Dimensionality reduction techniques: a brief introduction

Douglas S. Gonçalves

MTM - UFSC

Seminários de Aprendizado de Máquina

Novembro, 2016

Section 1

Introduction

Notation and basic concepts

Consider N data points in a high d-dimensional space:

$$X = [x_1 \; x_2 \; \ldots \; x_N] \in \mathbb{R}^{d \times N}$$

Dimensionality reduction problem

Given a target low dimension $K < d$, find a map $T : \mathbb{R}^d \to \mathbb{R}^K$ which minimizes some objective $f_X(.)$.

Low-dimensional transformed data: $\hat{x}_i = T(x_i)$

Notation and basic concepts

Consider N data points in a high d-dimensional space:

$$X = [x_1 \; x_2 \; \ldots \; x_N] \in \mathbb{R}^{d \times N}$$

Dimensionality reduction problem

Given a target low dimension $K < d$, find a map $T : \mathbb{R}^d \to \mathbb{R}^K$ which minimizes some objective $f_X(.)$.

Low-dimensional transformed data: $\hat{x}_i = T(x_i)$

Linear dimensionality reduction problem

Given a target low dimension $K < d$, find a linear map $P \in \mathbb{R}^{K \times d}$ which minimizes some objective $f_X(.)$.

Low-dimensional transformed data: $\hat{x}_i = Px_i$

Notation and basic concepts

Assume that the data are centered: $Xe = 0$, where $e = (1, 1, \ldots, 1)^\top \in \mathbb{R}^N$.

This can be done by

$$X \leftarrow X\left(I - \frac{1}{N}ee^\top\right) := XJ,$$

Notation and basic concepts

Assume that the data are centered: $Xe = 0$, where $e = (1, 1, \ldots, 1)^\top \in \mathbb{R}^N$.

This can be done by

$$X \leftarrow X\left(I - \frac{1}{N}ee^\top\right) := XJ,$$

Sample covariance matrix: $C = \frac{1}{N}XX^\top \in \mathbb{R}^{d\times d}$

Notation and basic concepts

Assume that the data are centered: $Xe = 0$, where $e = (1, 1, \ldots, 1)^\top \in \mathbb{R}^N$.

This can be done by

$$X \leftarrow X\left(I - \frac{1}{N}ee^\top\right) := XJ,$$

Sample covariance matrix: $C = \frac{1}{N}XX^\top \in \mathbb{R}^{d\times d}$

Gram matrix: $Y = X^\top X \in \mathbb{R}^{N\times N}$, matrix of inner products $Y_{ij} = \langle x_i, x_j\rangle$

Notation and basic concepts

Assume that the data are centered: $Xe = 0$, where $e = (1, 1, \ldots, 1)^\top \in \mathbb{R}^N$.

This can be done by

$$X \leftarrow X\left(I - \frac{1}{N}ee^\top\right) := XJ,$$

Sample covariance matrix: $C = \frac{1}{N}XX^\top \in \mathbb{R}^{d\times d}$

Gram matrix: $Y = X^\top X \in \mathbb{R}^{N\times N}$, matrix of inner products $Y_{ij} = \langle x_i, x_j\rangle$

Singular value decomposition: $X = U\Sigma V^\top$, where $\Sigma = \text{diag}(\sigma_1, \sigma_2, \ldots, \sigma_r)$, $r = \text{rank}(X)$, $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{N\times r}$ such that $U^\top U = I$, $V^\top V = I$, and:

$$\begin{aligned} Xv_i &= \sigma_i u_i, \quad i = 1, 2, \ldots, r \\ X^\top u_i &= \sigma_i v_i. \end{aligned}$$

Notation and basic concepts

EDM: Let $D \in \mathbb{R}^{N \times N}$ such that $D_{ij} = d_{ij}^2$. D is said a Euclidean distance matrix (EDM) if there exists $X \in \mathbb{R}^{d \times N}$ such that $\|x_i - x_j\|_2^2 = d_{ij}^2, \forall \{i, j\}$.

Notation and basic concepts

EDM: Let $D \in \mathbb{R}^{N\times N}$ such that $D_{ij} = d_{ij}^2$. D is said a Euclidean distance matrix (EDM) if there exists $X \in \mathbb{R}^{d\times N}$ such that $\|x_i - x_j\|_2^2 = d_{ij}^2, \forall\{i,j\}$.

From the identity $\|x_i - x_j\|^2 = \|x_i\|^2 - 2\langle x_i, x_j\rangle + \|x_j\|^2$:

$$D = \mathcal{K}(Y) := \text{diag}(Y)e^\top + e\text{diag}(Y)^\top - 2Y,$$

(where $Y = X^\top X$). Thus, given an EDM D we obtain

$$Y = \mathcal{K}^\dagger(D) := -\frac{1}{2}JDJ.$$

Notation and basic concepts

EDM: Let $D \in \mathbb{R}^{N \times N}$ such that $D_{ij} = d_{ij}^2$. D is said a Euclidean distance matrix (EDM) if there exists $X \in \mathbb{R}^{d \times N}$ such that $\|x_i - x_j\|_2^2 = d_{ij}^2, \forall \{i, j\}$.

From the identity $\|x_i - x_j\|^2 = \|x_i\|^2 - 2\langle x_i, x_j \rangle + \|x_j\|^2$:

$$D = \mathcal{K}(Y) := \text{diag}(Y)e^\top + e\text{diag}(Y)^\top - 2Y,$$

(where $Y = X^\top X$). Thus, given an EDM D we obtain

$$Y = \mathcal{K}^\dagger(D) := -\frac{1}{2} JDJ.$$

Theorem [Schoenberg, 1935]

D is EDM iff $\mathcal{K}^\dagger(D)$ is positive semidefinite. Moreover, the embedding dimension is given by $\text{rank}(\mathcal{K}^\dagger(D))$.

Section 2

Linear Dimensionality Reduction

PCA

Principal Component Analysis: find $P \in \mathbb{R}^{K \times d}$ such that the projected data $\hat{X} = PX$ preserves the variance in the original data X as much as possible; equivalently (minimizing reconstruction error):

$$\begin{aligned} \min_{M} \quad & \|X - MM^\top X\|_F^2 := f_X(M) \\ \text{s.t} \quad & M \in \mathbb{R}^{d \times K}, \quad M^\top M = I, \end{aligned}$$

where $P = M^\top$ (ie, $\hat{x}_i = M^\top x_i$).

[the objective is equivalent to min. $-\text{tr}(M^\top X X^\top M)$].

PCA

Principal Component Analysis: find $P \in \mathbb{R}^{K \times d}$ such that the projected data $\hat{X} = PX$ preserves the variance in the original data X as much as possible; equivalently (minimizing reconstruction error):

$$\begin{aligned} \min_{M} \quad & \|X - MM^\top X\|_F^2 := f_X(M) \\ \text{s.t} \quad & M \in \mathbb{R}^{d \times K}, \quad M^\top M = I, \end{aligned}$$

where $P = M^\top$ (ie, $\hat{x}_i = M^\top x_i$).

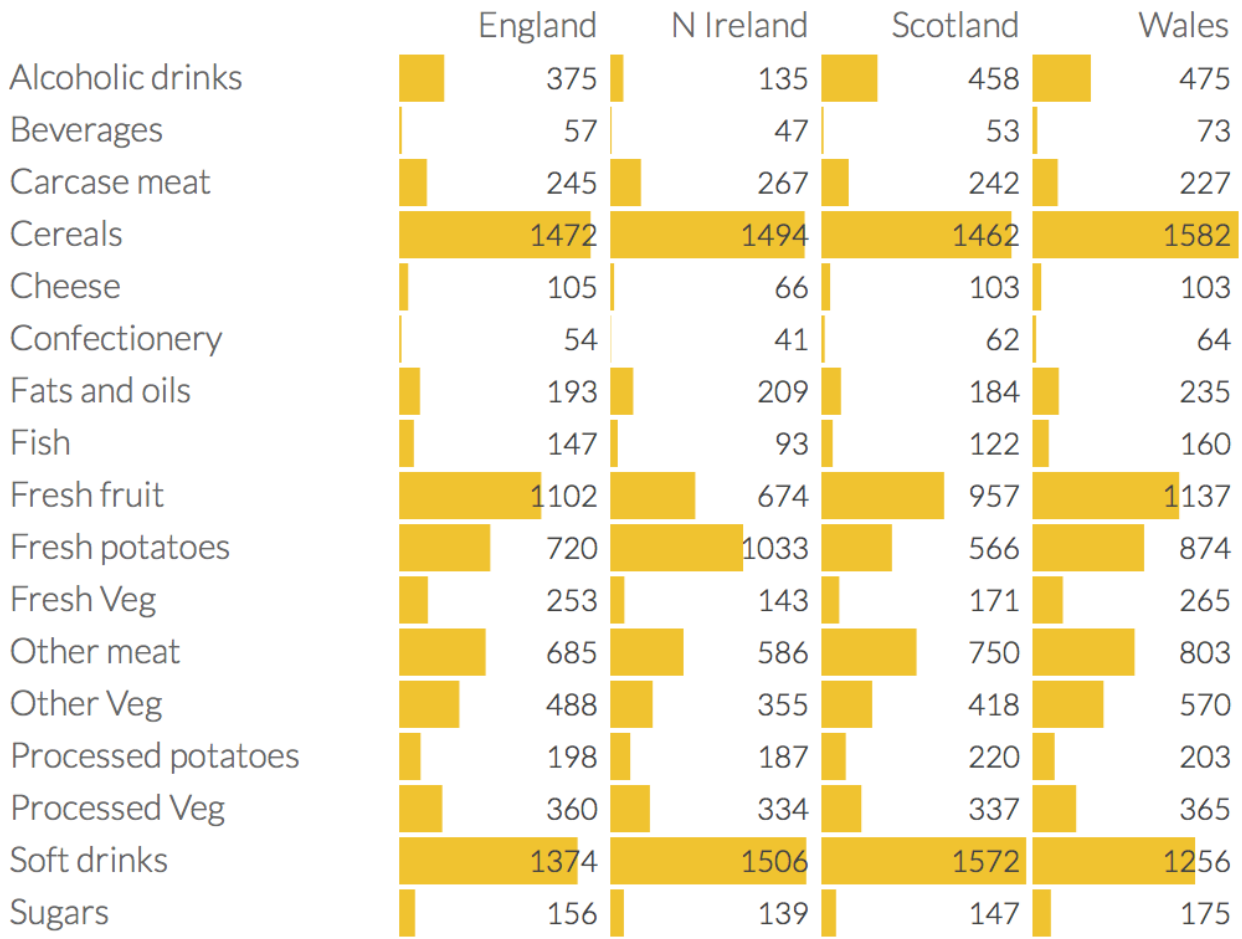
[the objective is equivalent to min. $-\text{tr}(M^\top XX^\top M)$].

From the reduced SVD of X: $XX^\top = U\Sigma^2 U^\top = U\Lambda U^\top$

It turns out that $M = \hat{U}$ is the solution, where $\hat{U} \in \mathbb{R}^{d \times K}$ with the columns of U corresponding to the K largest eigenvalues of $XX^\top$.

Thus $P = \hat{U}^\top$ and $\hat{X} = \hat{U}^\top X$.

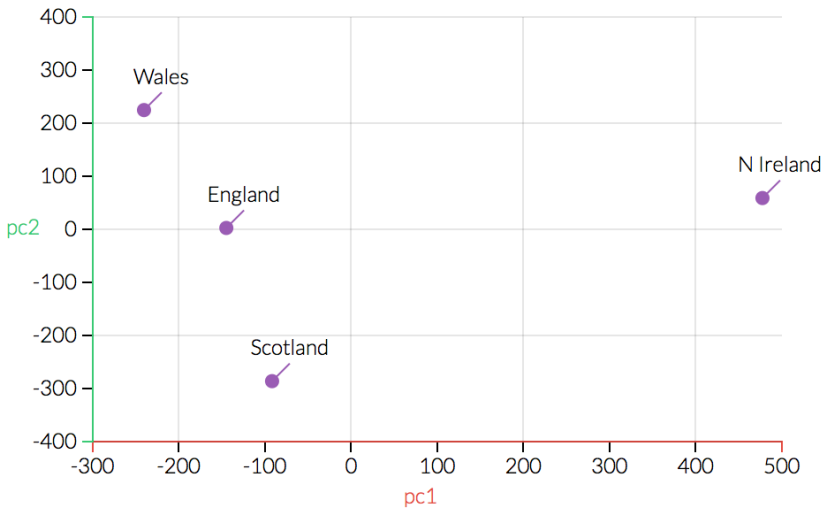
Simple example



	England	N Ireland	Scotland	Wales
Alcoholic drinks	375	135	458	475
Beverages	57	47	53	73
Carcase meat	245	267	242	227
Cereals	1472	1494	1462	1582
Cheese	105	66	103	103
Confectionery	54	41	62	64
Fats and oils	193	209	184	235
Fish	147	93	122	160
Fresh fruit	1102	674	957	1137
Fresh potatoes	720	1033	566	874
Fresh Veg	253	143	171	265
Other meat	685	586	750	803
Other Veg	488	355	418	570
Processed potatoes	198	187	220	203
Processed Veg	360	334	337	365
Soft drinks	1374	1506	1572	1256
Sugars	156	139	147	175

`http://setosa.io/ev/principal-component-analysis/`

$$\bar{\Sigma}\bar{V}^\top = \begin{bmatrix} -144.9932 & 477.3916 & -91.8693 & -240.5291 \\ 2.5330 & 58.9019 & -286.0818 & 224.6469 \\ 105.7689 & -4.8779 & -44.4155 & -56.4756 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$



http://setosa.io/ev/principal-component-analysis/

$$\begin{aligned} \sigma_1 &= 561.44, \\ \sigma_2 &= 368.49, \\ \sigma_3 &= 127.95, \\ \sigma_4 &= 0. \end{aligned}$$

$$\hat{X} = \begin{bmatrix} -144.99 & 477.39 & -91.86 & -240.52 \\ 2.5330 & 58.90 & -286.08 & 224.64 \end{bmatrix}$$

MDS

Classical Multidimensional Scaling

Given the dissimilarities(distances) between data points in d-dimensional space, find a representation $\hat{X}$ in a K low dimensional space such that the pairwise distances are preserved.

$$\min_{\hat{X} \in \mathbb{R}^{K \times N}} \quad \sum_{i,j} \left(\delta_{ij} - d(\hat{x}_i, \hat{x}_j)\right)^2$$

MDS

Classical Multidimensional Scaling

Given the dissimilarities(distances) between data points in d-dimensional space, find a representation $\hat{X}$ in a K low dimensional space such that the pairwise distances are preserved.

$$\min_{\hat{X} \in \mathbb{R}^{K \times N}} \quad \sum_{i,j} (\delta_{ij} - d(\hat{x}_i, \hat{x}_j))^2$$

If the data points are available: $Y = X^\top X = V \Sigma^2 V^\top = V \Lambda V^\top = V \sqrt{\Lambda} \sqrt{\Lambda} V^\top$.

Thus,

$$X_d = \sqrt{\bar{\Lambda}} \bar{V}^\top \in \mathbb{R}^{d \times N},$$
$$X_r = \sqrt{\Lambda} V^\top \in \mathbb{R}^{r \times N},$$

and the solution of the MDS is given by

$$\hat{X} = \Sigma(1:K, 1:K) V(:, 1:K)^\top$$

(which is precisely $\hat{U}^\top X$).

MDS

In fact, MDS can be stated as

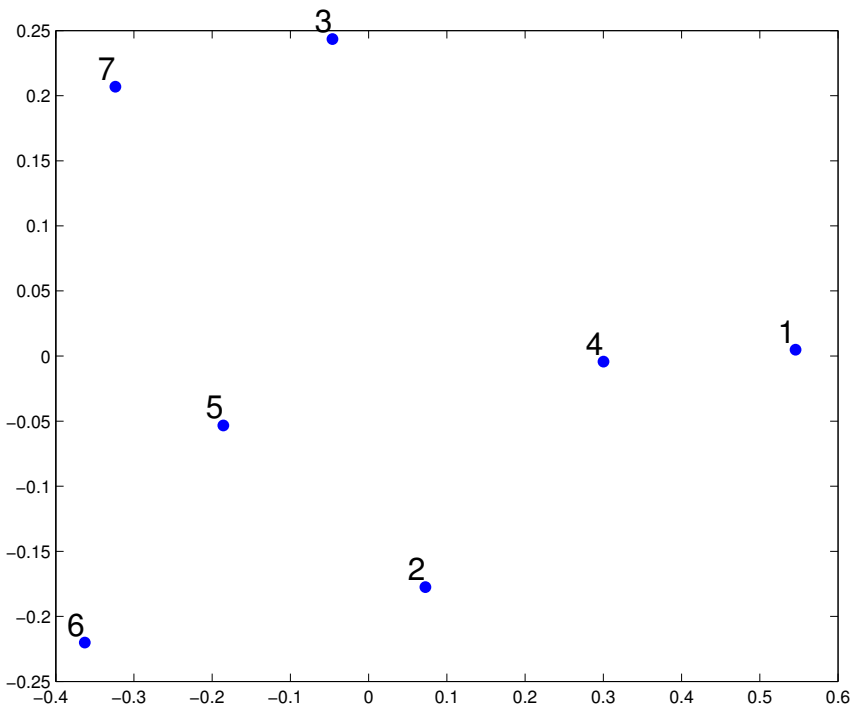
$$\begin{aligned} \min_{M} \quad & -\sum_{i,j} \|M^\top x_i - M^\top x_j\|_2^2 \\ \text{s.t} \quad & M \in \mathbb{R}^{d\times K}, \quad M^\top M = I. \end{aligned}$$

The above objective is equivalent to maximize $\sum_{i,j} \|\hat{x}_i - \hat{x}_j\|_2^2$, i.e., maximize the scatter of the projected points (which in its turn is the same of maximizing $\text{tr}(M^\top X X^\top M)$).

Example

		1	2	3	4	5	6	7
murder	1	1.00	0.52	0.34	0.81	0.28	0.06	0.11
rape	2	0.52	1.00	0.55	0.70	0.68	0.60	0.44
robbery	3	0.34	0.55	1.00	0.56	0.62	0.44	0.62
assault	4	0.81	0.70	0.56	1.00	0.52	0.32	0.33
burglary	5	0.28	0.68	0.62	0.52	1.00	0.80	0.70
larceny	6	0.06	0.60	0.44	0.32	0.80	1.00	0.55
car theft	7	0.11	0.44	0.62	0.33	0.70	0.55	1.00

distance from correlation: $d_{ij} = 1 - \rho_{ij}$



Section 3

Nonlinear data structures

Isomap

Main idea: preserve the intrinsic geometry of data as captured in the geodesic manifold distances d_M. [but we only have input-space distances d_X]

1. **Construct neighborhood graph:** using $d_X(i,j)$ construct a graph G, with and edge $\{i,j\}$ if i and j are closer than ϵ (or if i is one of the k-NN of j). Assign $d_G(i,j) = d_X(i,j)$ for those pairs.

2. **Compute shortest path distances:** apply Floyd's algorithm (shortest paths) to complete a distance matrix D ($D_{ij} = d_G(i,j)^2$) by setting $d_G(i,j)$ as the shortest path between i and j.

3. **Obtain a K-dimensional embedding by using MDS:**

$$Y = \mathcal{K}^{\dagger}(D) = V \Lambda V^{\top},$$

$$\hat{X} = \sqrt{\Lambda_{+}(1:K, 1:K)} V(:, 1:K)^{\top}.$$

[Tenenbaum et al., 2000]

Laplacian eigenmaps

1. Construct a weighted graph $G(V, E, W)$ with N nodes, set of edges E connecting neighboring points.

2. Build the Laplacian of such graph:

$$L = D - W,$$

where D is the degree matrix ($D_{ii} = \sum_j W_{ij}$) and W is a weight matrix ($W_{ij} > 0, \forall \{i, j\} \in E,\ W_{ij} = 0, \forall \{i, j\} \notin E$).

3. Solve the generalized eigenvalue problem:

$$Lv_i = \lambda_i D v_i,$$

set $V_K = [v_1\, v_2 \ldots v_K]$ corresponding to the K smallest eigenvalues (excluding the null one) and retrieve $\hat{X} = V_K^\top$.

[Belkin, Niyogi, 2003]

Laplacian eigenmaps

(underlying idea)

$$\begin{aligned} \min_{\hat{X}\in\mathbb{R}^{K\times N}} \quad & \sum_i \sum_j W_{ij}\|\hat{x}_i - \hat{x}_j\|^2 \\ \text{s.t} \quad & \hat{X}D\hat{X}^\top = I, \end{aligned} \tag{P}$$

where the constraint prevents the points from collapsing in a subspace with dimension less than $K-1$.

Using $L = D - W$ and $D_{ii} = \sum_j W_{ij}$, one can show that

$$\sum_i \sum_j W_{ij}\|\hat{x}_i - \hat{x}_j\|^2 = 2\text{tr}(\hat{X}L\hat{X}^\top),$$

and from optimality conditions for (P), it follows that the solution is provided by the matrix of eigenvectors corresponding to the K smallest eigenvalues of the generalized eigenvalue problem:

$$Lv = \lambda Dv.$$

Section 4

Random projections

Johnson-Lindenstrauss lemma

Lemma

Given $\varepsilon \in (0,1)$, and n points in $\mathbb{R}^d$, if $K = O\left(\frac{1}{\varepsilon^2}\log n\right)$, then there exists a map $P : \mathbb{R}^d \to \mathbb{R}^K$ such that

$$(1-\varepsilon)\left\|x_i - x_j\right\|^2 \leq \left\|Px_i - Px_j\right\|^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|^2, \quad \forall i, j.$$

Johnson-Lindenstrauss lemma

Lemma

Given $\varepsilon \in (0,1)$, and n points in $\mathbb{R}^d$, if $K = O\left(\frac{1}{\varepsilon^2}\log n\right)$, then there exists a map $P : \mathbb{R}^d \to \mathbb{R}^K$ such that

$$(1-\varepsilon)\left\|x_i - x_j\right\|^2 \leq \left\|Px_i - Px_j\right\|^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|^2, \quad \forall i,j.$$

Proof. (Sketch)

- Let $v \in \mathbb{R}^d$, such that $\|v\| = 1$. If $P \in \mathbb{R}^{K\times d}$ is a suitable random matrix then
$$\Pr\left[\left|\|Pv\|^2 - 1\right| > \varepsilon\right] < e^{-CK\varepsilon^2}.$$
- Applying this to $v = (x_i - x_j)/\|x_i - x_j\|$ for a fixed pair i,j yields
$$\Pr\left[\left|\|P(x_i - x_j)\|^2/\|x_i - x_j\|^2 - 1\right| > \varepsilon\right] < \frac{1}{n^2}$$
- taking the union bound over $\binom{n}{2}$ pairs:
$$\Pr\left[(1-\varepsilon)\left\|x_i - x_j\right\|^2 \leq \left\|Px_i - Px_j\right\|^2 \leq (1+\varepsilon)\left\|x_i - x_j\right\|^2\right] > 1/2.$$

Random projections

- P as an orthogonal projector [Johnson, Lidenstrauss, 82]: pick k random (i.i.d) orthonormal vectors in $\mathbb{R}^d$

- Gaussian random projector [Indyk, Motwani, 98], [Dasgupta, Gupta, 2003]:

$$P = \frac{1}{\sqrt{K}} R, \quad \text{where } R_{ij} \sim \mathcal{N}(0,1) \text{ i.i.d.}$$

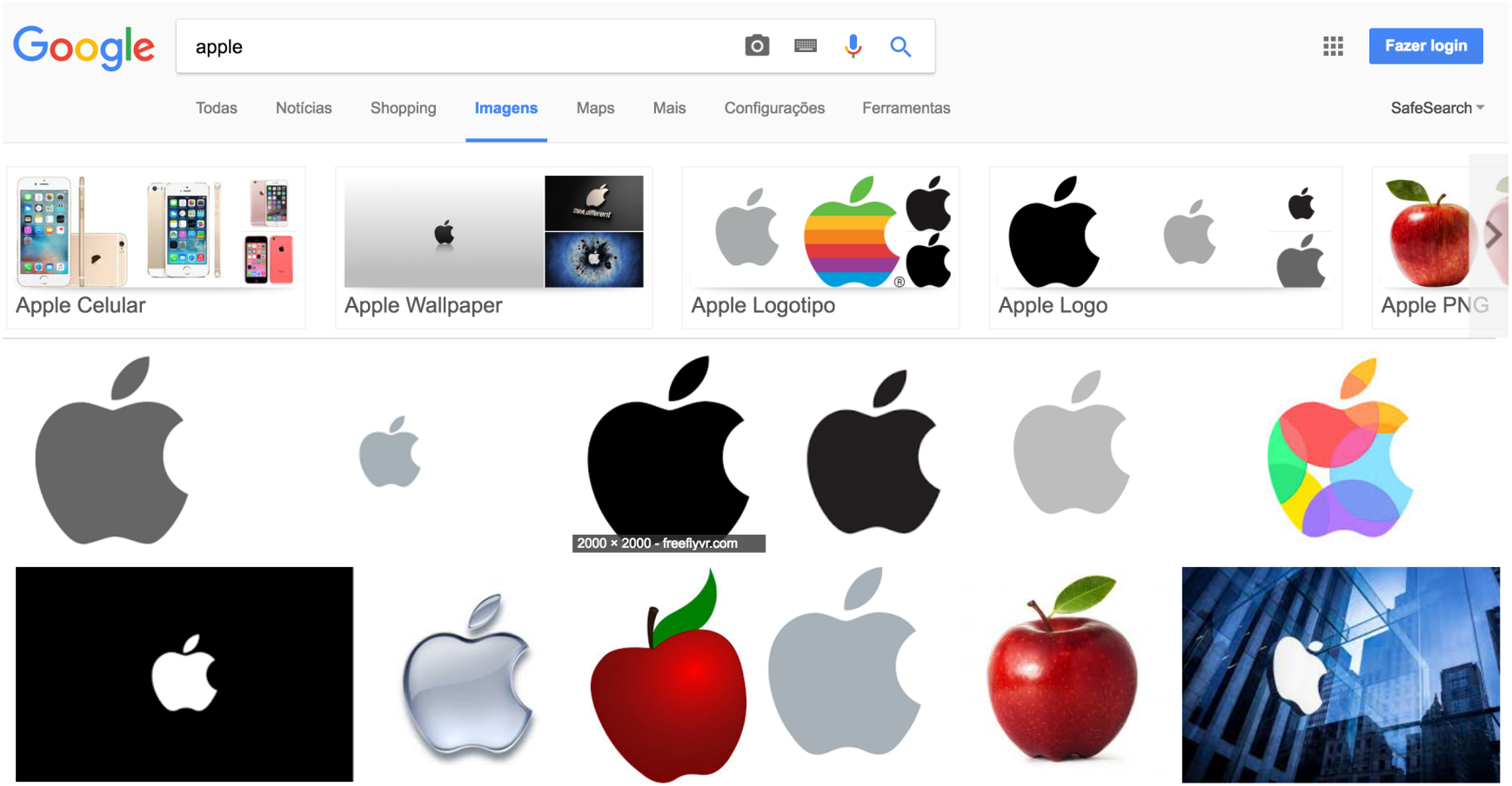
  Orthogonality: high dimension $\rightarrow$ high probability
  Normalization: exponential concentration bound

- Other practical choices:

  A very simple one: $R_{ij} = \pm 1$ with probability $1/2$.

  [Achlioptas, 2003]: $R_{ij} = \sqrt{3} \times \begin{cases} -1, & \text{w.p. } 1/6 \\ 0, & \text{w.p. } 2/3 \\ 1, & \text{w.p. } 1/6. \end{cases}$

Last example: Google images query



Clustering ...

Each image was rescaled to 200×200 pixels and, by using RGB scheme, represented by a vector $x \in \mathbb{R}^{120000}$.

- Clustering by k-means: ($k = 3$, runs=100)

```
>> tic; Vx = kmeans(X, 3, 100); toc
Elapsed time is 6.936064 seconds.
```

v = [1, 1, 2, 2, 2, 2, 3, 3]

- Preprocess by Random Projection ($\varepsilon = 0.1$, $K \approx 500$), then k-means:

```
>> tic; Vy = kmeans(Y, 3, 100); toc
Elapsed time is 0.058027 seconds.
```

v = [1, 1, 2, 2, 2, 2, 3, 3]

References

- J.P. Cunningham and Z. Ghahramani, Linear Dimensionality Reduction: Survey, Insights, and Generalizations, *Journal of Machine Learning Research* 16, 2859–2900, 2015.

- I. Dokmanic, R. Parhizkar, J. Ranieri, M. Vetterli, Euclidean distance matrices: Essential theory, algorithms, and applications, *IEEE Signal Processing* 32, 12–30, 2015.

- S. Dasgupta and A. Gupta, An Elementary Proof of a Theorem of Johnson and Lindenstrauss, *Random structures and algorithms* 22(1), 60–65, 2003.

- J. B. Tenenbaum, V. de Silva, J. C. Langford, A Global Geometric Framework for Nonlinear Dimensionality Reduction, *Science* 290, 2319–2323, 2000.

Obrigado pela atenção!!

douglas@mtm.ufsc.br