# SMAI-M20-04:Appreciating Data in High Dimension

C. V. Jawahar

IIIT Hyderabad

August 17, 2020

# Recap: Administrative

1. Scope and Course Plans:
   - Scope, Course Plans.
   - L01: https://www.dropbox.com/s/ltmyx9y15hxnmm8/L1.pdf?dl=0
   - L02: https://www.dropbox.com/s/7216t0zl39xgmeq/L2.pdf?dl=0
   - L03: https://www.dropbox.com/s/sw1mosvwsiv85ny/l3.pdf?dl=0
2. Logistics:
   - Started to use "shiksha" for questions.
   - Regular HW:
     - Approximately 3 questions will be posted on every leccture day. (Roughly 1 Q each of 1 pt, 2pt and 3pt) 1st set expected to be on Shikha today.
     - Submit within a week.
     - You need to do only 80%. (i.e., 160 points out of 200 expected.)
     - Use "TS&GH" channel for any difficulty. Don't wait for the last day.
   - Class Review Questions:
     - Will move to Shiksha systematically. Today in lecture slides.
     - First questions + answers on google forms (this week).
     - Later fully on Shiksha. (Need to understand the load/any issues).

# Summary: Till Now

1. Representation as a vector in $R^d$
2. Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
   - Notion of Training and Testing
3. Feature Transformation as a useful trick:
4. $\mathbf{x}' = \mathbf{W}\mathbf{x}$
   - Dimensionality Reduction
5. Two Simple Classification Schemes:
   - Nearest Neighbour Algorithm
   - Linear Classification
   - — $sign(\mathbf{w}^T\mathbf{x})$; Either +ve or -ve.
   - — Many ways to extend to more than 2 classes
6. Performance Metrics:
   - Classification: Accuracy, TP/FP etc., Confusion Matrix
   - Ranking: Precision, Recall, F-Score, AP

# This Lecture: Appreciating Data

- Data could come from a physical process
  - Data is not some random numbers.
  - Structure of the data allows us to learn
  - Reasonable assumption of Multivariate Gaussian
- Geometry of Representation:
  - Lines, Planes, Hyperplanes.
  - Geometry in high dimension.
  - What does it mean by

    $$\mathbf{w} \leftarrow \mathbf{w} + \nabla \mathbf{w}$$

    What is the geometric interpretation?
- Practical Challenges in High Dimension
  - Too many parameters to learn. Lot more samples required in High Dimension.
  - Computational and Practical Advantages.
  - Need of dimensionality reduction.

# Blank

# Discussions Point - I

Q: Consider a linear transformation $d \rightarrow d$ (i.e., $\mathbf{W}$ is a square matrix)

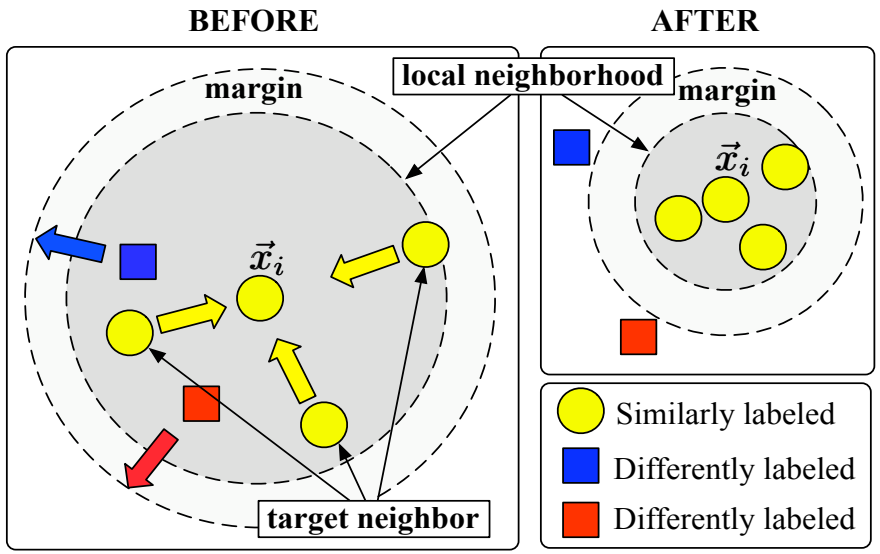
$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

We use a K-NN algorithm (the same K and distance as Euclidean distance) in original and new space.

- Will the performance (say accuracy) of the algorithm be same in both the space for any $\mathbf{W}$? i.e., with $\mathbf{x}$ and $\mathbf{x}'$? (Discuss)
- If no, what should be the condition on $\mathbf{W}$ to guarantee that?

# Discussions Point -Ib (Advanced)

If we can lean **W** for a K-NN so that performance improves, what should be our desirability?




Ans (Fig and Ans from the LMNN paper[1]): For each sample, we desire

- Its K (=3) target neighbors lie within a smaller radius after transforming with **W**
- Differently labeled inputs lie outside this smaller radius, with a margin of at least one unit distance.

~~How do we explain these requirements?~~

[1] Read initial sections of: "Distance Metric Learning for Large Margin Nearest Neighbor Classification", NIPS 2006.

# Blank

# Blank

# Discussions Point -II

An SMAI student (Raju) implements $K - NN$ and tested it on a popular data set. He conducted an experiment to vary $K$ (say from 3 to 15) and plot the performance.

Q1:

- Will he see a systematic increase in accuracy with K?
- Will he see a systematic decrease in accuracy with K?
- Will he see a systematic increase followed by a systematic decrease?

Q2: Can you help Raju in finding the best $K$?

# Blank

# Discussion Point - III

Over years, we have figured out HYD temperature in Jan and May are $\mathcal{N}(23, \sigma^2)$ and $\mathcal{N}(33, \sigma^2)$ (i.e, Normal, mean 23 and 33; Variance the same).

Q: We have 100 days of data from Jan and 100 days from May, but not labelled. We want a classifier as:

"If temp $< \theta$, then Jan else May"

What should be the value of $\theta$ intuitively?

- 28 $(= \frac{23+33}{2})$
- Less than 28.
- More than 28.

Why?

# Blank

# Blank

# Review Question - I (one, none or more correct)

Confusion matrix is:

(a) Square (b) Always Diagonal (c) Can never be diagonal (d) Can be diagonal (e) Always Symmetric (f) Can never be symmetric (g) Can be Symmetric

# Blank

# Review Question - II (one, none or more correct)

$$\frac{TP}{P}$$

is known as:
(a) Accuracy (b) Precision (c) Recall (d) None of the above

# Blank

# Review Question - III (one, none or more correct)

A disease occurs with a probability of 0.4 (i.e., it is present in 40% of the population). You have a test that detects the disease with a probability 0.6, and produces a false positive with probability of 0.1. What is the (posterior) probability that the test comes back positive.

Hint: $S$ is the event that you are sick; $P$ is the event that test comespositive.

$$P(S|P) = \frac{P(P|S)P(S)}{P(P)} = \frac{P(P|S)P(S)}{P(P|S)P(S) + P(P|\bar{S})P(\bar{S})}$$

(a) 0.6 (b) 0.7 (c) 0.8 (d) 0.9 (e) 0.95

# Blank

# Review Question - IV (one, none or more correct)

Two SMAI students (Raju and Sheela) worked on the same problem with the same measurements/features and samples, except that their feature orderings were different. (i.e., $\mathbf{x}$ and $\mathbf{x}'$ were permutations.) Identify correct statement(s).

- (a) Both got the same accuracy with KNN (same K and Eucli. distance)
- (b) Both got different accuracy with KNN (same K and Eucli. distance)
- (c) Their confusion matrices were different i.e., elements (cells) were swapped.
- (d) Both had the same Covariance Matrices (Hint: $\Sigma = \frac{1}{N}\sum_{i=1}^{N}[\mathbf{x}_i - \mu][\mathbf{x}_i - \mu]^T$)
- (e) Both had covariance matrices of the same Rank.
- (f) Both had covariance matrices where cells (elements) were swapped.

# What Next:?

- Logistics: Make sure that you can use "Shiksha" by Friday. No super hurry.
- Use Channels to Post Queries/Discussions
- Emails at: smai.m2020@gmail.com

1. Revise: Rank of a matrix. Interpretation of Rank.
2. Revise: Bayes Theorem
3. Revise: Eigen Values and Eigen Vectors

- Office Hour This week: Any queries/Doubts on Chapter 2 and 3 of the "Mathematics for Machine Leaning"