SMAI-M20-L14:Gradient Descent

C. V. Jawahar

IIIT Hyderabad

September 9, 2020

Class Review

We are given a set of 2D points X on/around a line. We compute the covariance matrix from this data; and its eigen values and eigen vectors. Then:

- What can we say about the mean μ?
- What can we say about the covariance matrix Σ?
- What can we say about the eigen values λ?
- What can we say about the eigen vectos $\mathbf{u}$?

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, Linear Classification; Linear Regression
  - Optimal Decision as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2
  - PCA
  - Gradient Descent Optimization

This Lecture:

1. Eigen Faces
   - A powerful application of PCA
   - Face representation and compression.
2. Appreciating Gradient Descent
   - Does gradient descent improve the solution in every step?
   - What is the optimal learning rate?
   - Is there a better update rule?
3. Perceptron Algorithm
   - An algorithm for linear classification
   - Assumes linear seperability.

Questions? Comments?

Appreciating PCA

Maximum Variance Direction: 1st PC a vector v such that projection on to this vector capture maximum variance in the data (out of all possible one dimensional projections)

$$\frac{1}{n}\sum_{i=1}^{n}(\mathbf{v}^T\mathbf{x}_i)^2 = \mathbf{v}^T\mathbf{X}\mathbf{X}^T\mathbf{v}$$

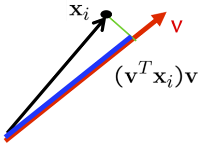
Minimum Reconstruction Error: 1st PC a vector v such that projection on to this vector yields minimum MSE reconstruction

$$\frac{1}{n}\sum_{i=1}^{n}\|\mathbf{x}_i - (\mathbf{v}^T\mathbf{x}_i)\mathbf{v}\|^2$$

blue² + green² = black²

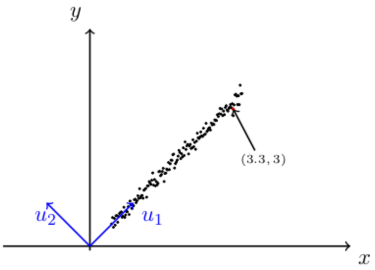
black² is fixed (it's just the data)

So, maximizing blue² is equivalent to minimizing green²



Slide from Nina Balcan

Reconstruction: Numerical Example



- $u_1 = [1, 1]$ and $u_2 = [-1, 1]$ are the new basis vectors
- Let us convert them to unit vectors $u_1 = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$ & $u_2 = \begin{bmatrix} \frac{-1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix}$

- Consider the point $x = [3.3, 3]$ in the original data
- $\alpha_1 = x^T u_1 = 6.3/\sqrt{2}$
  $\alpha_2 = x^T u_2 = -0.3/\sqrt{2}$
- the perfect reconstruction of x is given by (using $n = 2$ dimensions)

$$x = \alpha_1 u_1 + \alpha_2 u_2 = \begin{bmatrix} 3.3 & 3 \end{bmatrix}$$

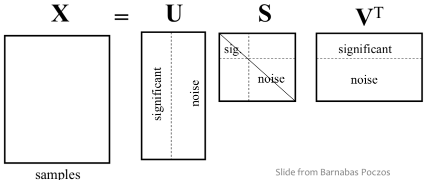
- But we are going to reconstruct it using fewer (only $k = 1 < n$ dimensions, ignoring the low variance u_2 dimension)

$$\hat{x} = \alpha_1 u_1 = \begin{bmatrix} 3.15 & 3.15 \end{bmatrix}$$

(reconstruction with minimum error)

SVD and PCA

$\mathbf{X}_{\text{features} \times \text{samples}} = \mathbf{USV}^T$



Slide from Barnabas Poczos

- **Columns of U**
  - the principal vectors, $\{ \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)} \}$
  - orthogonal and has unit norm – so $U^T U = I$
  - Can reconstruct the data using linear combinations of $\{ \mathbf{u}^{(1)}, \ldots, \mathbf{u}^{(k)} \}$

- **Matrix S**
  - Diagonal
  - Shows importance of each eigenvector

- **Columns of $\mathbf{V}^T$**
  - The coefficients for reconstructing the samples

Faces and Eigen Vectors

Representation and Reconstruction of Face from 16 EVs

Blank

Blank

How many coeff. are required to represent a block?

144 to 60, 16, 6, 3

Figure: Original, Blocks of size 12 X 12 in 60, 16, 6 and 3

Blank

Discussions Point - I

Consider images of size 100×100 and we have 200 such images. Assume means are subtracted.

1. What is the size of the covariance matrix?
2. What is the rank of the covariance matrix?
3. What is the size of XX^T and X^TX and what are their ranks?
4. How are the Eigen values of X^TX and XX^T related?
5. How are the Eigen vectors of X^TX and XX^T related?
6. How does the above help computationally in Eigen faces?

Ans/Hint:

- Let $\mathbf{A} = \mathbf{X^T X}$ and $\Sigma = \mathbf{X X^T}$
- If $\mathbf{v}$ is the EV of $\mathbf{A}$, then $\mathbf{Xv}$ is the EV of Σ.

Blank

Blank

Discussion Point - II

We know there are better update rules than gradient descent?

1. Write the newton's update rule?
2. Why is still Newton's method not preferred? [1]

[1] https://stats.stackexchange.com/questions/253632/why-is-newtons-method-not-widely-used-in-machine-learning

Blank

Blank

What Next:? (next two)

1. More about Gradient Descent
2. Neuron Model and Perceptrons
3. Analysis of Perceptron Algorithm