SMAI-M20-L08:SVD; MLE and MSE

C. V. Jawahar

IIIT Hyderabad

August 26, 2020

Announcements:

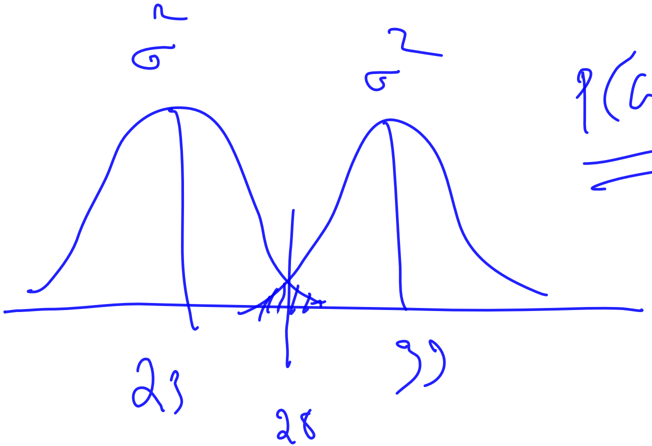
- Class Review Questions (wish to do in the lecture session:5-10 mins):
  - Five objective questions;; An average of 1 to 1.5 min per questions;
  - Submission (QA) will be now on Shiksha.
  - Enough buffer for missing lecture sessions/connectivity (only 80%)
  - Quick clarifications in the class; detailed doubts/queries in an OH.
- Home works:
  - Regular (we are lagging behind), handwritten or some programming.
  - Only 80% is required. Buffer for connectivity/personal schedules.
  - Assume by now: Comfortable with python and jypyter notebooks.
- MS Teams/Communication/Connectivity:
  - use smai.m2020@gmail.com for direct communication
  - use channels to post queries
  - avoid submission closest to the deadlines.
- Office Hours/Queries on:
  1. Chapter 2, 3 and 5 of the book
  2. Class Review Questions: L01-L08
  3. Micro-Lecture Videos: L01-L08

Review Questions: Let us submit in the first 10 mins

1. Numerically computing rank of a 3×3 matrix
2. The system of linear equations $Ax = b$ has?
3. Suppose a disease is prevalent in 1% of the population. Its medical diagnosis is 90% accurate in both directions. Given that a person tested positive, what is the chance, he actually has the disease (rounded to nearest integer)?
4. We know that the optimal classifier for two equally probable (equal Prior probability) classes (days of months) $N(23, \sigma^2)$ and $N(33, \sigma^2)$ is 28.

   If the variance of the second class becomes double, then the the optimial classification threshold will increase or decrease?
5. A man is known to speak truth 2 out of 3 times. He throws a die and reports that number obtained is a four. Find the probability that the number obtained is actually a four.

Blank



$P(\omega_1) = P(\omega_2)$

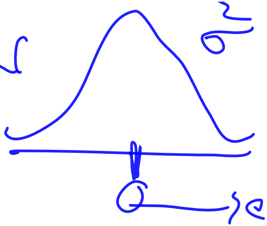
Recap:

- Problem Space:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
    - for classification
    - for regression
  - Learn useful features
    - feature transformations
    - dimensionality reduction
    - feature selection, feature extraction
- Supervised Learning:
  - Notion of Training and Testing
  - Notion of Loss Function and Optimization
  - Need of generalization and Worry of Overfitting
- Classification Algorithms:
  - Nearest Neighbour Algorithm
  - Linear Classification; Linear Regression
  - Decide as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2
  - Performance Metrics
- Mathematical Foundations: Linear Algebra, Probability, Optimization

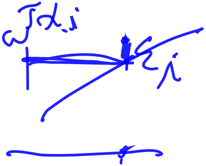
This Lecture:

$$\min_{w} \cancel{\max} \sum_{i=1}^{N} (y_i - w^T x_i)^2$$

λc



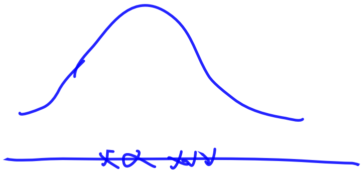
- SVD: Singular Value Decomposition
  - Connect to Eigen Decomposition
  - Connect to Data Matrix
  - Follow ups to come.
- MSE as MLE
  - Appreciate MLE as a general step.
  - Probabilistic interpretation of an intuitive expression.
- Geometry of Gaussians
  - Eigen Decomposition
  - Will lead to PCA.



$$\epsilon_i = y_i - w^T x_i$$

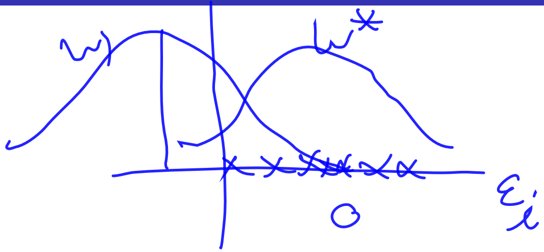
Blank

MLE



Find

μ, σ^2



$$\varepsilon_1 = y_1 - c^T x_1$$

Questions? Comments?

Blank

Discussions Point -I

Consider a situation when we continue to get one sample at a time. We have mean (μ_N) and variance (σ_N^2) computed and available at sample N.

Now we get the $N+1$ sample. How do we compute the new mean? Ans:

$$Ans : \mu_{N+1} = \frac{\mu_N \times N + x_{N+1}}{N+1}$$

$$Y_N = \frac{\sum_{i=1}^{N} x_i}{N}$$

How do we compute σ_{N+1}^2?
Where do we need such "online" computations?

Blank

$\mu = \frac{1}{N}\sum X_i$

$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}(x_i-\mu)^2$

$$\sum_i (x_i-\mu)^2 = \underline{\sum_i x_i^2} + \underline{\sum_i \mu^2} - \underline{\sum_i \mu \cdot x_i}$$

Discussion Point - II

$Q^{-1} = Q^T$ $\quad (Q\Lambda Q^T)^{-1}$ $\quad (Q^T)^{-1}\Lambda^{-1}Q^{-1}$ $\quad Q^{T}\Lambda^{-1}Q^{T}$ $\quad \boxed{Q\Lambda^{-1}Q^T}$

We know that:

- Eigen Decomposition of Symmetric Matrix **S**

$$\mathbf{S} = \mathbf{Q}\Lambda\mathbf{Q}^T = \sum_{i=1}^{n} \lambda_i \mathbf{q}_i \mathbf{q}_i^T$$

- SVD of **A** $n \times n$

rank 1

$$\mathbf{A} = \mathbf{U}\mathbf{D}\mathbf{V}^T = \sum_{i=1}^{n} d_i \mathbf{u}_i \mathbf{v}_i^T$$

UDV^T

1. How do we compute $\mathbf{S}^{-1}$ and $\mathbf{A}^{-1}$
2. If $\mathbf{A} = \mathbf{pq}^T$ is a 3×3 matrix, what is the SVD of **A**?

Blank

(b).

$$A = U D V^{-1} \qquad p \quad 3\times 1$$

$$A^{-1} = (V^{-1})^{-1} D^{-1} U^{-1} \qquad q \quad 3\times 1$$

$$= V D^{-1} U^{-1}$$

$$A = \begin{bmatrix} \bar{p} & \bar{0} & \bar{0} \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} \bar{q}^{T} \\ \bar{0}^{T} \\ \bar{0}^{T} \end{bmatrix} \begin{bmatrix} q_1 & q_2 & q_3 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

$$3\times 3 \qquad 3\times 3 \qquad 3\times 3$$

$$\begin{bmatrix} p_1 & 0 & 0 \\ p_2 & 0 & 0 \\ p_3 & 0 & \end{bmatrix}$$

Discussion Point - III

We are worried about outliers in the regression. Let us give a score γ_i as the "importance" or "confidence" of a sample that this is an inlier. We can now modify the loss/objective as:

$[Y - Xw]^T [Y - Xw]$

$$\frac{1}{N}\sum_{i=1}^{N} \gamma_i \times (y_i - \mathbf{w}^T \mathbf{x}_i)^2$$

$w = (X^T X)^{-1} X^T Y$

- a Write down the objective in matrix form. (Hint: use $\Gamma = Diag(\gamma_i)$). What is the final closed form expression for the $\mathbf{w}$?
- b (Advanced) Consider a two step Itertive algorithm:
  - b1 Assign γ_i as inversely proportional to the distance from the line. (distance $=0 \rightarrow \gamma$ as 1 and high distance $\rightarrow \gamma$ as 0)
  - b2 Compute $\mathbf{w}$ using the closed form expression (Q1).

  If we iterate the above two steps? (i) will it converge? (ii) will it take care of outliers? (Later: Try it out on a toy data of yours [1])

[1] A similar treatment in a different area: read "Sample weighted Clustering Methods", CMA, 2011

Blank

$$[y - Xw]^T [y - Xw]$$

$$[y - [\Gamma] Xw]^T [y - \Gamma Xw]$$

$$n \times n \qquad n \times 1$$

$$[y - Xw]^T \, \Gamma \, [y - Xw]$$

$$1 \times n \qquad n \times n \qquad n \times 1$$

$$\frac{\partial}{\partial w} = 0$$

What Next:? (next three)

- Application of SVD and Eigen Decomposition
- More Insights into Supervised Learning
- Bayesian View and Optimal Classification
- Practical Issues in Optimization