Foundations of Data Science and Machine Learning

SUMMARY ASSIGNMENT
Week 01
(CS5014-W-2021)

by

Devansh Singh Rathore (111701011)

under the guidance of

Prof. Deepak Rajendraprasad

INDIAN INSTITUTE
OF TECHNOLOGY
PALAKKAD

Contents

Chapter 1

Data to Vectors

1.1 How to convert data to vector?

→ Unstructured data → $\mathbb{R}^n$, where n is the number of parameters.

→ Example: Image of Dimensions (n x m x 3) with Pixel(i, j, k) $\in \mathbb{R}^{n.m.3}$

→ Example: graph → adjacency matrix can also be seen as vectorization of graph data.

→ Web Page Text Example: If the words of dictionary can be represented as $(w_1, w_2, ...w_n)$. Let x_i represent the frequency of word w_i in the given text, then the **word frequency vector** can be represented as $(x_1, x_2, ...x_n)$.

→ Lesser the distance between two different vectors, more is the similarity between their respective data points.

→ Typical ways in which a vector representation for a dataset is selected →

1. Domain Expert Driven - **Cepstral coefficient**
2. Don't Care - Choose any possible way. eg. in CNN, NN, etc.
3. Algo. Assisted - use Machine Learning Algorithms to find best suitable vector representation. eg. word2vec.

1.2 Why to convert data to vector?

Vector allows different applications such as -

→ Distance - similarity

→ Angle / Innerproduct

→ Separators (linear / non-linear) - eg. email spam classifier

$\rightarrow$ Geometry (Hulls / Boxes)

$\rightarrow$ Subspaces

$\rightarrow$ Algebra - eg. addition of vectors

$\rightarrow$ Limits

$\rightarrow$ Topology

Chapter 2

Linear Separators & Convex Hulls

2.1 Linear Separability

Definition 1: Two sets G and B of points in $\mathbb{R}^n$ are said to be **linearly separable** if there is a hyperplane l such that all points in G lie to one side of l, while all points in B lie to the other side of l.

where l is:

- in $\mathbb{R}^2$, $ax + by = c$
- in $\mathbb{R}^3$, $ax + by + cz = d$
- in $\mathbb{R}^n$, $a_1x_1 + a_2x_2 + ...a_nx_n = b$ or $\sum_{i=1}^n a_ix_i = b$ or $< a, x >= b$, where a = $(a_1, a_2, ...a_n)$ & x = $(x_1, x_2, ...x_n)$. **a is perpendicular to l**.

→ On one side of hyperplane l, $< a, x >> b$, while on the other side of l, $< a, x >< b$.

Definition 2: Two sets G and B of points in $\mathbb{R}^n$ are said to be **linearly separable** if there is a vector $a \in \mathbb{R}^n$ and a scalar $b \in \mathbb{R}^n$ s.t.

1. $\forall x \in G, < a, x >> b$, and
2. $\forall x \in B, < a, x >< b$

Here, the hyperplane defined by $< a, x >= b$ is called the **separating hyperplane**.

Definition 3: A set $S \subseteq \mathbb{R}^n$ is said to be a **convex set** if for every two points $x, y \in S$, the entire line segment connecting x and y is inside S.

Definition 4: $S \subseteq \mathbb{R}^n$ is said to be a **convex set** if for every two points $x, y \in S$ and $\forall \alpha \in [0, 1]$, $z = \alpha x + (1 - \alpha)y \in S$.

Definition 5: For a set $G \subseteq \mathbb{R}^n$, the **convex hull** of G is the smallest (minimal) convex set containing all the points of G. In the rubber band, pin example, the convex hull is represented by the rubber band boundary.

Definition 6: Convex Hull is a set $H_G \subseteq \mathbb{R}^n$ s.t.

1. $G \subseteq H_G$
2. H_G is convex
3. No convex proper subset of H_G contains all of G.

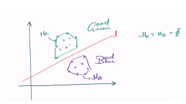
Theorm 2.1: Two sets of points $G, B \subseteq \mathbb{R}^n$ are linearly separable if and only if their convex hulls are disjoint. i.e. $H_G \cap H_B = \phi$.

Proof: $H_G \cap H_B = \phi \implies \exists (a, b)$ s.t. $\forall x \in G, < a, x >> b$ and $\forall x \in B, < a, x >< b$.

We will try to find out smallest vector a which is connecting H_G to H_B, which appears to be normal to l (if it exists).

So, a = the smallest vector in the set $\{x - y : x \in H_G, y \in H_B\}$

$\forall x \in G \cup B$, compute $< a, x >$. The values of $< a, x >$ for $x \in G$ are close to each other and separate from those of $x \in B$.



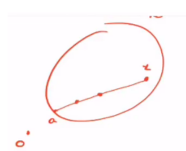
Claim: $\forall x \in G, \forall x \in B$, we have

$\rightarrow < x - y, a > \geq ||a||^2$

$\rightarrow < x, a > - < y, a > \geq ||a||^2$

Observation: $D = H_G - H_B$ is convex set. **(Proof exercise)**

Proof of Claim: If a is the smallest vector in convex set D, then $\forall z \in D$,



Part 1: $< z, a >>= ||a||^2/2$
let r be a point in line segment a to z. i.e. $r = \alpha z + (1 - \alpha)a \in D$
since a is closer to origin, $||a||^2 <= ||r||^2$
$= (\alpha z + (1 - \alpha)a)^2$
$= \alpha^2||z||^2 + (1 - \alpha)^2||a||^2 + 2\alpha(1 - \alpha) < z, a >$
$< z, a >>= (||a||^2 - (1 - \alpha)^2||a||^2 - \alpha^2||z^2||)/2\alpha(1 - \alpha)$
$= ((2 - \alpha)||a||^2 - \alpha||z||^2)/2(1 - \alpha)$, since a!=0
$>= ((2 - \alpha)||a||^2 - ||a||^2)/2(1 - \alpha)$, since $-\alpha||z||^2 >= -||a||^2$, hence $\alpha <= ||a||^2/||z||^2$
$= ||a||^2/2$
Now, if we choose $\alpha <= (1/2)(||a||^2/||z||^2)$, we get ¡z,a¿ ¿= $= ||a||^2$