The curse of dimensionality

The curse of dimensionality

- many applications require high dimensional data

The curse of dimensionality

- many applications require high dimensional data
- many algorithms become inefficient with high dimensional

The curse of dimensionality

- many applications require high dimensional data
- many algorithms become inefficient with high dimensional
- like to replace high dimensional data by smaller dimensional data without losing too much information

The curse of dimensionality

- many applications require high dimensional data
- many algorithms become inefficient with high dimensional
- like to replace high dimensional data by smaller dimensional data without losing too much information
- see two techniques for this task

The curse of dimensionality

- many applications require high dimensional data
- many algorithms become inefficient with high dimensional
- like to replace high dimensional data by smaller dimensional data without losing too much information
- see two techniques for this task
  1. Johnson-Lindenstrauss lemma

The curse of dimensionality

- many applications require high dimensional data
- many algorithms become inefficient with high dimensional
- like to replace high dimensional data by smaller dimensional data without losing too much information
- see two techniques for this task
  1. Johnson-Lindenstrauss lemma
  2. singular value decomposition / principal component analysis
- another technique is feature selection

Theorem 5.1

Let P be a set of n points in $\mathbb{R}^d$ and $0 < \epsilon < 1$. Then, for c large enough, there is an embedding $\pi : P \to \mathbb{R}^{c\log(n)/\epsilon^2}$, such that for all $p, q \in P$

$$(1-\epsilon)\cdot D_{l_2}(p,q) \le D_{l_2}(\pi(p),\pi(q)) \le (1+\epsilon)\cdot D_{l_2}(p,q).$$

Gaussian distribution

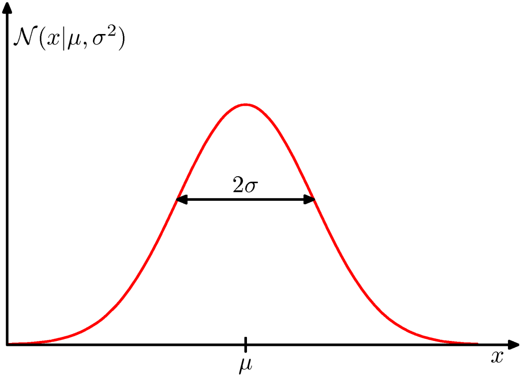
- $\mu \in \mathbb{R}, \sigma \in \mathbb{R}_{>0}$
- density function

$$\mathcal{N}(\cdot|\mu,\sigma^2) : \mathbb{R} \to \mathbb{R}_{>0}$$
$$\mathcal{N}(x|\mu,\sigma^2) \mapsto \frac{1}{\sqrt{2\pi\sigma^2}} \cdot \exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

- distribution with density function $\mathcal{N}(\cdot \mid \mu,\sigma^2)$ called Gaussian or normal distribution $\mathcal{N}(\mu,\sigma^2)$ with mean μ and standard deviation σ,i.e.

$$\forall l \in \mathbb{R} : \Pr[x \leq l] = \int_{-\infty}^{l} \mathcal{N}(x \mid \mu,\sigma^2)\mathrm{d}x.$$

Density function of Gaussian distribution



The Johnson-Lindenstrauss lemma - the construction

Random mapping

- $A = (r_{ij})_{1 \le i \le k, 1 \le j \le d} \in \mathbb{R}^{k \times d}$, where each r_{ij} is chosen according to $\mathcal{N}(0,1)$.
- $\forall x \in \mathbb{R}^d : \pi_A(x) = \frac{1}{\sqrt{k}} \cdot A \cdot x$.

Random mapping

- $A = (r_{ij})_{1 \leq i \leq k, 1 \leq j \leq d} \in \mathbb{R}^{k \times d}$, where each r_{ij} is chosen according to $\mathcal{N}(0,1)$.
- $\forall x \in \mathbb{R}^d : \pi_A(x) = \frac{1}{\sqrt{k}} \cdot A \cdot x$.

Lemma 5.2

Let $\pi_A : \mathbb{R}^d \to \mathbb{R}^k$ be a chosen as above, let $u \in \mathbb{R}^d$ be a vector, and let $0 < \epsilon < 1$. Then, for c large enough and $k = c \cdot \log(n)/\epsilon^2$:

$$\Pr\left[(1-\epsilon) \leq \frac{\|\pi_A(u)\|_2}{\|u\|_2} \leq (1+\epsilon)\right] \geq 1 - \frac{1}{3n^2}.$$