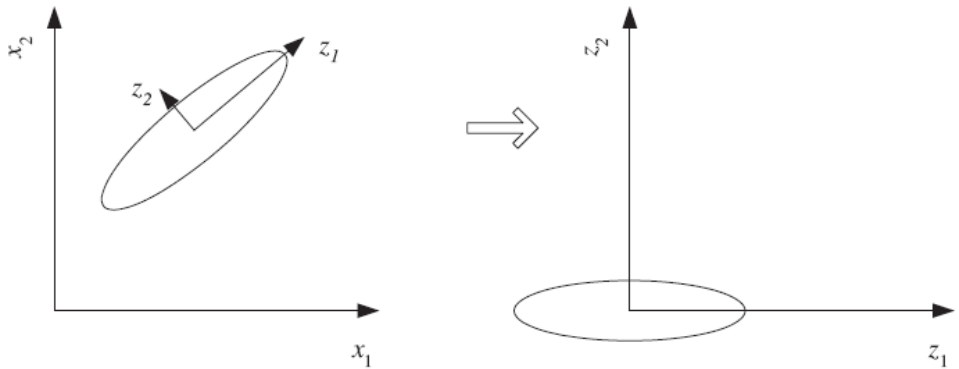
6. Solving this, results in $\beta_i = 0$ for all $i < j$.
7. To maximize the variance along w_j, we use the j^{th} largest eigenvalue of Σ.

Principal component analysis (Best k-dimensional approximation)

1. In summary, to find the best k-dimensional approximation to Σ, we compute the eigenvalues of Σ.
2. Because Σ is positive semidefinite, its eigenvalues must all be non-negative, and we can thus sort them in decreasing order
   $\lambda_1 \geq \lambda_2 \geq \ldots \lambda_{j-1} \geq \lambda_j \geq \ldots \geq \lambda_D \geq 0$
3. We then select the k largest eigenvalues, and their corresponding eigenvectors to form the best k-dimensional approximation.
4. Since Σ is symmetric, for two different eigenvalues, their corresponding eigenvectors are orthogonal. (Show it)
5. If Σ is positive definite ($x^\top \Sigma x > 0$ for all non-null vector x), then all its eigenvalues are positive.
6. If Σ is singular, its rank is k ($k < D$) and $\lambda_i = 0$ for $i = k+1, \ldots, D$.

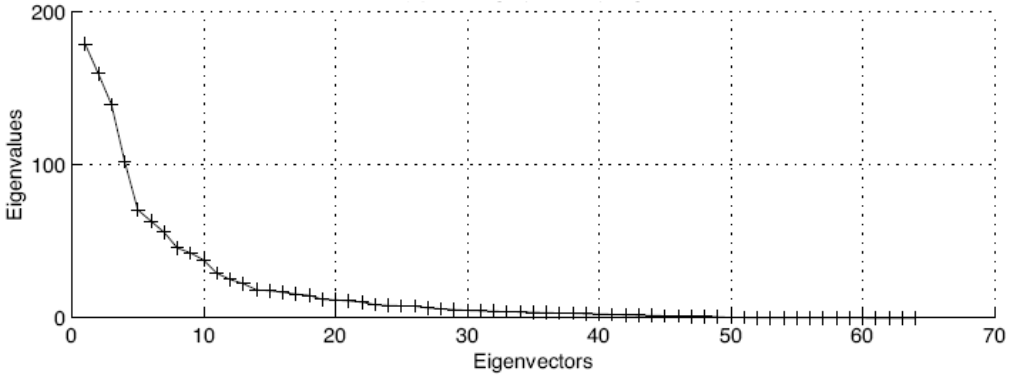
Principal component analysis (effect of centering data)

1. Define
$$Z = W^\top (X - \mathbf{m})$$
2. Then k columns of W are the k leading eigenvectors of S (the estimator of Σ).
3. $\mathbf{m}$ is the sample mean of X.
4. Subtracting $\mathbf{m}$ from X before projection centers the data on the origin.



Principal component analysis (selecting k)

1. How to select k?
2. Since all eigenvalues are positive and $|S| = \prod_{i=1}^{D} \lambda_i$ is small, then some eigenvalues have little contribution to the variance and may be discarded.
3. Scree graph is the plot of variance as a function of the number of eigenvectors.

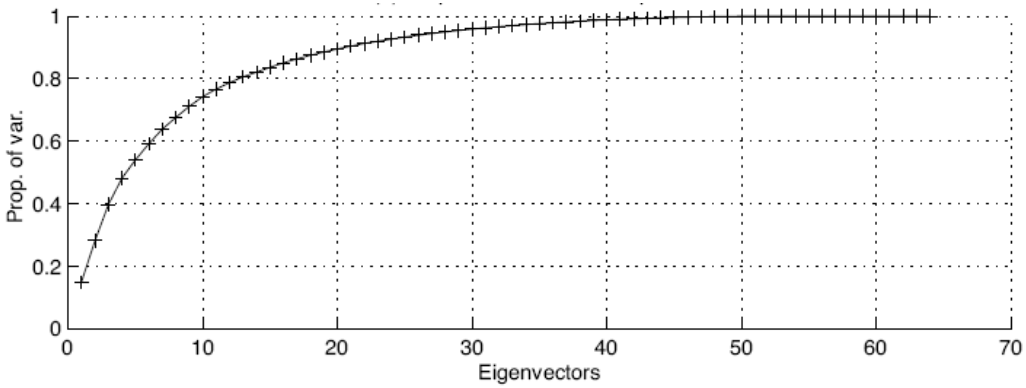


Principal component analysis (selecting k)

1. How to select k?
2. We select the leading k components that explain more than for example 95% of the variance.
3. The proportion of variance (POV) is

$$POV = \frac{\sum_{i=1}^{k} \lambda_i}{\sum_{i=1}^{D} \lambda_i}$$

4. By visually analyzing it, we can choose k.



References

Reading

1. Chapter 13 of Information Retrieval Book[2]

[2] Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

References

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze (2008). *Introduction to Information Retrieval*. New York, NY, USA: Cambridge University Press.

Questions?