SMAI-M20-L09: Aspects of Supervised Learning

C. V. Jawahar

IIIT Hyderabad

August 28, 2020

Class Review (L09)

1. Consider a matrix A of size $m \times n$. Rank of A is related m or n?
2. A and B are two independent events such that $P(\overline{A}) = 0.4$ and $P(A \cap B) = 0.2$ Then Find $P(A \cap \overline{B})$.
3. If $\mathbf{A}$ is a $n \times n$ matrix, with every pair of columns orthogonal i.e., $\mathbf{a_i} \cdot \mathbf{a_j} = \mathbf{0}$ $\forall i, j$ and $||\mathbf{a_i}|| = 1$.
4. Product of Eigen values of a real square matrix is known as ?
5. $X \sim N(0, 1)$, $Y \sim N(1, 1)$ and $Z = X + Y$. Then,

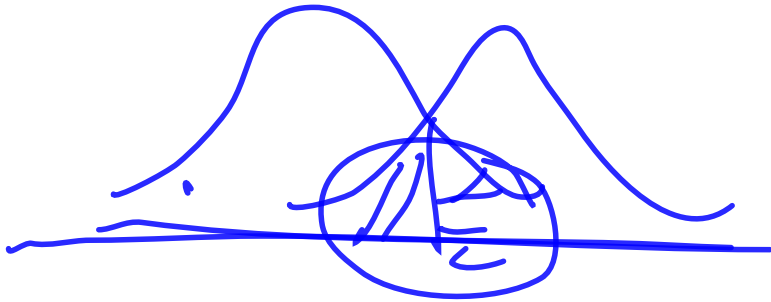
Blank

Recap:

- Problem Space:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
    - for classification
    - for regression
  - Learn useful features
- 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
  - SVD, Eigen Decomposition
  - MLE

Blank

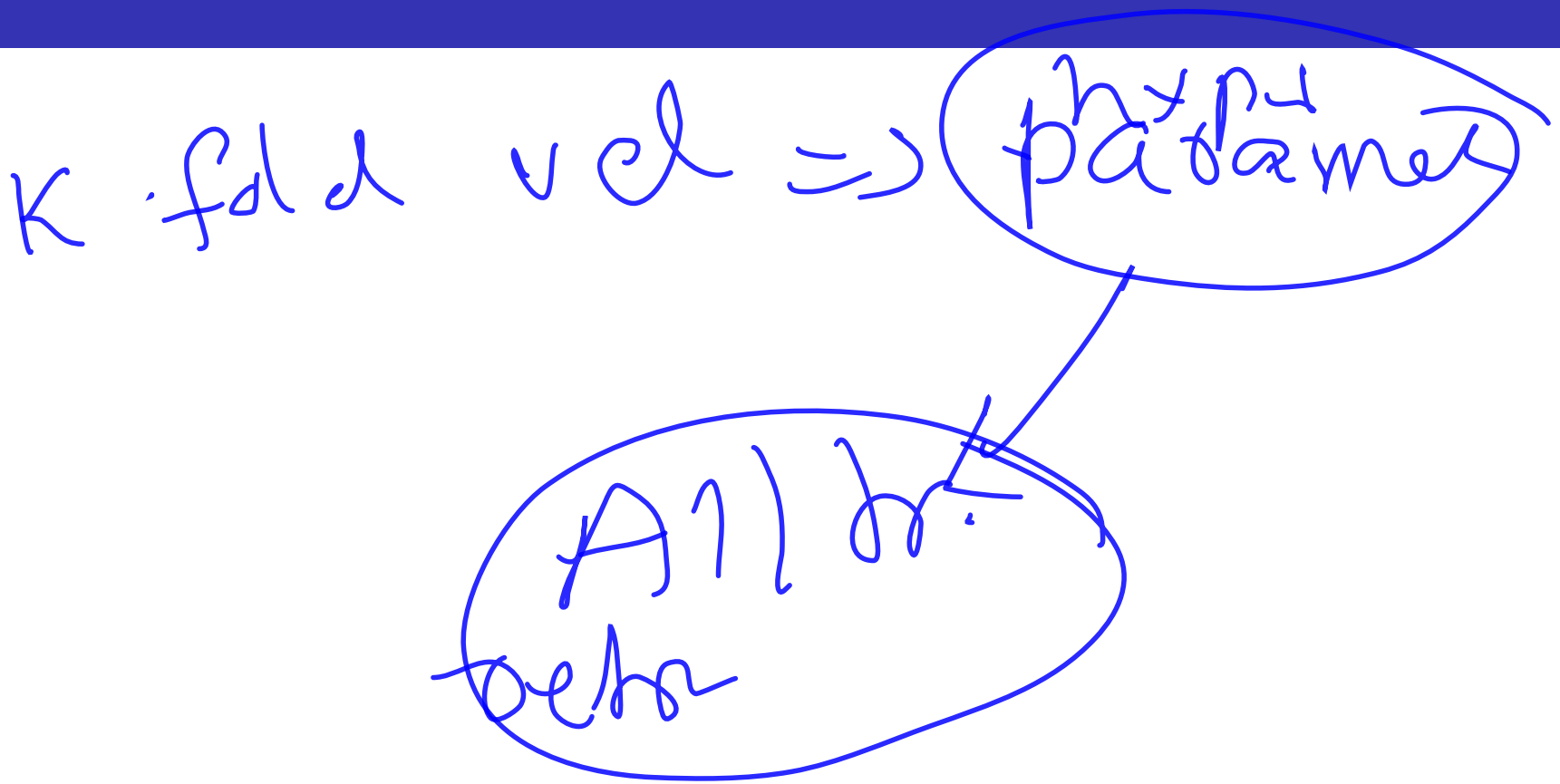
This Lecture Session:



Micro-Lecture Videos

1. Minimum Error Classification
   - The best we can ever achieve.
   - Q: Even Deep Learning can not do better. Sad. Isn't? :-)
2. Model complexity and Occam's razor
   - Simple, yet good model
   - New Key words: Regularization, Model Complexity
3. Validation Error, K-Fold and LOO
   - An estimate of the test error.
   - Q: How do we prefer one of the two solutions (say NN with K=3 and K=5)? Finding the right hyper parameters.

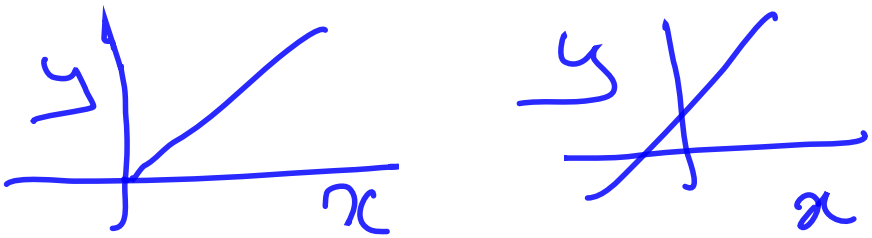
Blank



Questions? Comments?

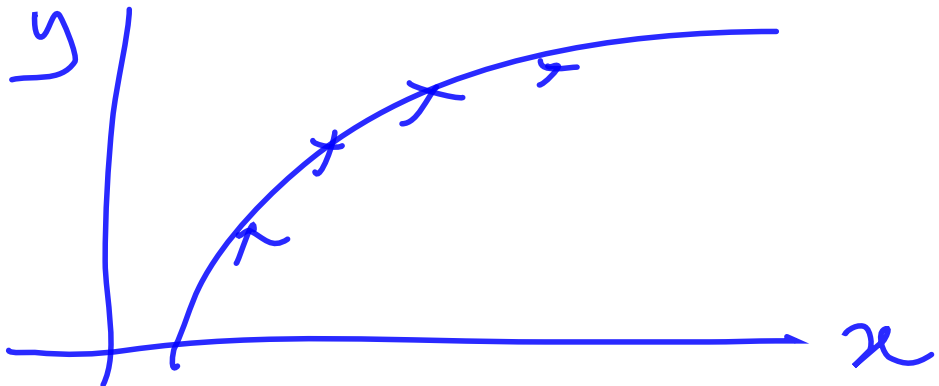
Blank

Discussions Point - I



In the context of regression (assume $\mathbf{x} \in R^1$ i.e., only one feature):

1. We know how to fit a line passing through origin with a model $y = \mathbf{w}^T\mathbf{x}$
2. We know how to fit a line, even if it is not passing through origin, with a model $y = \mathbf{w}^T\mathbf{x}'$. where $\mathbf{x}'$ is defined as $\begin{bmatrix} \mathbf{x} \\ 1 \end{bmatrix}$
3. How do we model the problem of fitting a quadratic (say a parabola) given a set of points?. What is $\mathbf{x}$? Is there a closed form expression?



Blank

Discussion Point - II

1. Can we guess/compute/complete the missing elements of the matrix:

$$\begin{bmatrix} 7 & ? & ? \\ ? & 8 & ? \\ ? & 12 & 6 \\ ? & ? & 2 \\ 21 & 6 & ? \end{bmatrix}$$

   if we know that this is a rank-1 matrix (or every row is a multiple of each other)[1]

2. If $\mathbf{A}$ is a $m \times n$ matrix and $\mathbf{A}_k$ is the nearest rank-k matrix, $\mathbf{A}_k$ can be computed using SVD as (i.e., $\mathbf{A}_k = \arg\min_{\mathbf{B}} ||\mathbf{A} - \mathbf{B}||_F^2$ and $rank(\mathbf{B}) = k$)

$$\mathbf{A}_k = \mathbf{U}_k \mathbf{D}_k \mathbf{V}_k^T$$

   Details:[2]

[1] Read later: https://web.stanford.edu/class/cs168/l/l9.pdf

[2] Read Later: https://courses.cs.washington.edu/courses/cse521/16sp/521-lecture-9.pdf

Blank

$m \times n$ $m \times n$

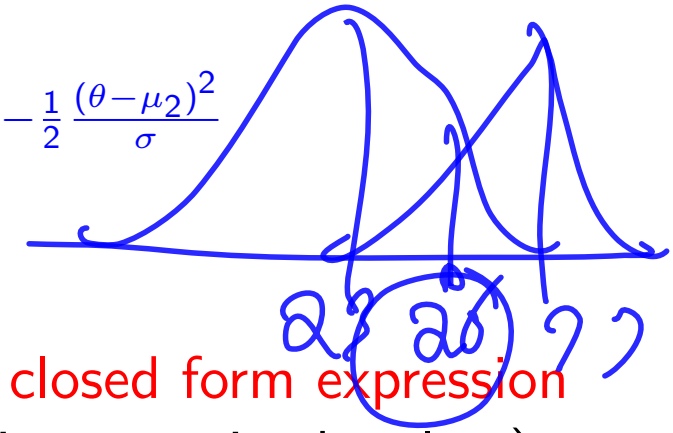
Discussion Point - III

Consider the binary classification problem where both classes are univariate Gaussian (assume $\mu_1 \leq \mu_2$) . i.e., $P(\omega_i|x) = \mathcal{N}(\mu_i, \sigma_i^2)$. Optimal decision is "**Decide as** ω_1 **if** $x \leq \theta$ **else** ω_2".

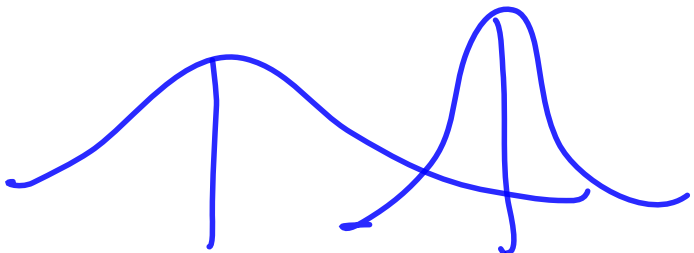
1. When $\sigma_1 = \sigma_2 = \sigma$, show that the optimal threshold (i.e., θ) is the mid point of means. Ans:

$$\frac{1}{2\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\theta-\mu_1)^2}{\sigma}} = \frac{1}{2\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\frac{(\theta-\mu_2)^2}{\sigma}}$$

or $\theta = \frac{\mu_1+\mu_2}{2}$



2. If $\sigma_1 \neq \sigma_2$, what will be the θ? Can we get a closed form expression for θ ? (for convenience, discard the normalizing term in the class)



Blank

What Next:? (next three)

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