SMAI-M20-06: Data, Distances and Learning

C. V. Jawahar

IIIT Hyderabad

August 21, 2020

Recap:

- Two problems of interest:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
  - Learn Feature Transformation as a step to find useful representations.
- Three Classification Schemes:
  - Nearest Neighbour Algorithm
  - Linear Classification
  - Decide as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2
- Performance Metrics:
  - Classification: Accuracy, TP/FP etc., Confusion Matrix; Ranking: Precision, Recall, F-Score, AP
- Supervised Learning:
  - Notion of Training and Testing
  - Notion of Loss Function
  - Role of Optimization

This Lecture:

- Knowing Matrices Better:
  - Rank, Determinant, Trace,
  - Eigen Values and Eigen Vectors
- Supervised Learning
  - Need to generalize
  - Difficulty due to overfitting
- Comparison
  - Distance vs Similarity
  - Samples in R^d
  - Comparison of Sets
  - Comparison of probabiliy distributions

Blank

Questions? Comments?

Comment

What is a good representative for a set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$
Ans: Some one who is close to all.!! Let it be $\mathbf{y}$

$$\min_{\mathbf{y}} \sum_{i=1}^{N} [\mathbf{x}_i - \mathbf{y}]^T [\mathbf{x}_i - \mathbf{y}]$$

$$\min_{\mathbf{y}} \sum_{i=1}^{N} [\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{y}^T \mathbf{x}_i + \mathbf{y}^T \mathbf{y}]$$

Differentiating wrt to $\mathbf{y}$ and equating to zero [1]

$$\sum_{i=1}^{N} -2\mathbf{x_i} + 2\sum_{i=1}^{N} \mathbf{y} = 0$$

or

$$\mathbf{y} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

[1] Tom Minka, "Old and New Matrix Algebra Useful for Statistics" (Read or refer)

Blank

Discussions Point -I

Consider two ways of comparing two samples $\mathbf{x}$ and $\mathbf{y}$

1. Weighted Euclidean Distance with a Symmetric Positive Definite (PD) matrix $\mathbf{W}$

$$[\mathbf{x} - \mathbf{y}]^T \mathbf{W} [\mathbf{x} - \mathbf{y}]$$

2. We know that PD matrix $\mathbf{W}$ can be decomposed as $\mathbf{L}\mathbf{L}^T$ (popularly known as Cholesky decomposition).
   - We transform the individual samples as:

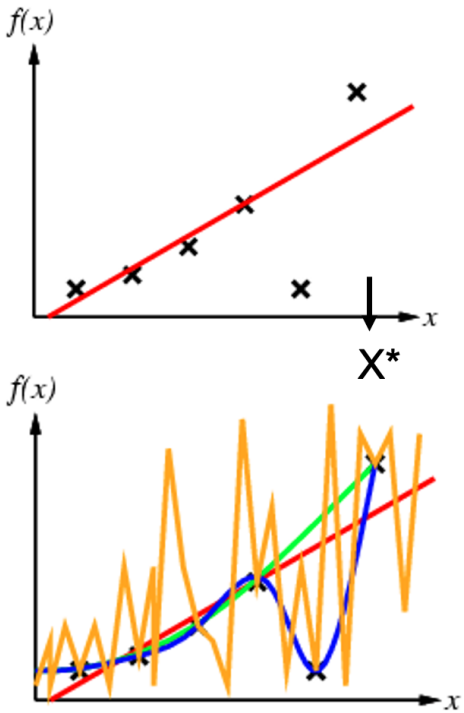
$$\mathbf{x}' = \mathbf{L}^T \mathbf{x}$$

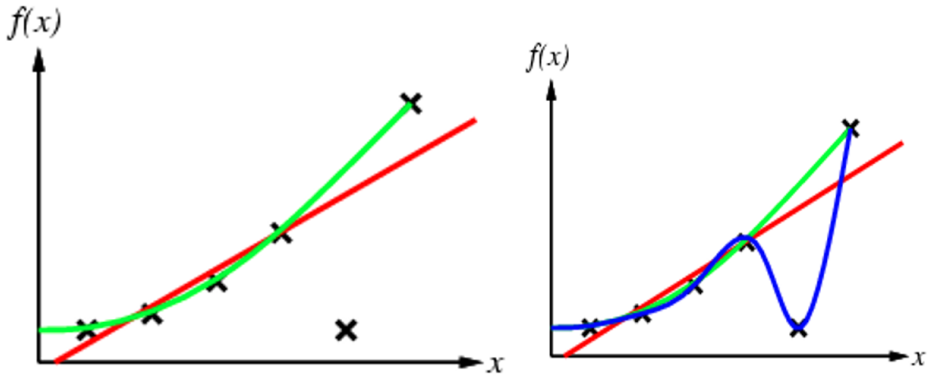
   - Compute the Euclidean distaance between $\mathbf{x}'$ and $\mathbf{y}'$
   - Show that both ways of computing give the same distance.
   - Which one is computationally attractive and when?[2] For example, consider a simple formulation of retrieving K nearest neighbours for a given query $\mathbf{q}$ from a database of N samples.

[2] Connect to the LMNN Paper/method, assuming you had read.

Blank

Discussion Point - II





Given six "training" (x_i, y_i) pairs, find the y corresponding to the new "test" x*

Which curve is the best?

Blank

Discussion Point - III

- We know the $d \times N$ data matrix $\mathbf{X}$ with every column as data elements. Show that $\mathbf{A} = \mathbf{X}\mathbf{X}^T$ is Symmetric. Show that $\mathbf{A}$ is PSD. [3]
- If $\mathbf{A}$ is real symmetric PSD matrix, we know $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathbf{T}$, then what is $\mathbf{V}^\mathbf{T}\mathbf{A}\mathbf{V}$? What is this process called?

[3] A matrix A is PSD if $z^T Az \geq 0 \ \forall z \in R^d$.

Blank

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

If

$$Ax = \lambda x$$

then What are the eigen values and eigen vectors of A^2
i.e., Find:

$$A^2x = ?$$

- (a) λ, x
- (b) λ^2, x
- (c) λ, $2x$
- (d) λ^2, $2x$
- (e) none of the above

Blank

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

Consider

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

$\mathbf{W}$ is constructed as below:

- We start with a $d \times d$ identity matrix
- We randomly permute (rearrange) the columns
- We remove half of the rows and create a $\frac{d}{2} \times d$ matrix $\mathbf{W}$

The process of creation of new representation is:

- (a) Feature subset selection; A random subset of the original features will be in $\mathbf{x}'$
- (b) Feature extraction; New features are linear combination of old ones, and not really a subset.
- (c) Dimensionality Reduction; New representation has smaller dimension than the original one.
- (d) This can not be done since these operations are illegal (or mathematically not defined).

Blank

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

Consider three sets

$$A = \{1, 3, 4, 5, 6, 7, 8\}$$

$$B = \{2, 4, 6, 8\}$$

$$C = \{1, 2, 3, 4, 5\}$$

We know Jacard index ($J = \frac{|A \cap B|}{|A \cup B|}$) as a good measure of similarity. Let us use $1 - J$ as a distance.

Does $1 - J$ obey triangular inequality for this set? [4]

(Hint Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y)$)

- (a) YES
- (b) NO
- (c) Triangular inequality is not applicable for this problem.
- (d) Can not be computed.

[4] Advanced (optional): How do we show this for any general three sets?

Blank

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

Consider the problem of feature transformation as

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

If $\mathbf{x}_i \in R^2$ and $\mathbf{W}$ is 2×2 matrix with rank as 1, then

the new points $\mathbf{x}'_i$

- lie on a line in 2D
- are also R^2
- undefined
- One coordinate (dimension) of all the $\mathbf{x}'_i$ will be always the same
- all points in 2D will collapse into a single point. (i.e., $\mathbf{x}'_i = \mathbf{x}'_j$ for all i, j)
- none of the above.

Blank

Review Question - V (one, none or more correct)

In a TV Game show, a contestant selects one of three doors; behind one of the doors there is a prize,and behind the other two there are no prizes. After the contestant selects a door, the game-show host opens one of the remaining doors, and reveals that there is no prize behind it. The host then asks the contestant whether the you want to SWITCH their choice to the other unopened door, or STICK to their original choice.
[5]

What should we advise? What is the prob. of win if the candidate switch:

- (a) $\frac{1}{3}$ since all the doors are equally likely. Don't switch
- (b) $\frac{1}{2}$ since there are only two left, both are equally likely, no advantage in switching.
- (c) $\frac{2}{3}$. Prefer switching. Bayes says so.
- (d) $\frac{1}{3}$. Don't switching. Bayes says so.
- (e) None of the above.

[5] A very popular problem on internet from khan academy to mit lecture notes!. Appreciate the role of evidence, specially if the answer is not intuitive.

Blank

What Next: Two Sessions?

- Eigen Values/Vectors, SVD, Rank and Data Matrix
- More into Supervised Learning and the associated issues
- Bayesian Optimal Classification