SMAI-M20-05: Features, Data and Learning

C. V. Jawahar

IIIT Hyderabad

August 19, 2020

Recap

1. Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data:
   - Representation as a vector in R^d
   - Learnable parameters $\mathbf{W}$
   - Notion of Training and Testing
2. Feature Transformation as a useful trick:
   - $\mathbf{x}' = \mathbf{W}\mathbf{x}$
   - Dimensionality Reduction
3. 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
4. Performance Metrics:
   - Classification: Accuracy, TP/FP etc., Confusion Matrix
   - Ranking: Precision, Recall, F-Score, AP

This Lecture: Appreciating Data

- Three Video Clips:
  1. Feature Selection and Feature Extraction
  2. "Features as a Vector" to "Data as a Matrix"
  3. Formulating Supervised Learning Task
- Progressing from "Raw" representation to "Useful features":
  - Examples of "raw" data; Eg. Pixels in an image
  - Subset selection
  - Linear transformation; including Dim. Reduction
  - Non-linear Transformations (more later)
  - Learning the features Eg. Deep Embeddings (more later)
- Data as matrix
  - Data is not "random"
  - "Structure in the data" leading to "data lie in a subspace".
  - Eg. Why are we not seeing samples with any $\frac{ht}{wt}$?

Blank

Questions? Comments?

Blank

Discussions Point - I

In supervised learning, we use the training data $\mathcal{D}_{Tr}$ to develop a solution and then evaluate it on the test data $\mathcal{D}_{Te}$

Q: Why is that a solution optimized only on training data as

$$\min_{\mathbf{W},} \mathcal{L}(\mathbf{W}, \mathbf{x}) \quad \mathbf{x} \in \mathcal{D}_{Tr}$$

works reasonably well on the test data also?

1. We believe in luck.
2. Once it is trained on $\mathcal{D}_{Tr}$, it then works well on any data.
3. Actually, we use $\mathcal{D}_{Te}$ also for training. Eqn. is incorrect.
4. What we care is actually the performance on the $\mathcal{D}_{Tr}$; not $\mathcal{D}_{Te}$.
5. It is some magic. No explanation found till today.

Blank

Discussions Point -II

Consider a problem where the original representation d (say 100) and N (say 1000) samples. Problem is binary classification. (Class 'A' Vs. Class 'B'). Assume each of these d features are in the range $[0, 10]$.

Q: We want to "select" top 10 features based on a "quality measure" (how good this feature could be for the classification) of the feature.

We rank features based on:

1. We rank each feature j based on $|\mu_A^j - \mu_B^j|$
2. We rank each feature j based on $\frac{(\sigma_A^j)^2 + (\sigma_B^j)^2}{2}$
3. We compute the probability of A and B in 10 intervals (say $[0, 1], [1, 2] \ldots [9, 10]$) as P_i and compute entropy for each feature as:

$$H = \sum_{i=1}^{10} P_i^A \log P_i^A + \sum_{i=1}^{10} P_i^B \log P_i^B$$

Blank

Discussion Point - III

Without numerically computing, what is the rank of the matrix? Show:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix}$$

1. 1
2. 2
3. 3
4. 4

Hint:

$$\begin{bmatrix} 1 & 2 & 3 \\ 4 & 5 & 6 \\ 7 & 8 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 2 & 3 \\ 1+3 & 2+3 & 3+3 \\ 1+6 & 2+6 & 3+6 \end{bmatrix}$$

Blank

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

We know that the rank of a 3×3 matrix formed by first 9 numbers arranged sequentially is 2.

What is the rank of a 5×5 matrix formed by first 25 numbers arranged sequentially?

(a) 1 (b) 2 (c) 3 (d) 4 (e) 5 (f) none of the above

Blank

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

A certain test for disease is known to have True positive of 0.6 and False Positive of 0.1.
A population of 100 people (where 60 of them are infected) undergoes this test.

What could be the confusion matrix?

(a) $\begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix}$ (b) $\begin{bmatrix} 0.6 & 0.4 \\ 0.9 & 0.1 \end{bmatrix}$ (c) $\begin{bmatrix} 0.6 & 0.2 \\ 0.1 & 0.3 \end{bmatrix}$ (d) $\begin{bmatrix} 0.58 & 0.42 \\ 0.15 & 0.85 \end{bmatrix}$
(e) None of the above

Blank

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

Covariance Matrix:

- (a) Can neve be diagonal.
- (b) Can be diagonal
- (c) Always Positive Semi Definite
- (d) Always full rank
- (e) Never full rank.
- (f) Always Symmetric
- (g) Not guaranteed to be symmetric

Blank

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

Let $\mu = \sum_{i=1}^{N} \mathbf{x}_i$ be the mean of $\mathbf{x}_1, \ldots, \mathbf{x}_N$. $\mathbf{x}_i \in R^d$

- (a) μ is also $\in R^d$
- (b) μ is always one of the N samples.
- (c) μ can not be one of the N samples
- (d) μ can be one of the N samples.
- (e) μ is equidistant from all the N samples
- (f) μ has the least sum of square error from all the samples (i.e., μ is $\arg\min_{\mathbf{y}} \sum_{i=1}^{N} ||\mathbf{x}_i - \mathbf{y}||_2^2$)

Blank

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

You are planning a picnic today, but the morning is cloudy

- 50% of all rainy days start off cloudy.
- But cloudy mornings are common (about 40% of days start cloudy)
- This is usually a dry month (only 3 of 30 days tend to be rainy, or 10%)

What is the chance of rain during the day?

(a) 10% (b) 12.5% (c) 15% (d) $> 20\%$ (e) $< 20\%$

Hint: Bayes Rule:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

Blank

What Next: Two Sessions?

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