Machine Learning 4: Dimensionality Reduction

Master 2 Computer Science

Aurélien Garivier

2019-2020

Table of contents

1. Dimension reduction: PCA

2. Dimension reduction: random projections

Dimensionality reduction

- Data: $X = \begin{pmatrix} x_1^T \\ \vdots \\ x_n^T \end{pmatrix} \in \mathcal{M}_{n,p}(\mathbb{R})$, $p \gg 1$.
- Dimensionality reduction: replace x_i with $y_i = Wx_i$, where $W \in \mathcal{M}_{d,p}(\mathbb{R})$, $d \ll p$.
- Hopefully, we do not loose too much by replacing x_i by y_i.
  2 approaches:
  - Quasi-invertibility: there exists a recovering matrix $U \in \mathcal{M}_{p,d}(\mathbb{R})$ such that for all $i \in \{1, \dots, n\}$,

    $$\tilde{x}_i = Uy_i \approx x_i \ .$$

  - More modest goal: distance-preserving property

    $$\forall 1 \leq i, j \leq n, \quad \|y_i - y_j\| \approx \|x_i - x_j\|$$

Dimension reduction: PCA

PCA

PCA aims at finding the compression matrix W and the recovering matrix U such that the total squared distance between the original and the recovered vectors is minimal:

$$\underset{W\in\mathcal{M}_{d,p}(\mathbb{R}),U\in\mathcal{M}_{p,d}(\mathbb{R})}{\arg\min} \sum_{i=1}^{n} \left\| x_i - UWx_i \right\|^2 .$$

Property. A solution (W, U) is such that $U^T U = I_d$ and $W = U^T$.

Proof. Let $W \in \mathcal{M}_{d,p}(\mathbb{R})$, $U \in \mathcal{M}_{p,d}(\mathbb{R})$, and let $R = \left\{ UWx : x \in \mathbb{R}^p \right\}$. $\dim(R) \leq d$, and we can assume that $\dim(R) = d$. Let $V = \begin{pmatrix} v_1 & \Big| & \dots & \Big| & v_d \end{pmatrix} \in \mathcal{M}_{p,d}(\mathbb{R})$ be an orthogonal basis of R, hence $V^T V = I_d$ and for every $\tilde{x} \in R$ there exists $y \in \mathbb{R}^d$ such that $\tilde{x} = Vy$. But for every $x \in \mathbb{R}^p$,

$$\underset{\tilde{x}\in R}{\arg\min} \|x - \tilde{x}\|^2 = V.\underset{y\in\mathbb{R}^d}{\arg\min} \|x - Vy\|^2 = V.\underset{y\in\mathbb{R}^d}{\arg\min} \|x\| + \|y\|^2 - 2y^T\left(V^T x\right) = VV^T x$$

(as can be seen easily by differentiation in y), and hence

$$\sum_{i=1}^{n} \left\| x_i - UWx_i \right\|^2 \geq \sum_{i=1}^{n} \left\| x_i - VV^T x_i \right\|^2 .$$

The PCA solution

Corollary: the optimization problem can be rewritten

$$\underset{U \in \mathcal{M}_{p,d}(\mathbb{R}) : U^T U = I_d}{\arg\min} \sum_{i=1}^{n} \left\| x_i - U U^T x_i \right\|^2 .$$

Since $\left\| x_i - U U^T x_i \right\|^2 = \| x_i \|^2 - \operatorname{Tr}\left(U^T x_i x_i^T U \right)$, this is equivalent to

$$\underset{U \in U \in \mathcal{M}_{p,d}(\mathbb{R}) : U^T U = I_d}{\arg\max} \operatorname{Tr}\left(U^T \sum_{i=1}^{n} x_i x_i^T U \right) .$$

Let $A = \sum_{i=1}^{n} x_i x_i^T$, and let $A = V D V^T$ be its spectral decomposition: D is diagonal, with $D_{1,1} \geq \cdots \geq D_{p,p} \geq 0$ and $V^T V = V V^T = I_p$.

Solving PCA by SVD

Theorem Let $A = \sum_{i=1}^{n} x_i x_i^T$, and let $u_1, \ldots, u_d$ be the eigenvectors of A corresponding to the d largest eigenvalues of A. Then the solution to the PCA optimization problem is $U = \left(\begin{array}{c|c|c} u_1 & \ldots & u_d \end{array} \right)$, and $W = U^T$.

Proof. Let $U \in \mathcal{M}_{p,d}(\mathbb{R})$ be such that $U^T U = I_d$, and let $B = V^T U$. Then $VB = U$, and $U^T A U = B^T V^T V D V^T V B = B^T D B$, hence

$$\mathrm{Tr}\left(U^T A U\right) = \sum_{j=1}^{p} D_{j,j} \sum_{i=1}^{d} B_{j,i}^2 \ .$$

Since $B^T B = U^T V V^T U = I_d$, the columns of B are orthonormal and $\sum_{j=1}^{p} \sum_{i=1}^{d} B_{j,i}^2 = d$.

In addition, completing the columns of B to an orthonormal basis of $\mathbb{R}^p$ one gets $\tilde{B}$ such that $\tilde{B}^T \tilde{B} = I_p$, and for every j one has $\sum_{i=1}^{p} \tilde{B}_{j,i}^2 = 1$, hence $\sum_{i=1}^{d} B_{j,i}^2 \leq 1$.

Thus,

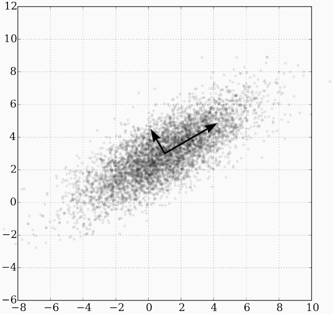
$$\mathrm{Tr}\left(U^T A U\right) \leq \max_{\beta \in [0,1]^p : \|\beta\|_1 \leq d} \sum_{j=1}^{p} D_{j,j} \beta_j = \sum_{j=1}^{d} D_{j,j} \ ,$$

which can be reached if U is made of the d leading eigenvectors of A.

PCA: comments

Interpretation: PCA aims at maximizing the projected variance.

Often, the quality of the result is measured by the proportion of the variance explained by the d principal components: $\frac{\sum_{i=1}^{d} D_{i,i}}{\sum_{i=1}^{p} D_{i,i}}$.



[Src: wikipedia.org]

In practice: if $p \geq n$, it is cheaper to diagonalize $B = XX^T \in \mathcal{M}_n(\mathbb{R})$, since if u is such that $Bu = \lambda u$ then for $v = X^T u / \|X^T u\|$ one has $Av = \lambda v$.

This remark is also at the basis of *kernel PCA*.

Computing the PCA: iteration method

Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be the eigenvalues of A, and let v be such that $\|v\| = 1$ and $Av = \lambda_1 v$. Goal: approximate v.

Algorithm: $u_0 = \left[\frac{\epsilon_1}{\sqrt{n}}, \ldots, \frac{\epsilon_n}{\sqrt{n}}\right]$ where $\epsilon_i \overset{iid}{\sim} \mathcal{U}\big(\{-1, 1\}\big)$, then $\|u_0\|^2 = 1$.
$u_{k+1} = \frac{Au_k}{\left\|Au_k\right\|}$.

Theorem

With probability at least 3/16,

$$\left|\langle u_t, v\rangle\right| \geq 1 - 2n\left(\frac{\lambda_2}{\lambda_1}\right)^{2t} .$$

Thus, it takes at most $t = \frac{\log \frac{2n}{\epsilon}}{2\log\frac{\lambda_1}{\lambda_2}}$ iterations to ensure that $\left|\langle u_t, v\rangle\right| \geq 1 - \epsilon$ with probability at least 3/16.

Remark: one can similarly show that with non-vanishing probability $\langle u_t, Au_t\rangle \geq \lambda_1 \times \frac{1-\epsilon}{1+4n(1-\epsilon)^{2t}}$. `http://theory.stanford.edu/~trevisan/expander-online/lecture03.pdf`.

Complexity of the iteration method 1/2

Observe that $\langle u_0, v\rangle$ has expectation 0 and variance $\sum_{i=1}^n (v_i)^2/n = 1/n$. Hence, $Z = \langle u_0, v\rangle^2$ has expectation $1/n$ and

$$n^2\,\mathbb{E}\big[Z^2\big] = \mathbb{E}\left[\sum_{1\le i,j,k,l\le d} \epsilon_i\epsilon_j\epsilon_k\epsilon_l\right] = \sum_{1\le j\le d}(v_j)^4 + 6\sum_{1\le j<k\le d}(v_j)^2(v_k)^2$$
$$= 3\left(\|v\|^2\right)^2 - 2\sum_{1\le j\le d}(v_j)^4 \le 3\,.$$

By the Cauchy-Schwartz inequality, for every $\delta\in(0,1)$

$$\mathbb{E}[Z] = \mathbb{E}\big[Z\mathbb{1}\{Z<\delta\mathbb{E}[Z]\}\big]+\mathbb{E}\big[Z\mathbb{1}\{Z\ge\delta\mathbb{E}[Z]\}\big] \le \delta\mathbb{E}[Z]+\sqrt{\mathbb{E}\big[Z^2\big]\mathbb{P}\big(Z\ge\delta\mathbb{E}[Z]\big)}\,.$$

and hence, for $\delta = 1/4$:

$$\mathbb{P}\big(Z\ge\delta\mathbb{E}[Z]\big) \ge (1-\delta)^2\frac{\mathbb{E}[Z]^2}{\mathbb{E}[Z^2]} \ge \left(\frac{3}{4}\right)^2\frac{1/n^2}{3/n^2} = \frac{9}{16}\times\frac{1}{3} \ge \frac{3}{16}\,.$$

Complexity of the iteration method 2/2

But since, if $v^1 = v$ and $\forall i \in \{2, \dots, n\}, \|v^i\| = 1$ and $Av^i = \lambda_i v^i$:

$$u_t = \frac{A^t u_0}{\|A^t u_0\|} = \frac{\sum_{i=1}^n \lambda_i^t \langle u_0, v^i \rangle v^i}{\sqrt{\sum_{i=1}^n \left(\lambda_i^t \langle u_0, v^i \rangle\right)^2}} \ ,$$

whenever $\langle u_0, v\rangle^2 > 1/(4n)$:

$$\begin{aligned} \left|\langle u_t, v\rangle\right| = \frac{\left|\langle u_0, v\rangle\right| \lambda_1^t}{\sqrt{\sum_{i=1}^n \langle u_0, v^i\rangle^2 \lambda_i^{2t}}} &= \frac{1}{\sqrt{1 + \frac{1}{\langle u_0, v\rangle^2} \sum_{i=2}^n \langle u_0, v^i\rangle^2 \left(\frac{\lambda_i}{\lambda_1}\right)^{2t}}} \\ &\geq \frac{1}{\sqrt{1 + 4n \sum_{i=2}^n \langle u_0, v^i\rangle^2 \left(\frac{\lambda_2}{\lambda_1}\right)^{2t}}} \\ &\geq 1 - 2n \left(\frac{\lambda_2}{\lambda_1}\right)^{2t} \ . \end{aligned}$$

Dimension reduction: random projections

Johnson-Lindenstrauss Lemma

Theorem

Let $x_1, \dots, x_n \in \mathbb{R}^p$, and let $\epsilon > 0$. Then, for every $d \geq \dfrac{4\log(n)}{\epsilon - \log(1+\epsilon)}$, there exists a matrix $A \in \mathcal{M}_{d,p}(\mathbb{R})$ such that

$$\forall 1 \leq i \leq j, \quad (1-\epsilon)\left\|x_i - x_j\right\|^2 \leq \left\|Ax_i - Ax_j\right\|^2 \leq (1+\epsilon)\left\|x_i - x_j\right\|^2 .$$

Remark 1: d is independent of p (!)

Remark 1: on the dependence on ϵ

$$\frac{4\log(n)}{\epsilon - \log(1+\epsilon)} \leq \frac{8\log(n)}{\epsilon^2}\left(1 + \frac{\epsilon}{3}\right)^2 .$$

Remark 2: how to find such a matrix A?

For every $d \geq \dfrac{4\log(n) + 2\log(1/\delta)}{\epsilon - \log(1+\epsilon)}$, the probability that a *random matrix* with entries $A_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{d}\right)$ satisfies the lemma is larger than $1 - \delta$.

Proof of the Johnson-Lindenstrauss Lemma

Method: (constructive) probabilistic method: we choose $A_{i,j} \overset{iid}{\sim} \mathcal{N}\left(0, \frac{1}{d}\right)$. Let $y \in \mathbb{R}^p$ and $Y = Ay$. Then, for all $1 \leq i \leq d$, $Y_i = \sum_{j=1}^{d} A_{i,j} y_j \sim \mathcal{N}\left(0, \frac{\|y\|^2}{d}\right)$. Hence $\mathbb{E}\big[\|Y\|^2\big] = \|y\|^2$. Besides, by the deviation bound for the χ^2 distribution given in the next slide,

$$\mathbb{P}\Big(\|Y\|^2 \geq (1{+}\epsilon)\|y\|^2\Big) = \mathbb{P}\left(\sum_{i=1}^{d}\left(\frac{\sqrt{d}Y_i}{\|y\|}\right)^2 \geq d(1+\epsilon)\right) \leq \exp\left(-d\,\phi^*(\epsilon)\right) \leq \frac{1}{n^2}$$

and similarly $\mathbb{P}\Big(\|Y\|^2 \leq (1-\epsilon)\|y\|^2\Big) \leq \exp\big(-d\,\phi^*(\epsilon)\big) \leq \dfrac{1}{n^2}$.

Applying this result to all $y_{i,j} = x_i - x_j$, $1 \leq i < j \leq n$, we obtain the conclusion by the union bound:

$$\mathbb{P}\Bigg(\bigcup_{1\leq i<j\leq n} \big\|A(y_i - y_j)\big\| \geq (1+\epsilon) \cup \big\|A(y_i - y_j)\big\| \leq (1-\epsilon)\Bigg)$$

$$\leq \frac{n(n-1)}{n^2} < 1\,,$$

and hence there exists at least a matrix A for which the lemma holds.

Deviations of the χ^2 distribution: rate function

Lemma

If $U \sim \mathcal{N}(0,1)$ and $X = U^2 - 1$, then

$$\phi^*(x) = \sup_\lambda \lambda x - \log \mathbb{E}\left[e^{\lambda X}\right] = \frac{x - \log(1+x)}{2} \geq \frac{x^2}{4\left(1+\frac{x}{3}\right)^2} \ .$$

Proof: For every $\lambda < 1/2$,

$$\mathbb{E}\left[e^{\lambda X}\right] = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{\lambda(u^2-1)} e^{-\frac{u^2}{2}} du = \frac{e^{-\lambda}}{\sqrt{2\pi}} \int_{\mathbb{R}} e^{-\frac{(1-2\lambda)u^2}{2}} du = e^{-\lambda} \frac{1}{\sqrt{1-2\lambda}} \ .$$

Hence $\phi(\lambda) = \log \mathbb{E}\left[e^{\lambda X}\right] = -\frac{1}{2}\log(1-2\lambda) - \lambda$. The concave function $\lambda \mapsto \lambda x - \phi(\lambda)$ is maximized at λ^* s.t. $x = \phi'(\lambda^*) = \frac{1}{1-2\lambda^*} - 1$, that is at $\lambda^* = \frac{1}{2}\left(1 - \frac{1}{1+x}\right) = \frac{x}{2(1+x)}$. Hence

$$\phi^*(x) = \lambda^* x - \phi(\lambda^*) = \frac{x - \log(1+x)}{2} \ .$$

The last inequality is obtained by "Pollard's trick" applied to $g(x) = x - \log(1+x)$: since $g(0) = g'(0) = 0$ and since $g''(x) = 1/(1+x)^2$ is convex, by Jensen's inequality

$$\frac{x - \log(1+x)}{x^2/2} = \int_0^1 g''(sx)2(1-s)ds \geq g''\left(\int_0^1 sx2(1-s)ds\right) = g''\left(\frac{x}{3}\right) \ .$$

Deviations of the $\chi^2(d)$ distribution

By Chernoff's method, if $Z \sim \chi^2(d) \overset{\mathrm{dist}}{=} U_1^2 + \cdots + U_d^2$ where $U_i \overset{iid}{\sim} \mathcal{N}(0,1)$:

$$\mathbb{P}\big(Z \geq d(1+\epsilon)\big) \leq \exp\big(-d\phi^*(\epsilon)\big) \leq \exp\left(-\frac{d\epsilon^2}{4\left(1+\frac{\epsilon}{3}\right)^2}\right) .$$

Moreover, since $\phi^*(-\epsilon) = -\frac{\epsilon+\log(1-\epsilon)}{2} = \frac{1}{2}\sum_{k\geq 2}\frac{\epsilon^k}{k} \geq \frac{1}{2}\sum_{k\geq 2}(-1)^k\frac{\epsilon^k}{k} = \phi^*(\epsilon)$, $\mathbb{P}\big(Z \leq d(1-\epsilon)\big) \leq \exp(-d\phi^*(\epsilon))$ and since $\phi^*(-\epsilon) = -\frac{\epsilon+\log(1-\epsilon)}{2} \geq \epsilon^2/4$,

$$\mathbb{P}\big(Z \leq d(1-\epsilon)\big) \leq \exp\left(-\frac{d\epsilon^2}{4}\right) .$$

Note: the Laurent-Massart inequality states that for every $u > 0$,

$$\mathbb{P}\big(Z \geq d + 2\sqrt{du} + 2u\big) \leq \exp\big(-u\big) .$$

It can be deduced from the previous bound by noting that for every $x > 0$

$$\begin{aligned}\phi^*\left(2\sqrt{x}+2x\right) &= x + \frac{1}{2}\left(2\sqrt{x} - \log\left(1 + 2\sqrt{x} + \frac{\left(2\sqrt{x}\right)^2}{2}\right)\right)\\ &\geq x + \frac{1}{2}\Big(2\sqrt{x} - \log\big(\exp(2\sqrt{x})\big)\Big) = x \text{ , and}\end{aligned}$$

$\mathbb{P}\big(Z \geq d + 2\sqrt{du} + 2u\big) = \mathbb{P}\big(\frac{1}{d}\sum_{i=1}^d (U_i^2 - 1) \geq 2\sqrt{\frac{u}{d}} + 2\frac{u}{d}\big) \leq \exp(-d\phi^*(2\sqrt{\frac{u}{d}} + 2\frac{u}{d})) \leq e^{-u}$.

The proof of Laurent and Massart (which takes elements from Birgé and Massart 1998) is a bit different: they note that $\phi(\lambda) = -\frac{1}{2}\log(1-2\lambda) - \lambda = \sum_{k=2}^{\infty}\frac{(2\lambda)^k}{2k} = \lambda^2\sum_{\ell=0}^{\infty}\frac{4(2\lambda)^\ell}{2(\ell+2)} \leq \lambda^2\sum_{\ell=0}^{\infty}(2\lambda)^\ell = \frac{\lambda^2}{1-2\lambda}$, and deduce that $\phi^*(x) \geq \psi^*(x) = \sup_\lambda \lambda x - \frac{\lambda^2}{1-2\lambda} = \frac{x+1-\sqrt{2x+1}}{2}$, while $x > 0$ and $\psi^*(x) = u$ implies $x = 2\sqrt{u} + 2u$. Also note in passing that by Pollard's trick $\phi^*(x) \geq \psi^*(x) \geq \frac{x^2}{4\left(1+\frac{2x}{3}\right)^{3/2}}$.