SMAI-M20-L13: PCA(Cont.)

C. V. Jawahar

IIIT Hyderabad

September 7, 2020

Class Review

Consider the vector $\mathbf{w} = [w_1, w_2]^T$ and the objective function to be minimized as:

$$\min_{\mathbf{w}} (3w_1 + 4w_2 - 12)^2 + \lambda g(\mathbf{w})$$

If $g(\mathbf{w})$ is Lp norm of $\mathbf{w}$, what is the optimal value of $\mathbf{w}$?

- for various p?
- for various λ?

Blank

Recap:

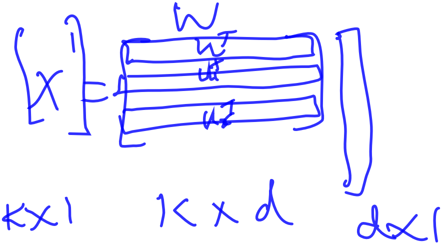
- Problem Space:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
  - Dimesnionality Reduction and Representation (Feature Selection, PCA, Neural Embeddings)
  - Matrix Factorization for Data Matrices: (LSI, Matrix Completion, Recommendation Systems)
- Supervised Learning:
  - Notions of Training, Validation and Testing; Loss Function and Optimization
  - Generalization, Overfitting, Occam's razor, Model Complexity, Bias and Variance, Regularization.
  - Performance Metrics, Estimating error using validation set.
- 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

This Lecture:

1. PCA as Compression
   - Dimensionality reduction that allow minimal loss in the data.
2. Intuitive Intro to Gradient Descent
   - An iterative optimization algorithm for minimizing loss functions.
3. Feature Normalization
   - Dimensions to be brought to some common scale and variation/range; Numerical and practical advantages.

Questions? Comments?

Blank

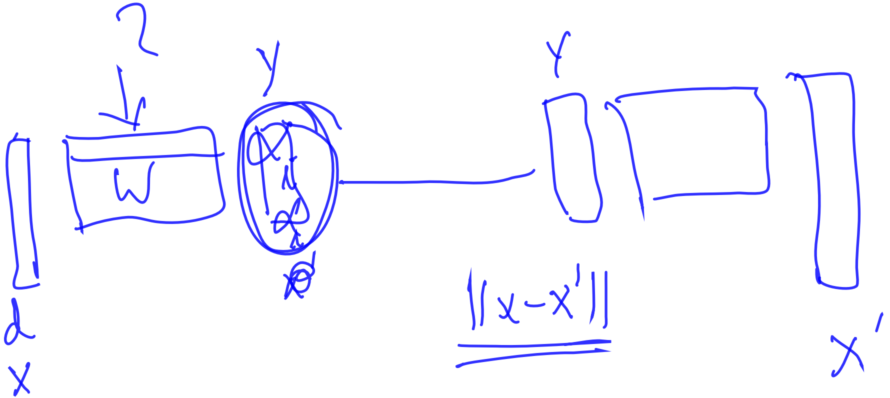


x $d \times 1$

Σ $d \times d$

u $d \times 1$

$a^T b = b^T a$



Blank

Discussions Point - I

- It is common to represent signal/data using some orthogonal basis. Discrete Fourier Transform (DFT) is an example that use complex exponential basis.
- Discrete Cosine Transform(DCT) (a close friend of DFT) is popular in the compression standards such as JPEG. We retain only a small number of coeff; rest as zero; thus obtain compression.
- The optimal linear transform to do this is popularly known as KL Transform (where eigen vectors of the covariance form the basis). Our PCA is a close relative.

Why is PCA/KL-Transform based compression algorithms not yet popular in the standards we all use?

DCT vs KLT

Blank

Blank

Comment about the effectiveness of GD for optimizing the loss functions. Initialization is also shown.

Blank

Blank

Discussions Point - III

Consider a two dimensional representation with each dimensions as x_1 and x_2. i.e., $\mathbf{x} = [x_1, x_2]^T$.
We know that we need to do a feature normalization to obtain $\mathbf{x}' = [x_1', x_2']^T$

A $x_i' = \frac{x_i}{\alpha_i}$

B $x_i' = x_i - \alpha_i$ (after the class?)

How do we represent

1. $\mathbf{x}' = \mathbf{L}\mathbf{x}$? What is $\mathbf{L}$
2. What should be the matrix $\mathbf{A}$ such that the Euclidean distance in $\mathbf{x}'$ is same as Mahalanobis distance in]$\mathbf{x}$. i.e.,

$$[\mathbf{x} - \mathbf{y}]^T \mathbf{A}^{-1} [\mathbf{x} - \mathbf{y}] = [\mathbf{x}' - \mathbf{y}']^T [\mathbf{x}' - \mathbf{y}']$$

Blank

Blank

What Next:? (next three)

1. Application of PCA
2. Perceptron Algorithms
3. Analysis of Gradient Descent Algorithm
4. Implementation of Gradient Descent
5. Analysis of Perceptron Algorithm