Class Notes

Dustin Leatherman

April 8, 2020

Contents

1 Review & Introduction (2020/03/31)

1.1 Review

Orthogonal: Vectors are orthogonal when the dot product = 0.

1.1.1 Basis

$$
\begin{aligned}
\underset{(n\times1)}{\vec{y}} &= \underset{(n\times p)}{A}\underset{(p\times1)}{\vec{x}} \\
&= B\vec{c} \\
&= \Sigma c_i \vec{b_i} \text{ (most } c_i = 0)
\end{aligned}
\tag{1}
$$

A: Basis Matrix

Properties of a Good Basis

- not all are orthogonal
- Allows for a sparse vector to be used ad the constant vector $\vec{c}$

Identity Matrices are the *worst* basis because most coefficients are non-zero.

2-Sparse Vector

$$
\vec{c} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 3 \\ 0 \\ 0 \\ 4 \end{bmatrix}
\tag{2}
$$

Very important!

> When dealing with Natural images and a good basis, there is a sparse vector.

1.1.2 Kernel

The kernel of a linear mapping is the set of vectors mapped to the 0 vector. The kernel is often referred to as the **null space**. Vectors should be linearly independent.

$$
Ker(A) = \vec{x} \in \mathbb{R}^n \colon A\vec{x} = \vec{0}
\tag{3}
$$

A must be designed such that the Kernel of A does not contain any s-sparse vector other than $\vec{0}$

Main Idea: For (1), reduce $\vec{y}$ to a K-Sparse matrix to reduce the amount of non-zero numbers.

1.2 Linear Algebra Review

$$\vec{u} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \vec{v} = \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \tag{4}$$

$$\begin{aligned} \underset{(1\times3)(3\times1)}{\vec{u}^T\vec{v}} &= \begin{bmatrix} 1 & 2 & -1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} = 1 + 2 - 2 = 1 \\ &= \vec{u} \cdot \vec{v} \end{aligned} \tag{5}$$

$$\underset{(3\times1)(1\times3)}{\vec{u}\ \vec{v}^T} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 2 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 2 \\ 2 & 2 & 4 \\ -1 & -1 & -2 \end{bmatrix} \tag{6}$$

$\vec{u}\ \vec{v}^T \neq \vec{u}^T\ \vec{v}$

1.2.1 Inner Product

$$\begin{aligned} < \vec{a}, \vec{b} > &= \vec{a} \cdot \vec{b} \\ &= \vec{a}^T \vec{b} \end{aligned} \tag{7}$$

1.2.2 Cauchy-Schwartz Inequality

$$\vec{a} = \begin{bmatrix} 1 \\ 2 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \\ 2 \end{bmatrix} \tag{8}$$

$$\begin{aligned} | < \vec{a}, \vec{b} > | &\leq \sqrt{1^2 + 2^2 + (-1)^2} \times \sqrt{1^2 + 1^2 + 2^2} \\ | < \vec{a}, \vec{b} > | &\leq ||\vec{a}||_2\ ||\vec{b}||_2 \text{ (euclidean/l2-norm)} \end{aligned} \tag{9}$$

1.2.3 Norms

Why is the l1 norm preferred for ML opposed to the classic l2 norm?

Philosophically,

If we looked at a sphere in l2 norm, the shadow casted would be a circle regardless of the direction of the light.

Looking at a sphere in the l1 norm is shaped as a tetrahedron. The shadow cast by a tetrahedron is different for different angles so observing the shadow provides a lot more context about the sphere.

1. Euclidean/l2

   Sphere: $||\vec{x}||_2 = \sqrt{(-4)^2 + 3^2} = \sqrt{25} = 5$

   (a) FOIL Given 2 fixed vectors x,y. Consider the l2-norm squared:

   $$f(t) = ||x + ty||_2^2$$

   $$\begin{aligned} f(t) =& ||x + ty||_2^2 \\ =& < x + ty, x + ty > \\ =& < x, x > +t < x, y > +t < y, x > +t^2 < y, y > \\ =& ||x||_2^2 + 2t < x, y > +t^2 ||y||_2^2 \end{aligned} \tag{10}$$

   Note: t<x,y> and t<y,x> can be combined because their dot-products are equivalent. $\vec{x} \cdot \vec{y} = \vec{y} \cdot \vec{x}$

   When using Machine Learning, don't use l2 norms. Use l1

   (b) Derivative

   $$\begin{aligned} \frac{d}{dt}(||x + ty||_2^2) =& 2 < x, y > +2t||y||_2^2 \\ =& 2x^T y + 2ty^T y \end{aligned} \tag{11}$$

2. Simplex/l1

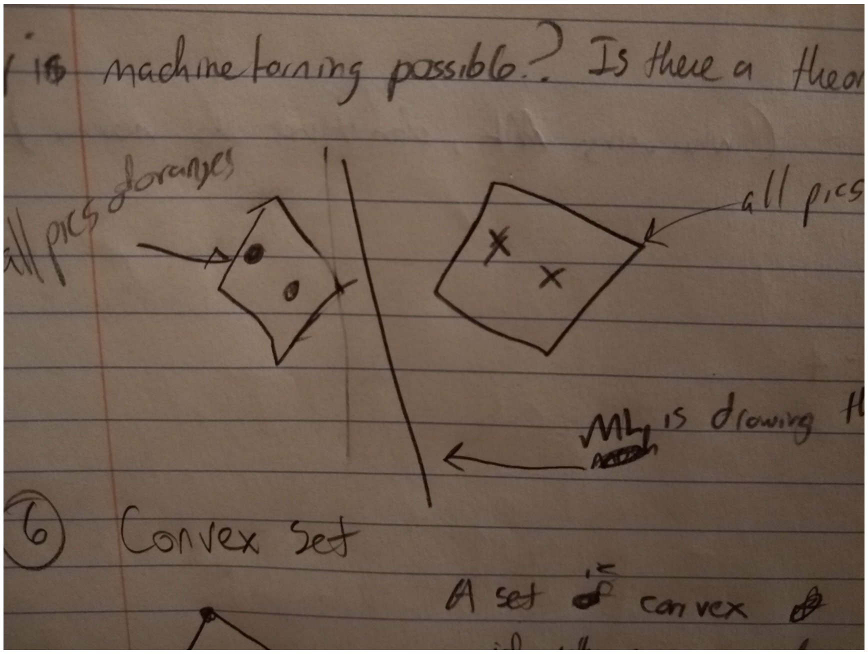
   Sphere: $||\vec{x}||_1 = | - 4| + |3| = 7$

3. Infinity

   Sphere: $||\vec{x}||_\infty = Max| - 4|, |3| = 4$

1.3 Optimization

Why is Machine Learning Possible? Is there a theoretical guarantee?



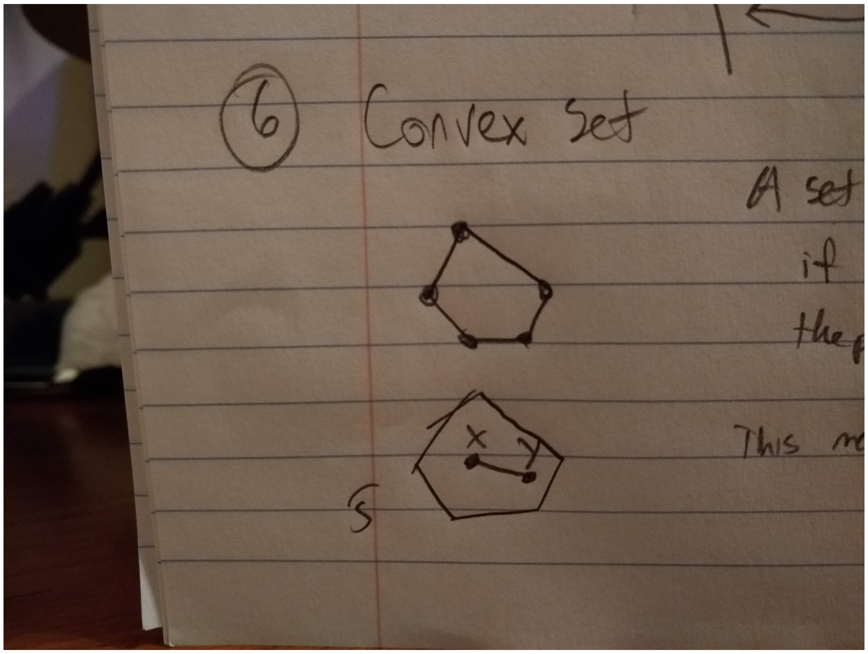
Imagine A is the set of all dogs and B is the set of all Cats

If the sets are convex and do not overlap, there exists a line between them which acts as a divider for determining whether a new pic belongs in A or B.

1.4 Convex Set

A set is convex if whenever X and Y are in the set, then for $0 \leq t \leq 1$ the points $(1-t)x + ty$ must also be in the set.

- #+ATTR$_{\text{LaTeX}}$: scale=0.5



1.5 Separating Hyper-plane Theorem

Let C and D be 2 convex sets that do not intersect. i.e. the sets are **disjoint**.

Then there exists a vector $\vec{a} \neq 0$ and a number b such that.

$$a^T x \leq b \forall x \in C$$

and

$$a^T x \geq b \forall x \in D$$

The Separating Hyper-plane is defined as $x\colon a^T x = b$ for sets C, D.
This is the theoretical guarantee for ML

vector a is perpendicular to the plane b.

2 Why Separating Hyperplane Theorem & Subspace Segmentation Example (2020/04/07)

2.1 Why is Separating Hyper-plane Theorem true?

2.1.1 Math Background

Let $x = d - c,\ y = u - d$

1. Square of the \$l₂\$-norm is the inner product

$$\|x\|_2^2 = \langle x, x\rangle = x^T x$$

$$(d-c)^T(d-c) = \|d-c\|_2^2$$

2. Expansion of Vectors

$$\begin{aligned} &\|x+ty\|_2^2 \\ =&\langle x+ty, x+ty\rangle \\ =&\|x\|_2^2 + 2t\langle x, y\rangle + t^2\|y\|_2^2 \end{aligned} \tag{12}$$

3. Derivative of vector products

$$\frac{d}{dt}(\|x+ty\|_2^2) = 2x^T y + 2ty^T y$$

$$\frac{d}{dt}(\|x+ty\|_2^2)|_{t=0} = 2x^T y$$

$$\frac{d}{dt}(\|d+t(u-d)-c\|_2^2)|_{t=0} = 2(d-c)^T(u-d)$$

2.1.2 Separating Hyper-plane Theorem

C, D are convex disjoint sets. Thus there exists a vecto $\vec{a} \neq 0$ and a number b such that

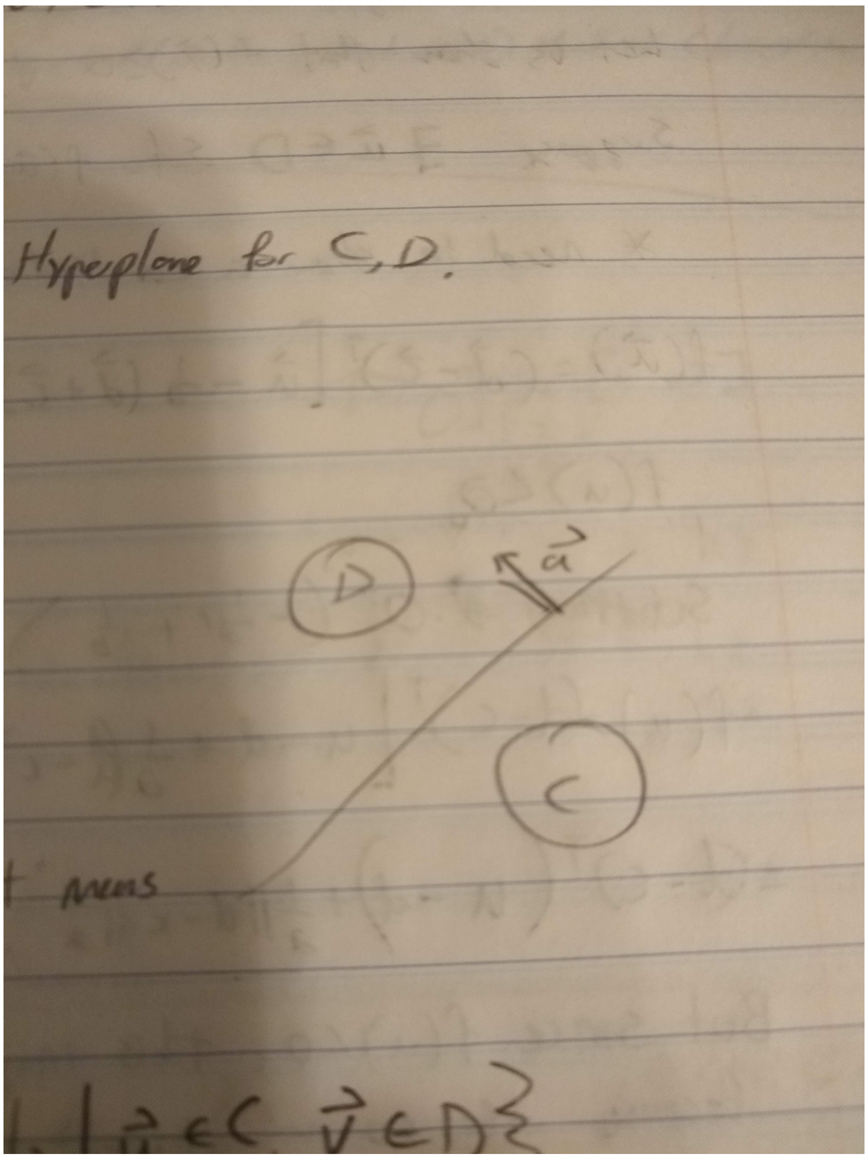
$$a^T x \leq b, \forall x \in C$$

and

$$a^T x \geq b, \forall x \in D$$

$x : a^T x = b$ is the separating hyper-plane for C,D.
When $b = 0$, then inconclusive answer.

2.1.3 Why is it true?



$$
\begin{aligned}
\vec{a}^T\vec{x} &\leq b \text{ on side C} \\
\vec{a^T}\vec{x} &\geq \text{ on side D}
\end{aligned}
\tag{13}
$$

Goal: Prove $\vec{a}$ exists as that means a separating hyperplane exists.

$$dist(C, D) = min\|\vec{u} - \vec{v}\|_2|\vec{u} \in C, \vec{v} \in D = \|\vec{c} - \vec{d}\|_2$$

where $\|\vec{u} - \vec{v}\|_2$ is the euclidean distance.
Let $\vec{a} = \vec{d} - \vec{c},\ b = \frac{1}{2}(\|\vec{d}\|_2^2 - \|\vec{c}\|_2^2)$
We will show that

$$f(\vec{x}) = a^T x - b$$

has the property that

$$f(\vec{x}) \leq 0,\ \forall \vec{x} \in C$$

and

$$f(\vec{x}) \geq 0,\ \forall \vec{x} \in D$$

Note: $(\vec{d} - \vec{c})^T \frac{1}{2}(\vec{d} + \vec{c}) = \frac{1}{2}(\|\vec{d}\|_2^2 - \|\vec{c}\|_2^2)$
What does showing something mean?
Let us show that $F(\vec{x}) \geq 0,\ \forall \vec{x} \in D$ (Argue by Contradiction)
Suppose $\exists \vec{u} \in D$ such that $f(\vec{x}) < 0$

$$f(\vec{u}) = (\vec{d} - \vec{c})^T[\vec{u} - \frac{1}{2}(\vec{d} + \vec{c})] = (\vec{d} - \vec{c})^T \vec{u} - \frac{1}{2}(\|\vec{d}\|_2^2 - \|\vec{c}\|_2^2)$$

Subtract 0

$$f(u) = (d - c)^T[u - d + \frac{1}{2}\|d - c\|]$$

$u - \frac{1}{2}d + \frac{1}{2}c$
$u - d + \frac{1}{2}d - \frac{1}{2}c$

$$f(u) = (d - c)^T(u - d) + \frac{1}{2}\|d - c\|_2^2$$

Now we observe that

$$\frac{d}{dt}(\|d + t(u - d) - c\|_2^2)|_{t=0} = 2(d - c)^T(u - d) < 0$$

and so for some small $t > 0$,

$$\|d + t(u - d) - c\|_2^2 < \|d - c\|_2^2$$

$g'(t) < 0$ means decreasing. Thus $g(t) < g(0)$.
Let's call point $p = d + t(u - d)$
Then

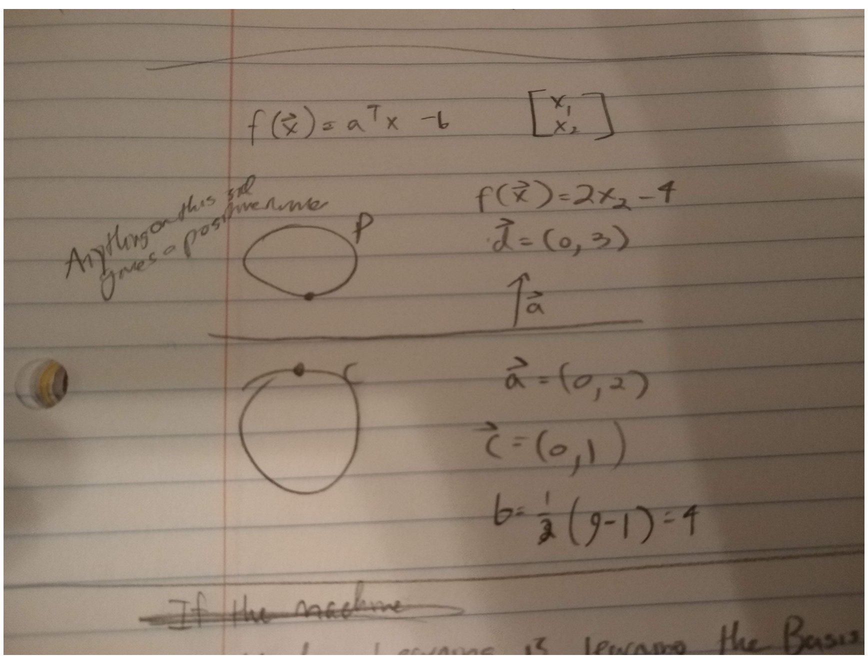
$$\|p - c\|_2^2 < \|d - c\|_2^2$$

This is a contradiction. Both d and u are in set D. Thus by the definition of convexity, $p = (1 - t)d + tu$

D is a convex set so p must also be in D. This situation is impossible since d is the point in D that is closest to c.

2.1.4 Example

Let $f(\vec{x}) = a^T x - b$



2.2 Subspace Segmentation Example

Machine Learning is learning the Basis A. If we can deduce that a vector $\vec{x}$ is a linear combination of A, then a vector is a subspace of Basis A and we know that it belongs to A.

$$V_1 = (x, y, z) \in R^3 : z = 0$$
$$V_2 = (x, y, z) \in R^3 : x = 0, y = 0$$

V_i is the affine variety (it is also a Ring, Module)
Apply a Veronase map with degree 2 to lift up from 3 to 6 dimensions.

$$\nu_n \begin{bmatrix} x \\ y \\ z \end{bmatrix} = \begin{bmatrix} x^2 \\ y^2 \\ z^2 \\ xy \\ xz \\ yz \end{bmatrix}, \nu_n : R^3 \to R^6$$

$$\begin{aligned} z_1 = (3, 4, 0), z_2 = (4, 3, 0), \\ z_3 = (2, 1, 0), z_4 = (1, 2, 0), \\ z_5 = (0, 0, 1), z_6 = (0, 0, 3), z_7 = (0, 0, 4) \end{aligned} \tag{14}$$

Plug the sample points into the Veronase map to produce a matrix L

$$L = \begin{bmatrix} 9 & 16 & 4 & 1 & 0 & 0 & 0 \\ 16 & 9 & 1 & 4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 9 & 6 \\ 12 & 12 & 2 & 2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \in R^{6 \times 7}$$

solve for $\vec{c}$, where $\vec{c}^T L = \vec{0}$

$$\vec{c_1} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 0 \end{bmatrix}, \vec{c_2} = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \end{bmatrix}$$

Rank(L) = 4 (since there are 4 linearly independent rows)

$$\begin{aligned} q_1(X) =& \vec{c}^T \nu_n(X) \\ =& xz \\ q_2(X) =& \vec{c}_2^T \nu_n(X) \\ =& yz \end{aligned} \tag{15}$$

We have:

$$\begin{aligned} q_1(X) &= xz \quad V_1 = (z = 0) \\ q_2(X) &= yz \quad V_2 = (x = 0, y = 0) \end{aligned} \tag{16}$$

Observe:
$V_1 \cup V_2 = ((x, y, z) \in R^3 : q_1(X) = 0, q_2(X) = 0)$
Construct the Jacobian matrix
J(Q)(X) $= \begin{bmatrix} \frac{\partial q_1}{\partial x} & \frac{\partial q_1}{\partial y} & \frac{\partial q_1}{\partial z} \\ \frac{\partial q_2}{\partial x} & \frac{\partial q_2}{\partial y} & \frac{\partial q_2}{\partial z} \end{bmatrix} = \begin{bmatrix} z & 0 & x \\ 0 & z & y \end{bmatrix}$

1. When $z = z_1 = (3, 4, 0), J(Q)(z_1) = \begin{bmatrix} 0 & 0 & 3 \\ 0 & 0 & 4 \end{bmatrix}$

   When $z = z_3 = (2, 1, 0),\ J(Q)(z_3) = \begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & 1 \end{bmatrix}$

   The right null space of $J(Q)(z_1)$ has basis $\vec{b}_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$, $\vec{b}_2 = \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}$

2. When $z = z_5 = (0, 0, 1),\ J(Q)(z_5) = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}$

   When $z = z_7 = (0, 0, 4),\ J(Q)(z_7) = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 4 & 0 \end{bmatrix}$ The right null space of $J(Q)(z_5)$ has basis $\vec{b} = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}$

$C = [\vec{c}_1 | \vec{c}_2]$