High-Dimensional Space

Prof. Asim Tewari
IIT Bombay

Dimensionality reduction

- Feature selection: Feature selection approaches try to find a subset of the original variables (also called features or attributes)
  - Filter strategy (e.g. information gain)
  - Wrapper strategy (e.g. search guided by accuracy)
  - Embedded strategy (features are selected to add or be removed while building the model based on the prediction errors)

- Feature projection: Feature projection transforms the data in the high-dimensional space to a space of fewer dimensions. The data transformation may be linear, as in principal component analysis (PCA), but many nonlinear dimensionality reduction techniques also exist. For multidimensional data, tensor representation can be used in dimensionality reduction through multilinear subspace learning.

Markov's inequality

- Let x be a nonnegative random variable. Then for a > 0,

$$Prob(x \geq a) \leq \frac{E(x)}{a}$$

$$E(x) = \int_0^\infty xp(x)dx = \int_0^a xp(x)dx + \int_a^\infty xp(x)dx$$

Chebyshev's inequality

- Let x be a random variable. Then for c > 0,

$$Prob\Big(\,|x - E(x)| \geq c\Big) \leq \frac{Var(x)}{c^2}$$

Proof $\text{Prob}(|x - E(x)| \geq c) = \text{Prob}(|x - E(x)|^2 \geq c^2)$. Note that $y = |x - E(x)|^2$ is a nonnegative random variable and $E(y) = Var(x)$, so Markov's inequality can be applied giving:

$$\text{Prob}(|x - E(x)| \geq c) = \text{Prob}\left(|x - E(x)|^2 \geq c^2\right) \leq \frac{E(|x - E(x)|^2)}{c^2} = \frac{Var(x)}{c^2}.$$

Law of Large Numbers

- Let $x_1, x_2, \ldots, x_n$ be n independent samples of a random variable x. Then

$$Prob\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) \leq \frac{Var(x)}{n\epsilon^2}$$

Law of Large Numbers

$$\text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) = \text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)\right| \geq \epsilon\right)$$

By Chebyshev's inequality,

$$\begin{aligned}\text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E(x)\right| \geq \epsilon\right) &= \text{Prob}\left(\left|\frac{x_1 + x_2 + \cdots + x_n}{n} - E\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)\right| \geq \epsilon\right) \\ &\leq \frac{Var\left(\frac{x_1 + x_2 + \cdots + x_n}{n}\right)}{\epsilon^2} \\ &= \frac{1}{n^2\epsilon^2} Var(x_1 + x_2 + \cdots + x_n) \\ &= \frac{Var(x)}{n\epsilon^2}.\end{aligned}$$

Law of Large Numbers

Implications of the Law of Large Numbers:

- If we draw a point x from a d-dimensional Gaussian with unit variance, then it will lie on a sphere of radius sqrt(d).

This is because

$$|x|^2 \approx d$$

Law of Large Numbers

Implications of the Law of Large Numbers:

- If we draw two point y and z from a d-dimensional Gaussian with unit variance, then they would be approximately orthogonal

This is since for all i, $E(y_i - z_i)^2 = E(y_i^2) + E(z_i^2) - 2E(y_i z_i)$

$$= Var(y_i) + Var(z_i) - 2E(y_i)E(z_i) = 2$$

Therefore, $|\mathbf{y} - \mathbf{z}|^2 = \sum_{i=1}^{d} (y_i - z_i)^2 \approx 2d$

Thus by the Pythagorean theorem, the two points y and z must be approximately orthogonal.

Volume in objects of High Dimensions

An important property of high-dimensional objects is that most of their volume is near the surface.
to produce a new object $(1-\epsilon)A = \{(1-\epsilon)x | x \in A\}$. Then the following equality holds:

$$\text{volume}\big((1-\epsilon)A\big) = (1-\epsilon)^d \text{volume}(A).$$

To see that this is true, partition A into infinitesimal cubes. Then, $(1-\varepsilon)A$ is the union of a set of cubes obtained by shrinking the cubes in A by a factor of $1-\varepsilon$. When we shrink each of the $2d$ sides of a d-dimensional cube by a factor f, its volume shrinks by a factor of f^d. Using the fact that $1-x \leq e^{-x}$, for any object A in R^d we have:

$$\frac{\text{volume}\big((1-\epsilon)A\big)}{\text{volume}(A)} = (1-\epsilon)^d \leq e^{-\epsilon d}.$$

Fixing ϵ and letting $d \to \infty$, the above quantity rapidly approaches zero. This means that nearly all of the volume of A must be in the portion of A that does not belong to the region $(1-\epsilon)A$.

Volume of the Unit Ball

To calculate the volume $V(d)$ of the unit ball in R^d, one can integrate in either Cartesian or polar coordinates. In Cartesian coordinates the volume is given by

$$V(d) = \int_{x_1=-1}^{x_1=1} \int_{x_2=-\sqrt{1-x_1^2}}^{x_2=\sqrt{1-x_1^2}} \cdots \int_{x_d=-\sqrt{1-x_1^2-\cdots-x_{d-1}^2}}^{x_d=\sqrt{1-x_1^2-\cdots-x_{d-1}^2}} dx_d \cdots dx_2 dx_1.$$

Since the limits of the integrals are complicated, it is easier to integrate using polar coordinates. In polar coordinates, $V(d)$ is given by

$$V(d) = \int_{S^d} \int_{r=0}^{1} r^{d-1} dr d\Omega.$$

$$V(d) = \int_{S^d} d\Omega \int_{r=0}^{1} r^{d-1} dr = \frac{1}{d} \int_{S^d} d\Omega = \frac{A(d)}{d}$$

where $A(d)$ is the surface area of the d-dimensional unit ball. For instance, for $d = 3$ the surface area is 4π and the volume is $\frac{4}{3}\pi$. The question remains how to determine the surface area $A(d) = \int_{S^d} d\Omega$ for general d.

Volume of the Unit Ball

Consider a different integral,

$$I(d) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} e^{-(x_1^2 + x_2^2 + \cdots x_d^2)} dx_d \cdots dx_2 dx_1.$$

Including the exponential allows integration to infinity rather than stopping at the surface of the sphere. Thus, $I(d)$ can be computed by integrating in both Cartesian and polar coordinates. Integrating in polar coordinates will relate $I(d)$ to the surface area $A(d)$. Equating the two results for $I(d)$ allows one to solve for $A(d)$.

First, calculate $I(d)$ by integration in Cartesian coordinates.

$$I(d) = \left[\int_{-\infty}^{\infty} e^{-x^2} dx \right]^d = \left(\sqrt{\pi}\right)^d = \pi^{\frac{d}{2}}.$$

Here, we have used the fact that $\int_{-\infty}^{\infty} e^{-x^2} \, dx = \sqrt{\pi}$

Next, calculate $I(d)$ by integrating in polar coordinates. The volume of the differential element is $r^{d-1} d\Omega dr$. Thus,

$$I(d) = \int_{S^d} d\Omega \int_0^{\infty} e^{-r^2} r^{d-1} dr.$$

Volume of the Unit Ball

$$I(d) = \int_{S^d} d\Omega \int_0^\infty e^{-r^2} r^{d-1} dr.$$

The integral $\int_{S^d} d\Omega$ is the integral over the entire solid angle and gives the surface area, $A(d)$, of a unit sphere. Thus, $I(d) = A(d) \int_0^\infty e^{-r^2} r^{d-1} dr$. Evaluating the remaining integral gives

$$\int_0^\infty e^{-r^2} r^{d-1} dr = \int_0^\infty e^{-t} t^{\frac{d-1}{2}} \left(\frac{1}{2} t^{-\frac{1}{2}} dt\right) = \frac{1}{2} \int_0^\infty e^{-t} t^{\frac{d}{2}-1} dt = \frac{1}{2} \Gamma\left(\frac{d}{2}\right),$$

and hence, $I(d) = A(d) \frac{1}{2} \Gamma\left(\frac{d}{2}\right)$ where the Gamma function $\Gamma(x)$ is a generalization of the factorial function for non-integer values of x. $\Gamma(x) = (x-1)\Gamma(x-1), \Gamma(1) = \Gamma(2) = 1$, and $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. For integer $x, \Gamma(x) = (x-1)!$.

Combining $I(d) = \pi^{\frac{d}{2}}$ with $I(d) = A(d) \frac{1}{2} \Gamma\left(\frac{d}{2}\right)$ yields

$$A(d) = \frac{\pi^{\frac{d}{2}}}{\frac{1}{2}\Gamma\left(\frac{d}{2}\right)},$$

Volume of the Unit Ball

The surface area $A(d)$ and the volume $V(d)$ of a unit-radius ball in d-dimensions are given by

$$A(d) = \frac{2\pi^{\frac{d}{2}}}{\Gamma(\frac{d}{2})} \quad \text{and} \quad V(d) = \frac{2\pi^{\frac{d}{2}}}{d\,\Gamma(\frac{d}{2})}.$$

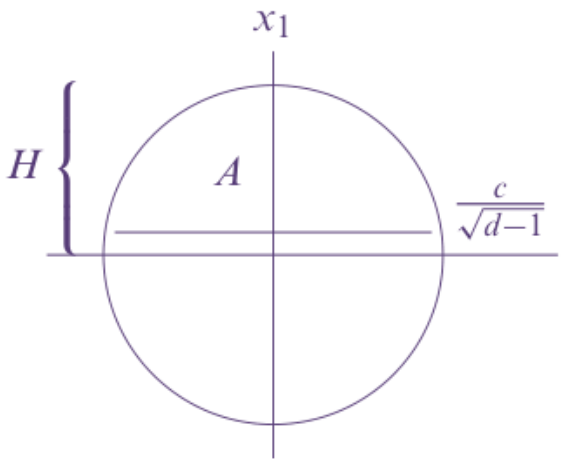
To check the formula for the volume of a unit ball, note that $V(2) = \pi$ and $V(3) = \frac{2}{3}\frac{\pi^{\frac{3}{2}}}{\Gamma\left(\frac{3}{2}\right)} = \frac{4}{3}\pi$, which are the correct volumes for the unit balls in two and three dimensions. To check the formula for the surface area of a unit ball, note that $A(2) = 2\pi$ and $A(3) = \frac{2\pi^{\frac{3}{2}}}{\frac{1}{2}\sqrt{\pi}} = 4\pi$, which are the correct surface areas for the unit ball in two and three dimensions. Note that $\pi^{\frac{d}{2}}$ is an exponential in $\frac{d}{2}$ and $\Gamma\left(\frac{d}{2}\right)$ grows as the factorial of $\frac{d}{2}$. This implies that $\lim_{d\to\infty} V(d) = 0$

Volume near the Equator

- An interesting fact about the unit ball in high dimensions is that most of its volume is concentrated near its “equator.” In particular, for any unit-length vector v defining “north,” most of the volume of the unit ball lies in the thin slab of points whose dot- product with v has magnitude O(1/ √d).

Volume near the Equator

For $c \geq 1$ and $d \geq 3$, at least a $1 - \frac{2}{c}e^{-c^2/2}$ fraction of the volume of the d-dimensional unit ball has $|x_1| \leq \frac{c}{\sqrt{d-1}}$.



We can then show that the ratio of the volume of A to the volume of H goes to zero by calculating an upper bound on volume(A) and a lower bound on volume(H) and proving that

$$\frac{\text{volume}(A)}{\text{volume}(H)} \leq \frac{\text{upper bound volume}(A)}{\text{lower bound volume}(H)} = \frac{2}{c}e^{-\frac{c^2}{2}}$$

Volume near the Equator

$$\text{volume}(A) \leq \int_{\frac{c}{\sqrt{d-1}}}^{\infty} \frac{x_1\sqrt{d-1}}{c} e^{-\frac{d-1}{2}x_1^2} V(d-1)dx_1$$

$$= V(d-1)\frac{\sqrt{d-1}}{c}\int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-\frac{d-1}{2}x_1^2} dx_1$$

Now

$$\int_{\frac{c}{\sqrt{d-1}}}^{\infty} x_1 e^{-\frac{d-1}{2}x_1^2} dx_1 = -\frac{1}{d-1} e^{-\frac{d-1}{2}x_1^2}\Big|_{\frac{c}{\sqrt{(d-1)}}}^{\infty} = \frac{1}{d-1} e^{-\frac{c^2}{2}}$$

Thus, an upper bound on volume(A) is $\frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}}$.

Volume near the Equator

The volume of the hemisphere below the plane $x_1 = \frac{1}{\sqrt{d-1}}$ is a lower bound on the entire volume of the upper hemisphere, and this volume is at least that of a cylinder of height $\frac{1}{\sqrt{d-1}}$ and radius $\sqrt{1 - \frac{1}{d-1}}$. The volume of the cylinder is $V(d-1)(1 - \frac{1}{d-1})^{\frac{d-1}{2}} \frac{1}{\sqrt{d-1}}$. Using the fact that $(1-x)^a \geq 1 - ax$ for $a \geq 1$, the volume of the cylinder is at least $\frac{V(d-1)}{2\sqrt{d-1}}$ for $d \geq 3$.

Thus,

$$\text{ratio} \leq \frac{\text{upper bound above plane}}{\text{lower bound total hemisphere}} = \frac{\frac{V(d-1)}{c\sqrt{d-1}} e^{-\frac{c^2}{2}}}{\frac{V(d-1)}{2\sqrt{d-1}}} = \frac{2}{c} e^{-\frac{c^2}{2}}.$$

Thus: *For $c \geq 1$ and $d \geq 3$, at least a $1 - \frac{2}{c} e^{-c^2/2}$ fraction of the volume of the d-dimensional unit ball has $|x_1| \leq \frac{c}{\sqrt{d-1}}$.*

Random Projection

Johnson-Lindenstrauss Lemma

- The projection $f : R^d \rightarrow R^k$. Pick k Gaussian vectors u_1, u_2, . . . , u_k in Rd with unit-variance coordinates. For any vector v, define the projection f (v) by:

$$f(v) = (u_1 \cdot v, u_2 \cdot v, \ldots, u_k \cdot v).$$

(The projection f (v) is the vector of dot products of v with the u_i) . Then, the Johnson-Lindenstrauss Lemma

States that:

- With high probability, $|f(v)| \approx \sqrt{k}|v|$
- And for any two vectors v_1 and v_2,

$$f(v_1 - v_2) = f(v_1) - f(v_2)$$