MATHEMATICS OF DATA SCIENCE

November 27, 2023

Today:

Vector statistics

Johnson-Lindenstrauss

⚠ Notations for random variables/vectors and deterministic variables/vectors will be identical

↳ Last time: we saw concentration inequalities (Hoeffding) for subgaussian random variables

Q) What can we show for vectors of such random variables?

What happens to the norms of such vectors?

Ⓐ Subexponential random variables

Def: A random variable y (with values in $\mathcal{Y} \subseteq \mathbb{R}$) is subexponential if $\exists c > 0$, $\exists K > 0$ such that for every $t \geq 0$

$$(1) \quad \mathbb{P}(|y| \geq t) \leq 2\exp\left(-\frac{ct}{K}\right)$$

$$(2) \quad \mathbb{E}\left[\exp\left(\frac{|y|}{K}\right)\right] \leq 2$$

In that case, the value K is called the subexponential norm of y, and we denote it by $\|y\|_{\psi_1}$

N.B.: For a subgaussian random variable, we have:

$$\mathbb{P}(|y| \geq t) \leq 2\exp\left(-\frac{ct^2}{\|y\|_{\psi_2}^2}\right)$$

Proposition: A random variable y is subgaussian if and only if y^2 is subexponential

Th | Bernstein 's inequality

Let $y_1, \dots, y_N$ be independent, mean zero ($\mathbb{E}[y_1] = \dots = \mathbb{E}[y_N] = 0$), subexponential random variables.

For any $a \in \mathbb{R}^N$ and any $t \geq 0$,

$$\mathbb{P}\left(\left|\sum_{i=1}^{N} a_i y_i\right| \geq t\right)$$

$\|a\|_\infty = \max_{1 \leq i \leq N} |a_i|$

$$\leq 2\exp\left(-c\min\left\{\frac{t^2}{K_{max}^2 \|a\|^2}, \frac{t}{K_{max}\|a\|_\infty}\right\}\right)$$

for some $c > 0$, with $K_{max} = \max_{1 \leq i \leq N} \|y_i\|_{\psi_1}$

Corollary ($a = \begin{bmatrix} 1/N \\ \vdots \\ 1/N \end{bmatrix}$)

Under the same assumptions than above,

$$\mathbb{P}\left(\left|\frac{1}{N}\sum_{i=1}^{N} y_i\right| \geq t\right) \leq 2\exp\left(-cN\min\left(\frac{t^2}{K_{max}^2}, \frac{t}{K_{max}}\right)\right)$$

with $K_{max} = \max_{1 \leq i \leq N} \|y_i\|_{\psi_1}$

↳ the right-hand side goes to 0 as $N \to \infty$.
We say that in that case Bernstein's inequality is similar to a law of large numbers

↳ Just like Hoeffding's, there are many versions of Bernstein's inequality depending on the distributions of the y_i s and /or the vector $a \in \mathbb{R}^N$

(B) Random vectors

Def: A random vector y with values in $\mathcal{Y} \subseteq \mathbb{R}^m$ is defined by the joint distribution of its coordinates $y = \begin{bmatrix} y_1 \\ \vdots \\ y_m \end{bmatrix}$

$$p: \mathcal{Y}_1 \times \dots \times \mathcal{Y}_m \longrightarrow [0,1]$$
$$(z_1, \dots, z_m) \longmapsto \mathbb{P}(y_1 = z_1, \dots, y_m = z_m)$$

(Every y_i is a random variable with values in $\mathcal{Y}_i \subseteq \mathbb{R}$, $\mathcal{Y} = \mathcal{Y}_1 \times \dots \times \mathcal{Y}_m$)

- The expected value of y is a vector in $\mathbb{R}^m$ denoted by $\mathbb{E}[y]$
- The "variance" of y is represented by its covariance matrix, denoted by Σ_y and

defined by

$$\forall (i,j) \in \{1, \dots, m\}^2, \quad [\Sigma_y]_{ij} = \overbrace{\mathbb{E}\left[y_i - \mathbb{E}[y_i]\right] \mathbb{E}\left[y_j - \mathbb{E}[y_j]\right]}^{\mathrm{Cov}(y_i, y_j)}$$

$$\Leftrightarrow \Sigma_y = \mathbb{E}\Big[\underbrace{(y - \mathbb{E}[y])}_{m \times 1} \underbrace{(y - \mathbb{E}[y])^T}_{1 \times m} \Big] \in \mathbb{R}^{m \times m}$$

Ex) Gaussian vector

$$y \sim \mathcal{N}(0, \Sigma) \quad \text{with} \quad \Sigma \in \mathbb{R}^{m \times m}$$

$$\Sigma = \Sigma^T \succ 0$$

$\forall z \in \mathbb{R}^m$,

$$\mathbb{P}(y = z) = \frac{1}{\sqrt{(2\pi)^m \det \Sigma}} \exp(-z^T \Sigma^{-1} z)$$

$\mathbb{E}[y] = 0, \ \Sigma_y = \Sigma$

If the coordinates of y are independent identically distributed $(y_i \sim \mathcal{N}(0, \sigma^2), \sigma^2 > 0)$ then

$$y \sim \mathcal{N}(0, \sigma^2 I)$$

$$\mathbb{E}[y] = 0 \qquad \Sigma_y = \sigma^2 I$$

$\forall z \in \mathbb{R}^m$,

$$\mathbb{P}(y = z) = \frac{1}{\sqrt{(2\pi)^m \sigma^{2m}}} \exp\left(- \frac{\sum_{i=1}^{m} z_i^2}{2\sigma^2}\right)$$

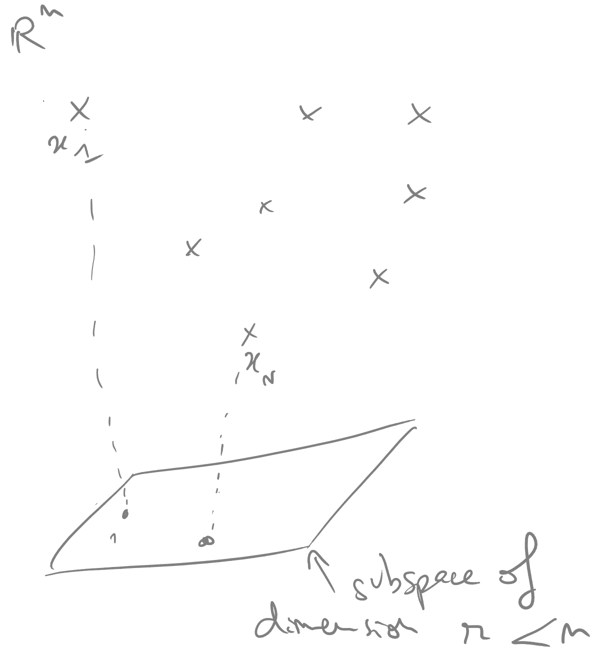
Def: A random vector y is subgaussian if $\forall v \in \mathbb{R}^m$, the random variable $y^T v$ is subgaussian.

Ex) A random vector with independent subgaussian coordinates is subgaussian and we can define

$$\|y\|_{\psi_2} = \max_{\substack{v \in \mathbb{R}^n \\ \|v\|=1}} \|y^T v\|_{\psi_2}$$

↳ If y is a subgaussian random vector, then for any $v \in \mathbb{R}^n$, the random variable $(y^T v)^2$ is subexponential

C) Johnson – Lindenstrauss lemma



Goal : Given $\mathcal{X} = \{x_1, \ldots, x_N\} \subseteq \mathbb{R}^n$, find a subspace of dimension $r < n$ such that projecting $\mathcal{X}$ onto the subspace preserves distances between the points up to a desired accuracy $\varepsilon \in (0,1)$

Ideally, we would like to find projections $\tilde{x}_1, \ldots, \tilde{x}_N$ such that $\forall (i,j) \in \{1, \ldots, N\}^2$,

$$(1-\varepsilon)\|x_i - x_j\| \leq \|\tilde{x}_i - \tilde{x}_j\| \leq (1+\varepsilon)\|x_i - x_j\|$$

↳ Not possible to guarantee deterministically, but possible to satisfy with high probability

Theorem: [Johnson - Lindenstrauss lemma]

Let $\mathcal{X} = \{x_1, \dots, x_N\}$ be a set of points in $\mathbb{R}^n$ and let $\varepsilon > 0$.

Let $y_1, \dots, y_r$ be m random vectors in $\mathbb{R}^m$ such that

- $y_1, \dots, y_r$ are independent, subgaussian
- $\mathbb{E}[y_i] = 0 \quad \forall i = 1..r$
- $\Sigma_{y_i} = I_{\mathbb{R}^{n \times m}}$ (y_i is "isotropic")

y vector of iid Rademacher variables (± 1)

Define P as the matrix in $\mathbb{R}^{r \times m}$ given by $P = \frac{1}{\sqrt{r}} \times \begin{bmatrix} y_1^T \\ \vdots \\ y_r^T \end{bmatrix}$.

Defines a random subspace in $\mathbb{R}^m$ when $r \leq m$

Then there exists $C > 0$ that does not depend on m, r, or N such that

$$r \geq C \varepsilon^{-2} \ln(N),$$

The subspace dimension only depends on ε and on $\ln(N)$

then

$$\mathbb{P}\Big((1-\varepsilon)\|x_i - x_j\| \leq \|Px_i - Px_j\| \leq (1+\varepsilon)\|x_i - x_j\| \quad \forall (i,j)\Big) \geq 0.99$$

High-probability result

Projection of x_i onto the random subspace

$$\underbrace{P}_{r \times m}\underbrace{x_i}_{m \times 1} = \underbrace{\frac{1}{\sqrt{r}} \begin{bmatrix} y_1^T x_i \\ \vdots \\ y_r^T x_i \end{bmatrix}}_{\in \mathbb{R}^r}$$