PCMI Lecture # 10

Pedagogical Lecture: Clustering, SVD, FLD, SVM, concentration of measure

For this lecture we will be looking at images as points in high (infinite even) dimensional spaces. It is in fact very natural to think of images as vectors, whether we are modeling them as functions $u : [0,1]^2 \to \mathbb{R}$ or $\mathbb{R}^3$ or as $n \times n$ arrays of intensities.

Tasks such as detection, classification, inference are tasks we would like to do automatically.

And both the Theory and Computation required are accessible to good undergraduates.

$\boxed{\text{PCA / SVD}}$ In this lecture we begin with a classic method for generating a reduced complexity model of data in high dimensional spaces: principal component Analysis (PCA)

$$V \equiv n \overset{N}{\begin{bmatrix} | & | & & | \\ v_1 & v_2 & \cdots & v_N \\ | & | & & | \end{bmatrix}}$$

← data, arranged in columns

data $\subset \mathbb{R}^n$

$$C \equiv \frac{1}{(N-1)} V V^T$$

(where we have assumed that the mean has been removed --- $\frac{1}{N} V \begin{bmatrix} 1 \\ \vdots \\ 1 \end{bmatrix} [1\,1 \cdots 1]$ has been subtracted from V)

$\uparrow N \times 1$ $\quad$ $\uparrow 1 \times N$

Now we diagonalize C. $\quad C = O_L \Sigma O_L^T$.

Let's use the SVD we met earlier (Lecture 4)

$$V = \begin{array}{c} \\ n \end{array}\overset{n}{\Big[\; O_L \;\Big]}\begin{bmatrix} \sigma_1 & & & 0 & \vdots \\ & \sigma_2 & & & \vdots \\ 0 & & \ddots & & \vdots \\ & & & \sigma_n & \vdots \end{bmatrix}\overset{N}{} \overset{N}{\Bigg[\; O_R^T \;\Bigg]}\, N$$

$$\Rightarrow C = \frac{1}{N-1} V V^T = \frac{1}{N-1} O_L \Sigma \Sigma^T O_L^T$$

$\Rightarrow$ eigenvalues of C are $\sigma_1^2, \sigma_2^2, \ldots \sigma_n^2$ from the SVD of V

An exceedingly useful property of the SVD

$$V_K \equiv O_L \cdot [\Sigma_K] \cdot O_R^T \quad \text{where} \quad \Sigma_K = \begin{bmatrix} \sigma_1 & & & & & \vdots & \\ & \ddots & & 0 & & \vdots & \\ & & \sigma_K & & & \vdots & 0 \\ & 0 & & 0 & & \vdots & \\ & & & & \ddots\, 0 & \vdots & \end{bmatrix}$$

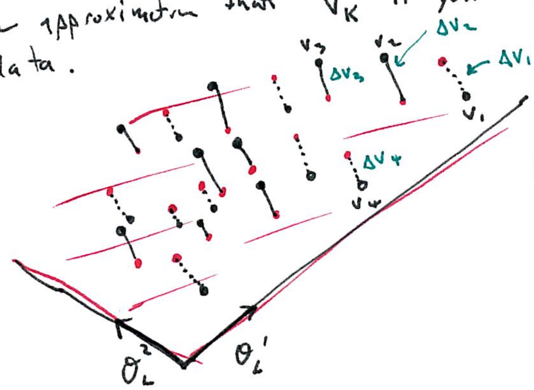
is the best rank K approximation of V in both the L^2 operator norm and the Frobenius norm.

$\|V\|_F = \sqrt{\sum_{i,j} |V_{ij}|^2}$

$\|V\|_2 \equiv \max_{X \neq 0} \frac{\|VX\|}{\|X\|}$

* $\|V - V_K\|_F \leq \|V - W\|_F$ any W with rank at most K
* $\|V - V_K\|_2 \leq \|V - W\|_2$ " " "

The first inequality says that using the L^2 distance, one cannot do better in linear approximation than V_K if you want a K-dimensional approximation of the data.



span(O_L^1, O_L^2) give best 2 dim approx subspace for the data ... the sum of the lengths to the subspace from the data, squared is minimal

$\sum_i \|\Delta v_i\|_2^2$ is minimal

Comment: Because of its ubiquitous usefulness and its power of illumination, students should learn about the SVD as soon as possible.

Thinking about Matrices, Linear Algebra and Data with the SVD in your toolset is much easier than the same sans-SVD!

what is the SVD good for?

1. as we have already seen, optimal dimension reductions ... optimal in the L^2 sense.

2. ... and before even seeing that, we saw it's use in regularization of inverse problems.

3. reduction in computational costs. This is really a corollary or side effect of 1, but it is so important I give it its own bullet. Simply reducing the dimension of the representation can make a huge difference because of the resulting reduction in computational costs.

4. understanding: what are the important components in my data?

Fisher Linear Discriminant (FLD)

Suppose our data $v_1, v_2, \dots, v_N$ are each given either the label 0 or the label 1. Assuming that these labeled data points represent some distribution of type 0 and type 1 points well, we can try to build a classifier that separates the labeled data well, believing that it will then be a good classifier for new, unlabeled data.

Data: $\{(v_i, y_i)\}_{i=1}^{N}$, $v_i \in \mathbb{R}^n$, $y_i \in \{0, 1\}$

(1) $$m_0 \text{ or } m_1 = \frac{\sum_{y_i = 0 \text{ or } 1} V_i}{\#\{y_i = 0 \text{ or } 1\}} \rightarrow N_0, N_1 \qquad N_0 + N_1 = N$$

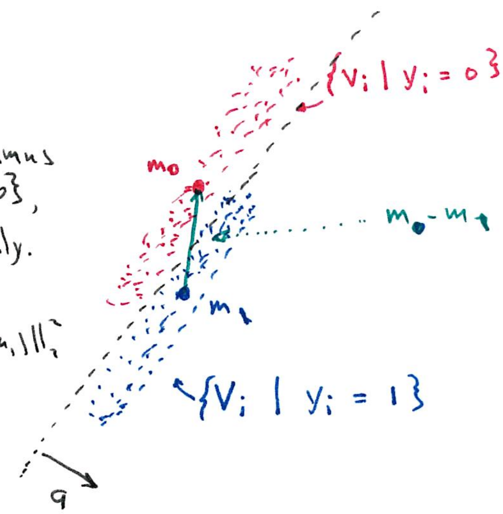
(2) $a \in \mathbb{R}^n$

(3) W_0, W_1 = matrices with columns equal to $\{V_i \mid y_i = 0\}$, $\{V_i \mid y_i = 1\}$ respectively.

I use $W_0 - m_0$ to denote $W_0 - \begin{bmatrix} | \\ m_0 \\ | \end{bmatrix} \underset{1\ 2\ \cdots\ N_0}{[1\ 1 \cdots 1]}$ and analogously for $W_1 - m_1$

(4) $\sigma_0^2, \sigma_1^2 \equiv \|a^T(W_0 - m_0)\|_2^2, \|a^T(W_1 - m_1)\|_2^2$

$$\|a^T(W-m)\|_2^2 = a^T \underbrace{(W-m)(W-m)^T}_{S} a$$



FLD: "Best" a given by

$$\max_a J(a) \equiv \frac{a^T(m_0 - m_1)(m_0 - m_1)^T a}{\sigma_0^2 + \sigma_1^2}$$

$$= \frac{a^T(m_0 - m_1)(m_0 - m_1)^T a}{a^T(S_1 + S_2)a}$$

Solution: $a = (S_1 + S_2)^{-1}(m_0 - m_1)$ (all we care about is the unsigned direction)

quick proof: $\min\ a^T(S_1 + S_2)a$ subject to $a^T(m_0 - m_1) = 1$

$\Rightarrow (S_1 + S_2)a + \lambda(m_0 - m_1) = 0$

$\Rightarrow a = (S_1 + S_2)^{-1}(m_0 - m_1)$ works

Caution: this can work pretty badly!

But: The ideas encountered by students when seeing this are all important ... and it is often the first thing you should try with data classification problems.

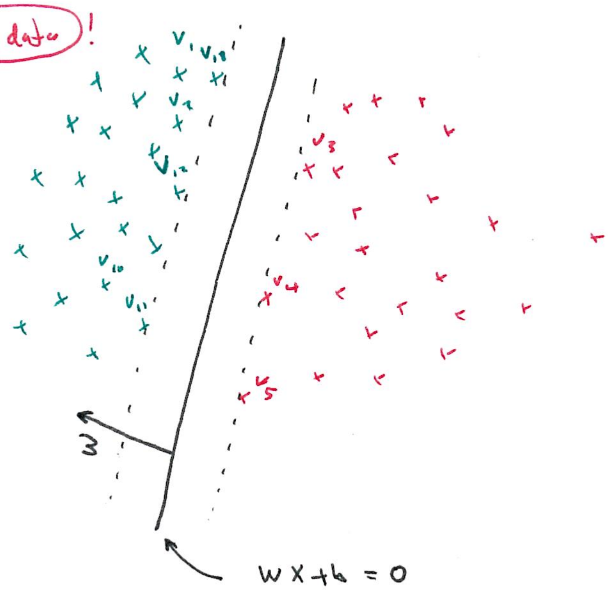
Support Vector Machines (SVMs)

The next trick in the bag should be support vector machines.

For separable data!

* now for convenience *

$y_i \in \{-1, 1\}$



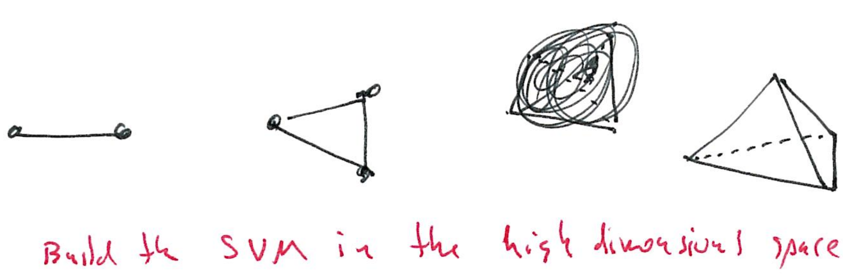
$$\min \|w\| \quad \text{subject to} \quad y_i\,(wx + b) \geq 1$$

$\Rightarrow$ this results in a $w, b \ni$ margin is maximized; margin = min distance from separating hyperplane to data.

Kernel trick: all data is separable!

Idea: map data to a high (infinite) dimensional space where we know how to compute inner products.

why?: The data will, generically, land on the vertices of a simplex $\Rightarrow$ any subset is separable from its complement by a hyperplane in the high dimensional space



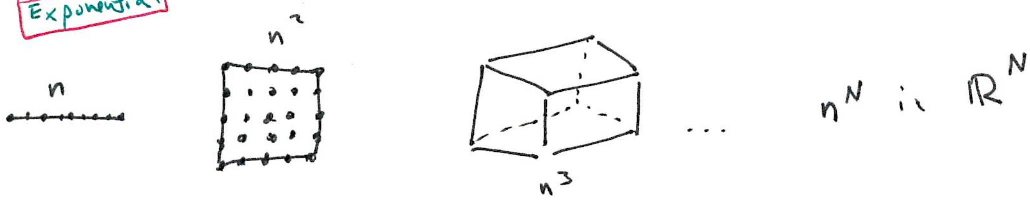
Build the SVM in the high dimensional space.

Concentration of measure and high dimensions

working high dimensions brings both blessings and curses:

First, the cursing ☺

Exponential



... even if you want to "cover" the surface of a cube by only putting points at the vertices, in $\boxed{100 \text{ dimensions}}$ you need $\boxed{2^{100}}$ points which is about $\boxed{10^{30}}$ points !!

... and as high dimensional spaces go $\mathbb{R}^{100}$ is <u>very</u> low dimensional !

$$\boxed{C^N \text{ and } N! \text{ get huge fast as } N \to \infty}$$

but there are blessings !!

Concentration phenomena

These phenomena are both illuminating and entertaining ... students will find this something truly new and surprising ... and accessible to playing around themselves.

We start with volumes of balls: spheres and balls in $\mathbb{R}^n$

$$V(B_r) = \alpha(n)\, r^n \qquad \alpha(n) \equiv V(B_1)$$

$$V(S_r) = n\alpha(n)\, r^{n-1} \qquad = \frac{\pi^{n/2}}{\Gamma(\frac{n}{2}+1)}$$

Sterling $\longrightarrow$ $$n! = \sqrt{2\pi n}\left(\frac{n}{e}\right)^n e^{\theta(n)/12n} \qquad = \frac{\pi^{n/2}}{\frac{n}{2}\Gamma(\frac{n}{2})}$$

$n = 1, 2, \dots$

$0 < \theta(n) < 1$

(playing around with spheres and balls in $\mathbb{R}^n$ can teach a great deal actually)

1

notice: $$V(S_r) = \frac{\partial}{\partial r} V(B_r)$$

what is $\frac{\partial}{\partial r} V(S_r)$?

answer:

$$\frac{\partial}{\partial r} V(S_r) = n \cdot n-1\, \alpha(n)\, r^{n-2}$$

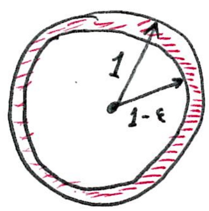
$$= \frac{n-1}{r}\left(n\,\alpha(n)\, r^{n-1}\right)$$

$$= \frac{n-1}{r} V(S_r)$$

$$= \int_{\partial B_r = S_r} \vec{H}\cdot\vec{n}\, d\sigma = \int_{S_r} \vec{H}\cdot\vec{n}\, d\mathcal{H}^{n-1}$$

total mean curvature of sphere of radius r

2



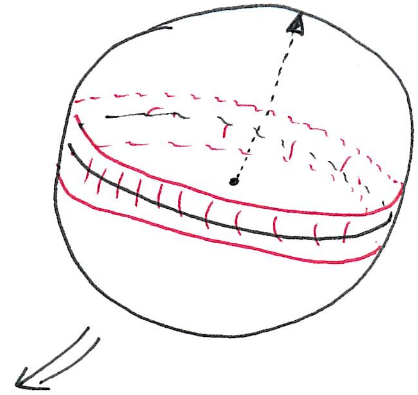
$$\frac{V(B_{1-\epsilon})}{V(B_1)} \xrightarrow[n\to\infty]{} 0 \;!$$ (for any $\epsilon > 0$ no matter how small)

i.e. $(1-\epsilon)^n \underset{n\to\infty}{\to} 0$

As $n\to\infty$ almost all the volume of a ball is in a very thin shell on the boundary!

This is the essence of concentration phenomena

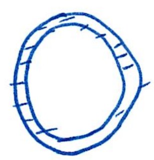
③ Equitorial Bands on Spheres

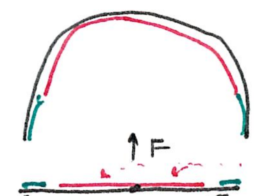


contains almost all the area of the surface as $n \to \infty$ for any $\varepsilon > 0$!

Explains: random vectors in high dimension are almost orthogonal.

get hemisphere from warped disk





Jacobian of mapping F is bigger here than here.

④ Tons of other things to explore

* Gaussian dist in $\mathbb{R}^n \Rightarrow$ most vectors have length $\sqrt{n}$ when $C = \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}$
* volume of unit Ball $B_1 \to 0$ $\quad n \to \infty$
* cubes and balls interacts...

⋮

* Lipschitz functions on spheres are approximately constant as $n \to \infty$ (i.e. on a set of large measure fraction it is within ε of a median value).

⋮