SMAI-M20-01: CSE-471: Statistical Methods in AI

C. V. Jawahar

IIIT Hyderabad

August 10, 2020

About the course: Scope

- A **fundamental** course on machine learning.
- **Not an** advanced course on ML/AI
- Emphasis on:
  - Basics Concepts, Principles, Basic Maths
  - Connect to Practical and Effective ML Algorithms
- Note: Class could be highly heterogeneous:
  - UG, PG, Working Professionals, New Students, People who have taken an ML course else where, students who use these ideas in their own research/work etc.

About the course: Getting Ready

- Exposure:
  - Maths in the form of UG Courses and Schools.
  - Programming/Algorithms: Comfortable with adaptation/hand-on.
- Maths:
  - Typical Engg Maths. Not super advanced.
  - Topics: Linear Algebra, Probability, Differential Calculus
- Programming:
  - Most tools/libraries are in python. No plan to teach python or programming. Familiarity with plotting (graphs) etc. will be useful.
  - We will use: **jupyter notebook**. If you have not used, try it out before the next class.
  - No super-heavy compute expt. planned. Some TA support will be provided.
- Infrastructure:
  - Some compute (say a laptop) advisable. Though VPN/Remote/Cloud may also work in many cases.
  - Some internet connection (say 4G) expected. Course accounts buffers to take care of unfortunate network failures.

About the course: Course Coverage

1. **Part 1: Basics**
   - Mathematical Foundations; Role of Linear Algebra and Probability; Supervised Leaning Formulation and Challenges, Sample Algorithms
2. **Part II: Fundamental Algorithms**
   - Linear Methods in Machine Learning; Regression, PCA, Logistic Regression, Perceptrons, Gradient Descent, Multiclass
3. **Part III: Powerful Ideas**
   - SVMs, Kernels, Nonlinear Methods, Ensembe, Semi-Supervised, Unsupervised and Self-Supervised Learning
4. **Part IV: Neural Network Learning**
   - Artificial Neural Networks, MLP and Back Propagation; Intro to Deep Learning, Intro to CNN, RNN

Approximately 25% emphasis/time on each part.

Course Structure and Evaluation

- Course will be fully <u>online</u>
- What does it may imply?
  - You need to pay more attention to all content/input. You can't just pick up from the lectures.
  - You need to be more regular, and more disciplined.
  - Communicate and discuss technical stuff in online mode.
- How is this course (SMAI-M2020)?
  - Will try to align to the traditional models: Regular meetings, some lectures/explanations, some in person QA, etc.
  - Some specific material and background: (i) Lecture Notes (ii) Some Videos (iii) Some Links
  - Learn by Regular Work: Home Works, Assignments, In Class Reviews, Quizzes etc.

Course Structure and Evaluation

- Lectures, Tutorials (as per time table)
  - Lec: M,W,F: 9.30-10.30AM IST
  - Tut: ?? (TBD)
- Office Hours (OH) (Additional Support; Weekly (?)):
  - Logistics, Exceptions, Semi-Technical, Administrative
  - TA (One Common; One on Reserving)
- Emails:
  - Use Office Hours, if your request can wait. Get personal attention.
  - Any course specific emails: please send to: **smai.m2020gmail.com**
- Expected course load:
  - Regular activities (Homeworks, Assignments)
  - Some preparation (20-30 mins before the lecture)
  - Reading, Thinking, Discussing
  - Meetings: Lec (*), Tut, OH (Optional)
- **If attending, attend seriously and disciplined.** This is a large class. May be not everyone was not given opportunity.

Course Structure and Evaluation

- Regular Homeworks: 40%
  - Handwritten and Programming
  - Approximately 200 Points. Your best 80% (Approximately out of 160) will be used for grading
  - Stretch beyond the class/lecture. Learning+Evaluation is the goal.
- Quiz (3): 25%
  - Closest to the traditional exams.
  - Evaluation is the primary purpose
- Assignment (3): 25%
  - Programming and Exploration
  - Exposure to the depth is the focus.
- In Class Review/Recap: 10%
  - Objective questions and regular review. Light weight. Learning is the primary objective.

Problem Space

Let us start with an Introductory Video

- How Can Data be Useful in Solving Problems?
- `https://www.youtube.com/watch?v=8xniRSjRyCQ&feature=youtu.be`
- Around 10 mins; Let us wait for 15 mins.

Problem Space

$$y = f(\mathbf{W}, \mathbf{x})$$

- $\mathbf{W}$ is the learnable parameters.
- ML algorithms aims to learn/find $\mathbf{W}$ from the Data $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ $i = 1, \ldots, N$.
- Usually only
  - a portion of the data is used for "Training" (developing the solution or computing $\mathbf{W}$) and
  - the rest of the data is used for "Testing" (or measuring the performance.

What Next:?

- We will use this week (until 15 Aug) for streamlining the Course:
  - New students
  - Teething problems in online mode.
- Topics:
  - K Nearest neighbour Algorithm
  - Performance Metrics of ML solutions.
- Recap of (do it yourself!):
  - Vectors and Matrices
  - Probability Distributions and Bayes Theorem
- Logistics:
  - Formal Details on Moodle
  - TAs on Board
  - Office Hours to Start
  - Mechanism for submitting regular homeworks (shiksha.iiit.ac.in)
- A good text book:
  - Mathematics of Machine Learning (pdf available) https://mml-book.github.io/
  - More on Moodle

SMAI-M20-02: Representation and Classification

C. V. Jawahar

IIIT Hyderabad

August 12, 2020

Agenda

1. Recap/Repeat of L01
   - https://www.dropbox.com/s/ltmyx9y15hxnmm8/L1.pdf?dl=0
2. This lecture:
   - Representation as a vector
   - Nearest Neighbour Algorithm
   - Linear Classification
3. Discussions and Extensions
4. Reviews and Next Step

Recap

- Exposure:
  - Maths in the form of UG Courses and Schools.
  - Programming/Algorithms: Comfortable with adaptation/hand-on.
- Maths:
  - Typical Engg Maths. Not super advanced.
  - Topics: Linear Algebra, Probability, Differential Calculus
- Tools:
  - Most tools/libraries are in python. No plan to teach python or programming.
  - `https://jupyter.org/`
- Infrastructure:
  - Some compute (say a laptop) advisable. Though VPN/Remote/Cloud may also work in many cases.
  - Some internet connection (say 4G) expected. Course accounts buffers to take care of unfortunate network failures.

About the course: Course Coverage

1. **Part 1: Basics**
   - Mathematical Foundations; Role of Linear Algebra and Probability; Supervised Leaning Formulation and Challenges, Sample Algorithms
2. **Part II: Fundamental Algorithms**
   - Linear Methods in Machine Learning; Regression, PCA, Logistic Regression, Perceptrons, Gradient Descent, Multiclass
3. **Part III: Powerful Ideas**
   - SVMs, Kernels, Nonlinear Methods, Ensembe, Semi-Supervised, Unsupervised and Self-Supervised Learning
4. **Part IV: Neural Network Learning**
   - Artificial Neural Networks, MLP and Back Propagation; Intro to Deep Learning, Intro to CNN, RNN

Approximately 25% emphasis/time on each part.

Course Structure and Evaluation

- Lectures, Tutorials (as per time table)
  - Lec: M,W,F: 9.30-10.30AM IST
  - Tut: Wed. 5.30-7.30 (may be 1 Hr batches?)(Watch for Announcements)
- Office Hours (OH) (Additional Support; Weekly (?)):
  - Logistics, Exceptions, Semi-Technical, Administrative
  - TA (One Common; One on Reserving)
- Emails:
  - Use Office Hours, if your request can wait. Get personal attention.
  - Any course specific emails: please send to: **smai.m2020@gmail.com**
- Expected course load:
  - Regular activities (Homeworks, Assignments)
  - Some preparation (20-30 mins before the lecture)
  - Reading, Thinking, Discussing
  - Meetings: Lec (*), Tut, OH (Optional)
- **If attending, attend seriously and disciplined.** This is a large class. May be not everyone was not given opportunity.

Course Structure and Evaluation

- Regular Homeworks: 40%
  - Handwritten and Programming
  - Approximately 200 Points. Your best 80% (Approximately out of 160) will be used for grading
  - Stretch beyond the class/lecture. Learning+Evaluation is the goal.
- Quiz (3): 25%
  - Closest to the traditional exams.
  - Evaluation is the primary purpose
- Assignment (3): 25%
  - Programming and Exploration
  - Exposure to the depth is the focus.
- In Class Review/Recap: 10%
  - Objective questions and regular review. Light weight. Learning is the primary objective.
- 5-10% changes due to operational issues of uncertainties.

Recap: Technical Summary

- We had seen how data can lead to learnable parameterized functions $f(\mathbf{W}, \mathbf{x})$
- The notion of "Training" and "Testing"
- We started with an overview at: `https://www.youtube.com/watch?v=8xniRSjRyCQ&feature=youtu.be` on how data can help in solving problems.

This Lecture

- How to represent in the form of $\mathbf{x} \in R^d$
- Nearest neighbour Algorithm for Classification
  - Classify based on the majority labels in the neighbourhood.
- A simple Linear Classifier
  - $sign(\mathbf{w}^T\mathbf{x})$
  - Either +ve or -ve.
- If you have not see the pre-lecture videos, please do. At least the one on "Linear Classifier"
  - Go to MS Teams
  - Go to the channel"Lectures - Information". See the posts
  - Watch the last video/link.
  - `https://www.youtube.com/watch?v=P92mkhzt6Hg&feature=youtu.be`
- (We will take 10 mins for people who might have missed; But as we move forward, do plan 30 mins of pre-lecture preparation previous day.)

Blank

Discussions

Q: We know, if we have two classes, we can classify as:

$$sign(\mathbf{w}^T\mathbf{x})$$

Assume we have three classes, how do we extend this decision making logic?

Blank

Discussions

Q: $\mathbf{w}^T\mathbf{x}$ is a line passing through origin. How do we characterize "general" line?

- Ans1: $\mathbf{w}^T\mathbf{x} + b$
- Ans2: $\mathbf{w}'^T\mathbf{x}'$ where $\mathbf{x}'$ is an augmented vector. i.e., $\mathbf{x}' = [\mathbf{x}^T, 1]^T$. (There is notational convenience for this.)
- (Refer the usage of "bias" in neural networks. You will appreciate this better)
- In general, when we write as $\mathbf{w}^T\mathbf{x}$, often $\mathbf{x}$ is augmented and bias is absorbed.

Blank

Feature Transformation

- In general, this feature transformation

$$\phi : \mathbf{x} \rightarrow \mathbf{x}'$$

  is a useful trick.
  - **Goal:** Get better features starting from "raw" measurements/features.
- The feature transformations like:

$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

  can lead to dimensionality reduction, when the matrix $\mathbf{W}$ has more columns than rows.

Blank

Review Question

What is the angle between the two lines characterized by

$$\mathbf{w}_1 = [1, 1]^T$$

$$\mathbf{w}_2 = [1, -1]^T$$

(a) 0 (b) 45 (c) 90 (d) 120

Review Question

We stop two people at random. What is the probability that they were born on the same day of the week?

(a) $\frac{1}{7}$ (b) $\frac{1}{7^2}$ (c) $\frac{1}{7+7}$ (d) $\frac{1}{2}$ (e) None of the above

What Next:?

- We will use this week (until 15 Aug) for streamlining the Course:
- Topics: Performance Metrics of ML solutions.
- Logistics:
  - Office Hours (Watch MS Team Announcements and Channels)
  - Start using Skikha (look for instructions in the Channels)
  - Formal Details on Moodle; TAs active (by Weekend or Early next week)
- Watch Channels in "Teams" regularly.

SMAI-M20-03: Performamnce Metrics

C. V. Jawahar

IIIT Hyderabad

August 14, 2020

Recap

1. Recap/Repeat of L01
   - Scope, Course Plans.
   - L01: https://www.dropbox.com/s/ltmyx9y15hxnmm8/L1.pdf?dl=0
   - L02: https://www.dropbox.com/s/7216t0zl39xgmeq/L2.pdf?dl=0
2. Scope:
   - Representation as a vector in R^d
   - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
     - Notion of Training and Testing
   - Feature Transformation as a useful trick:

$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

     - Dimensionality Reduction
   - Two Simple Classification Schemes:
     - Nearest Neighbour Algorithm
     - Linear Classification
     - — $sign(\mathbf{w}^T\mathbf{x})$; Either +ve or -ve.
     - — Many ways to extend to more than 2 classes

This Lecture: Performance Metrics

- In the classification setting:
  - Accuracy, Confusion Matrix, Precision, Recall
- In the ranking/retrieval setting:
  - Precision, Recall, AP, F-score.
- Many more performance metrics see for example:
  - `https://en.wikipedia.org/wiki/Confusion_matrix`
- Pre-Class Videos in the Channel (12 min) (Please make sure you see it before the class)
- (*Let us spend 5 mins today to watch for those who have missed. But make sure from the next class you spend 20-30 mins before the lecture.*)

Blank

Discussions Point - I

Q: We have heard in recent months:

- The PCR-Test (or RT-PCR) for COVID (TEST-1) has some "false negatives". (PCR-Test tests whether a patient has COVID right now by looking at the swab.)
- The Antibody test used or COVID (TEST-II) has very high "false positives". (Antibody test is used to know the immunity level. It checks whether a person has already got COVID, (usually after recovered or infected unnoticed).)

What does these sentences mean? Can you please explain the implication of these two in two sentences S1 and S2 (say to your grand mother!). Support (like) the best explanations!!.
(Type S1 and S2 in chat section)

Blank

Discussions Point -II

Q: Let us consider that FN rate of TEST-I is 10% and the FP of rate of TEST-II is 50%.
An SMAI student gave a recommendation to the Govt that Every Person should be tested three times (say in a day) and majority label should be assigned.

- Is this student, technically sound?
- Why is that Govt is not ready to listen to his/her recommendation?

Blank

Discussion Point - III

Q: There was a proposal to an airport from a supplier:
"We have a multi-sensor product that can recognize terrorists reliably. We can classify a normal person vs terrorist (two class classification problem) with an accuracy of 99.5% "

- Is this a good product for detecting terrorists?
- As an SMAI student, who evaluate this product, what performance metrics you will prefer?

Blank

Discussions Points - IV

$$\begin{bmatrix} 10^{-4} & 0 \\ 0 & 10^{3} \end{bmatrix}$$

In a problem of classification of "cricket ball vs Tennis ball" there were two features (i) size/radius measured in micometer (ii) Weight measured in Tons.

- Fact: SMAI student felt that size should be in cm (1 cm $= 10^4$ mico meter) and weight in Kg (1 Ton $= 10^3$ Kg) for good performance of the algorithm she uses.
- Suggest the 2×2 matrix $\mathbf{W}$ to obtain the new feature representation from the old one.

$$\begin{bmatrix} 10^{-4} & 0 \\ 0 & 10^{-9} \end{bmatrix} \qquad \mathbf{x}' = \mathbf{W}\mathbf{x} \qquad \begin{bmatrix} 10^{4} & 0 \\ 0 & 10^{-3} \end{bmatrix}$$

Blank

Blank

Review Question (one, none or more correct)

We are given a set of 2D points from two classes, as shown in the Figure.
Q: To make the computations efficient, we want to do a dimensionality reduction from 2D to 1 D with the help of a 1×2 matrix $\mathbf{W}$

$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

What should be the $\mathbf{W}$ matrix be in this case?
(Indeed the goal is to get good classification performance in the new feature space $\mathbf{x}'$, while the computations could be efficient)

(a) $[1, 0]$ (b) $[-1, 0]$ (c) $[2, 0]$ (d) $[1, 1]$ (e) $[0, 1]$ (f) $[0, -1]$ (g) $[1, 0]^T$

Blank

Blank

Review Question (one, none or more correct)

We are given a set of 2D points from two classes, as shown in the Figure. We want to "rorate" the data so that points are spread across first (x) axis (i.e., something <u>like</u> rotate clockwise by 45°)

$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

What should be the 2×2 matrx $\mathbf{W}$ be in this case?

(a) $\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ (b) $\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$ (c) $\begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}$ (d) $\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ (e) $\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$
(f) $\begin{bmatrix} 1 & 1 \\ -1 & 1 \end{bmatrix}$ (g) Any one of the above (h) None of the above

Blank

Review Question (one, none or more correct)

What is the probability that, in a room of 30 people, there is a pair of people who have the same birthday. (use calculator, pick the closest)
(a) 0.11 (b) 0.31 (c) 0.51 (d) 0.71 (e) 0.91

What Next:?

- Logistics: Watch MS-Teams for use of “shiksha”.
  - Announcements Channel for General Announcements
  - Lecture related material on "Lecture Announcements"
  - Emails at: smai.m2020@gmail.com
- Topics and Preparation:
  - Revise: Rank of a matrix. Interpretation of Rank.
  - Continue on Features and Supervised Learning (Next Week)
  - Budget time to watch videos before class.
- Next Week:
  - Start submission of regular HWs (a week to submit) and Class Reviews (in class 5-10 mins)

SMAI-M20-04:Appreciating Data in High Dimension

C. V. Jawahar

IIIT Hyderabad

August 17, 2020

Recap: Administrative

1. Scope and Course Plans:
   - Scope, Course Plans.
   - L01: https://www.dropbox.com/s/ltmyx9y15hxnmm8/L1.pdf?dl=0
   - L02: https://www.dropbox.com/s/7216t0zl39xgmeq/L2.pdf?dl=0
   - L03: https://www.dropbox.com/s/sw1mosvwsiv85ny/l3.pdf?dl=0
2. Logistics:
   - Started to use "shiksha" for questions.
   - Regular HW:
     - Approximately 3 questions will be posted on every leccture day. (Roughly 1 Q each of 1 pt, 2pt and 3pt) 1st set expected to be on Shikha today.
     - Submit within a week.
     - You need to do only 80%. (i.e., 160 points out of 200 expected.)
     - Use "TS&GH" channel for any difficulty. Don't wait for the last day.
   - Class Review Questions:
     - Will move to Shiksha systematically. Today in lecture slides.
     - First questions + answers on google forms (this week).
     - Later fully on Shiksha. (Need to understand the load/any issues).

Summary: Till Now

1. Representation as a vector in R^d
2. Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
   - Notion of Training and Testing
3. Feature Transformation as a useful trick:
4. $\mathbf{x}' = \mathbf{W}\mathbf{x}$
   - Dimensionality Reduction
5. Two Simple Classification Schemes:
   - Nearest Neighbour Algorithm
   - Linear Classification
   - — $sign(\mathbf{w}^T\mathbf{x})$; Either +ve or -ve.
   - — Many ways to extend to more than 2 classes
6. Performance Metrics:
   - Classification: Accuracy, TP/FP etc., Confusion Matrix
   - Ranking: Precision, Recall, F-Score, AP

This Lecture: Appreciating Data

- Data could come from a physical process
  - Data is not some random numbers.
  - Structure of the data allows us to learn
  - Reasonable assumption of Multivariate Gaussian
- Geometry of Representation:
  - Lines, Planes, Hyperplanes.
  - Geometry in high dimension.
  - What does it mean by

    $$\mathbf{w} \leftarrow \mathbf{w} + \nabla \mathbf{w}$$

    What is the geometric interpretation?
- Practical Challenges in High Dimension
  - Too many parameters to learn. Lot more samples required in High Dimension.
  - Computational and Practical Advantages.
  - Need of dimensionality reduction.

Blank

Discussions Point - I

Q: Consider a linear transformation $d \to d$ (i.e., $\mathbf{W}$ is a square matrix)

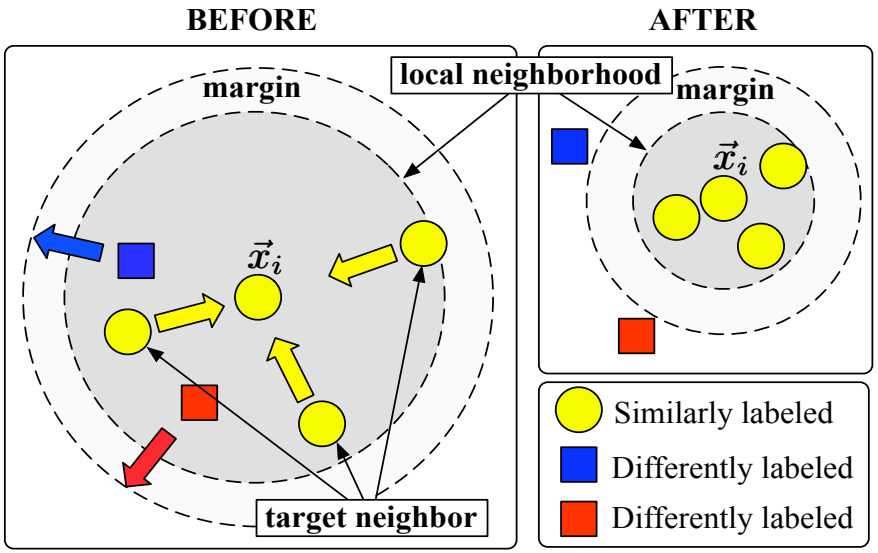
$$\mathbf{x}' = \mathbf{W}\mathbf{x}$$

We use a K-NN algorithm (the same K and distance as Euclidean distance) in original and new space.

- Will the performance (say accuracy) of the algorithm be same in both the space for any $\mathbf{W}$? i.e., with $\mathbf{x}$ and $\mathbf{x}'$? (Discuss)
- If no, what should be the condition on $\mathbf{W}$ to guarantee that?

Discussions Point -Ib (Advanced)

If we can lean **W** for a K-NN so that performance improves, what should be our desirability?



Ans (Fig and Ans from the LMNN paper[1]): For each sample, we desire

- Its K (=3) target neighbors lie within a smaller radius after transforming with **W**
- Differently labeled inputs lie outside this smaller radius, with a margin of at least one unit distance.

How do we explain these requirements?

[1] Read initial sections of: "Distance Metric Learning for Large Margin Nearest Neighbor Classification", NIPS 2006.

Blank

Blank

Discussions Point -II

An SMAI student (Raju) implements $K - NN$ and tested it on a popular data set. He conducted an experiment to vary K (say from 3 to 15) and plot the performance.

Q1:

- Will he see a systematic increase in accuracy with K?
- Will he see a systematic decrease in accuracy with K?
- Will he see a systematic increase followed by a systematic decrease?

Q2: Can you help Raju in finding the best K?

Blank

Discussion Point - III

Over years, we have figured out HYD temperature in Jan and May are $\mathcal{N}(23, \sigma^2)$ and $\mathcal{N}(33, \sigma^2)$ (i.e, Normal, mean 23 and 33; Variance the same).

Q: We have 100 days of data from Jan and 100 days from May, but not labelled. We want a classifier as:

"If temp $< \theta$, then Jan else May"

What should be the value of θ intuitively?

- 28 $(= \frac{23+33}{2})$
- Less than 28.
- More than 28.

Why?

Blank

Blank

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

Confusion matrix is:

(a) Square (b) Always Diagonal (c) Can never be diagonal (d) Can be diagonal (e) Always Symmetric (f) Can never be symmetric (g) Can be Symmetric

Blank

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

$$\frac{TP}{P}$$

is known as:

(a) Accuracy (b) Precision (c) Recall (d) None of the above

Blank

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

A disease occurs with a probability of 0.4 (i.e., it is present in 40% of the population). You have a test that detects the disease with a probability 0.6, and produces a false positive with probability of 0.1. What is the (posterior) probability that the test comes back positive.

Hint: S is the event that you are sick; P is the event that test comespositive.

$$P(S|P) = \frac{P(P|S)P(S)}{P(P)} = \frac{P(P|S)P(S)}{P(P|S)P(S) + P(P|\bar{S})P(\bar{S})}$$

(a) 0.6 (b) 0.7 (c) 0.8 (d) 0.9 (e) 0.95

Blank

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

Two SMAI students (Raju and Sheela) worked on the same problem with the same measurements/features and samples, except that their feature orderings were different. (i.e., $\mathbf{x}$ and $\mathbf{x}'$ were permutations.) Identify correct statement(s).

- (a) Both got the same accuracy with KNN (same K and Eucli. distance)
- (b) Both got different accuracy with KNN (same K and Eucli. distance)
- (c) Their confusion matrices were different i.e., elements (cells) were swapped.
- (d) Both had the same Covariance Matrices (Hint: $\Sigma = \frac{1}{N}\sum_{i=1}^{N}[\mathbf{x}_i - \mu][\mathbf{x}_i - \mu]^T$)
- (e) Both had covariance matrices of the same Rank.
- (f) Both had covariance matrices where cells (elements) were swapped.

What Next:?

- Logistics: Make sure that you can use "Shiksha" by Friday. No super hurry.
- Use Channels to Post Queries/Discussions
- Emails at: smai.m2020@gmail.com

1. Revise: Rank of a matrix. Interpretation of Rank.
2. Revise: Bayes Theorem
3. Revise: Eigen Values and Eigen Vectors

- Office Hour This week: Any queries/Doubts on Chapter 2 and 3 of the "Mathematics for Machine Leaning"

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 <u>any</u> data.
3. Actually, we use $\mathcal{D}_{Te}$ <u>also</u> 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

SMAI-M20-06: Data, Distances and Learning

C. V. Jawahar

IIIT Hyderabad

August 21, 2020

Recap:

- Two problems of interest:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
  - Learn Feature Transformation as a step to find useful representations.
- 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
- Performance Metrics:
  - Classification: Accuracy, TP/FP etc., Confusion Matrix; Ranking: Precision, Recall, F-Score, AP
- Supervised Learning:
  - Notion of Training and Testing
  - Notion of Loss Function
  - Role of Optimization

This Lecture:

- Knowing Matrices Better:
  - Rank, Determinant, Trace,
  - Eigen Values and Eigen Vectors
- Supervised Learning
  - Need to generalize
  - Difficulty due to overfitting
- Comparison
  - Distance vs Similarity
  - Samples in R^d
  - Comparison of Sets
  - Comparison of probabiliy distributions

Blank

Questions? Comments?

Comment

What is a good representative for a set $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

Ans: Some one who is close to all.!! Let it be $\mathbf{y}$

$$\min_{\mathbf{y}} \sum_{i=1}^{N} [\mathbf{x}_i - \mathbf{y}]^T [\mathbf{x}_i - \mathbf{y}]$$

$$\min_{\mathbf{y}} \sum_{i=1}^{N} [\mathbf{x}_i^T \mathbf{x}_i - 2\mathbf{y}^T \mathbf{x}_i + \mathbf{y}^T \mathbf{y}]$$

Differentiating wrt to $\mathbf{y}$ and equating to zero [1]

$$\sum_{i=1}^{N} -2\mathbf{x_i} + 2\sum_{i=1}^{N} \mathbf{y} = 0$$

or

$$\mathbf{y} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i$$

[1] Tom Minka, "Old and New Matrix Algebra Useful for Statistics" (Read or refer)

Blank

Discussions Point -I

Consider two ways of comparing two samples $\mathbf{x}$ and $\mathbf{y}$

1. Weighted Euclidean Distance with a Symmetric Positive Definite (PD) matrix $\mathbf{W}$

$$[\mathbf{x} - \mathbf{y}]^T \mathbf{W} [\mathbf{x} - \mathbf{y}]$$

2. We know that PD matrix $\mathbf{W}$ can be decomposed as $\mathbf{L}\mathbf{L}^T$ (popularly known as Cholesky decomposition).
   - We transform the individual samples as:

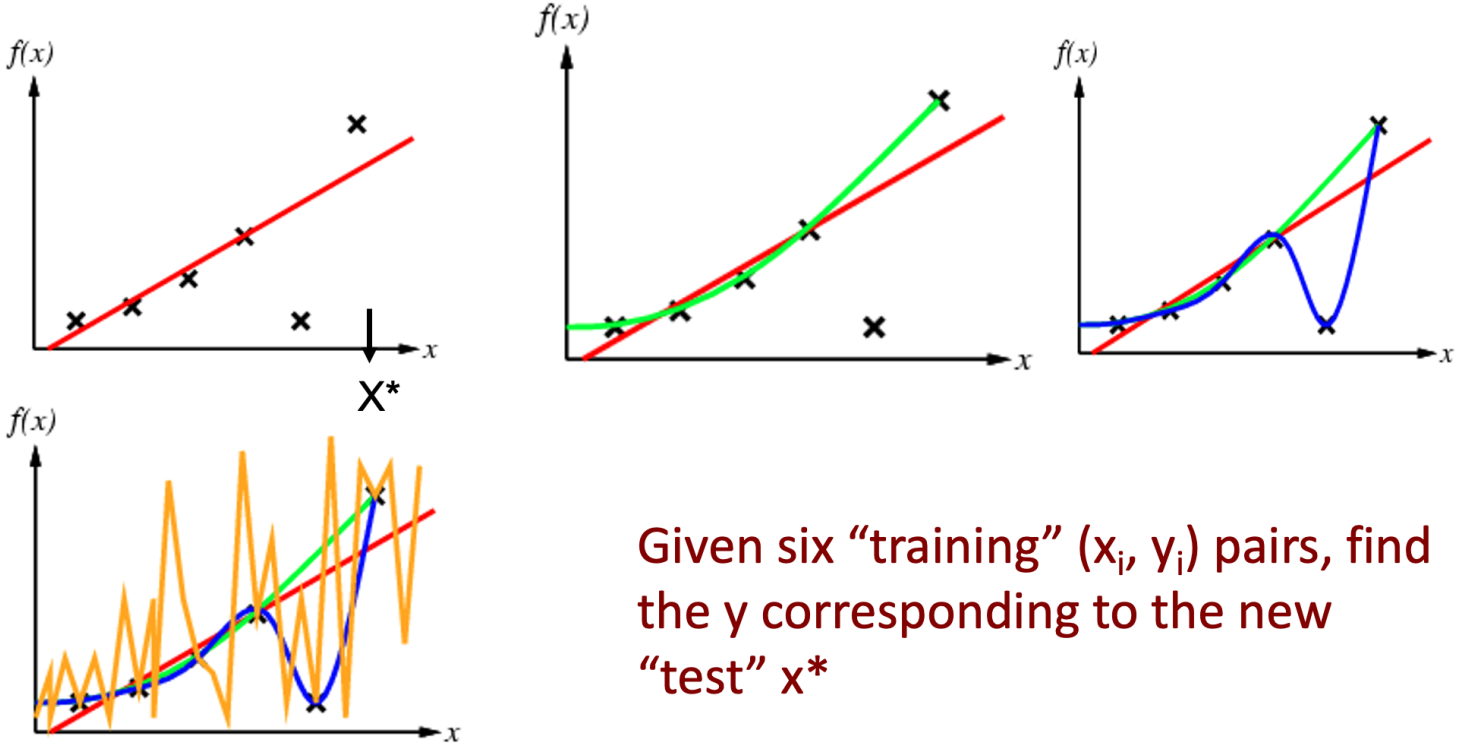
$$\mathbf{x}' = \mathbf{L}^T \mathbf{x}$$

   - Compute the Euclidean distaance between $\mathbf{x}'$ and $\mathbf{y}'$
   - Show that both ways of computing give the same distance.
   - Which one is computationally attractive and when?[2] For example, consider a simple formulation of retrieving K nearest neighbours for a given query $\mathbf{q}$ from a database of N samples.

[2] Connect to the LMNN Paper/method, assuming you had read.

Blank

Discussion Point - II



Given six "training" (x_i, y_i) pairs, find the y corresponding to the new "test" x*

Which curve is the best?

Blank

Discussion Point - III

- We know the $d \times N$ data matrix $\mathbf{X}$ with every column as data elements. Show that $\mathbf{A} = \mathbf{X}\mathbf{X}^T$ is Symmetric. Show that $\mathbf{A}$ is PSD. [3]
- If $\mathbf{A}$ is real symmetric PSD matrix, we know $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathbf{T}$, then what is $\mathbf{V}^\mathbf{T}\mathbf{A}\mathbf{V}$? What is this process called?

[3] A matrix A is PSD if $z^T A z \geq 0 \ \forall z \in R^d$.

Blank

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

If

$$Ax = \lambda x$$

then What are the eigen values and eigen vectors of A^2
i.e., Find:

$$A^2 x = ?$$

- (a) λ, x
- (b) λ^2, x
- (c) λ, $2x$
- (d) λ^2, $2x$
- (e) none of the above

Blank

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

Consider

$$\mathbf{x}'_i = \mathbf{W}\mathbf{x}_i$$

W is constructed as below:

- We start with a $d \times d$ identity matrix
- We randomly permute (rearrange) the <u>columns</u>
- We remove half of the <u>rows</u> and create a $\frac{d}{2} \times d$ matrix **W**

The process of creation of new representation is:

- (a) Feature subset selection; A random subset of the original features will be in $\mathbf{x}'$
- (b) Feature extraction; New features are linear combination of old ones, and not really a subset.
- (c) Dimensionality Reduction; New representation has smaller dimension than the original one.
- (d) This can not be done since these operations are illegal (or mathematically not defined).

Blank

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

Consider three sets

$$A = \{1, 3, 4, 5, 6, 7, 8\}$$

$$B = \{2, 4, 6, 8\}$$

$$C = \{1, 2, 3, 4, 5\}$$

We know Jacard index ($J = \frac{|A \cap B|}{|A \cup B|}$) as a good measure of similarity. Let us use $1 - J$ as a distance.

Does $1 - J$ obey triangular inequality for this set? [4]

(Hint Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y)$)

- (a) YES
- (b) NO
- (c) Triangular inequality is not applicable for this problem.
- (d) Can not be computed.

[4] Advanced (optional): How do we show this for any general three sets?

Blank

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

Consider the problem of feature transformation as

$$\mathbf{x}'_i = \mathbf{W}\mathbf{x}_i$$

If $\mathbf{x}_i \in R^2$ and $\mathbf{W}$ is 2×2 matrix with rank as 1, then

the new points $\mathbf{x}'_i$

- lie on a line in 2D
- are also R^2
- undefined
- One coordinate (dimension) of all the $\mathbf{x}'_i$ will be always the same
- all points in 2D will collapse into a single point. (i.e., $\mathbf{x}'_i = \mathbf{x}'_j$ for all i, j)
- none of the above.

Blank

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

In a TV Game show, a contestant selects one of three doors; behind one of the doors there is a prize,and behind the other two there are no prizes. After the contestant selects a door, the game-show host opens one of the remaining doors, and reveals that there is no prize behind it. The host then asks the contestant whether the you want to SWITCH their choice to the other unopened door, or STICK to their original choice.
[5]

What should we advise? What is the prob. of win if the candidate switch:

- (a) $\frac{1}{3}$ since all the doors are equally likely. Don't switch
- (b) $\frac{1}{2}$ since there are only two left, both are equally likely, no advantage in switching.
- (c) $\frac{2}{3}$. Prefer switching. Bayes says so.
- (d) $\frac{1}{3}$. Don't switching. Bayes says so.
- (e) None of the above.

[5] A very popular problem on internet from khan academy to mit lecture notes!. Appreciate the role of evidence, specially if the answer is not intuitive.

Blank

What Next: Two Sessions?

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

SMAI-M20-07:Examples of Optimization Formulations

C. V. Jawahar

IIIT Hyderabad

August 24, 2020

Recap

- Two problems of interest:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
  - Learn Feature Transformation as a step to find useful representations.
- Algorithms:
  - Nearest Neighbour Algorithm
  - Linear Classification
  - Decide as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2
  - Performance Metrics
- Vectors, Matrices and Data Matrix
  - Comparison of Vectors.
  - Matrix vector multiplication. Differentiation.
  - Properties, Rank, Eigen Values and Eigen Vectors
- Supervised Learning:
  - Notion of Training and Testing
  - Notion of Loss Function
  - Role of Optimization
  - Need of generalization
  - Worry of Overfitting

This Lecture:

- This Lecture:
  - MLE
  - Linear Regression
  - Eigen Vector based optimization
- Problems:
  - How optimization arise?
  - Objective/Loss and Problem Statement
- Solutions:
  - Linear Regression: Differentiation of quadratic function. Unconstrained.
  - Constrained Optimization Problems that lead to solution as eigen vectors
  - Probabilistic formulations, intuitive solutions, interpretations.
  - Closed form solutions (today); iterative in future.

Blank

Questions? Comments?

Blank

Discussions Point -I

We are interested in fitting a solution in presence of outliers. A student came up with an original, yet simple idea. (i) Detect outliers from the data set $\mathcal{D}$ and remove them (ii) Build models on what is left out $\mathcal{D}'$. Here are the formal steps:

1. Find $\mathbf{w}$, the solution for linear regression problem with data set as $\mathcal{D}$.
2. Remove all samples where $|y_i - \mathbf{w}^T\mathbf{x}_i| > \theta$ from $\mathcal{D}$ to create an inlier set $\mathcal{D}'$.
3. Now find a new $\mathbf{w}$ with $\mathcal{D}'$.

- Student tested it with θ as very small (say zero) and very large (say infinity). In both cases, the algorithm came out to be ineffective. What could be the reason?
- Another student helped to solve the problem by coming with a nice intuitive way of finding θ. It did reasonably well. Suggest a simple and effective way to address outliers (or finding θ).

Blank

Discussion Point - II

We are intersted in solving an overdetermined system of homogenous equations

$$\mathbf{A}\mathbf{x} = \mathbf{0}$$

where $\mathbf{A}$ is $m \times n$. $m > n$

Problem is formulated as:

$n \times n$

$(Ax)^T(Ax)$

$x^T x = 1$

$$\arg\min_{\mathbf{x}} ||\mathbf{A}\mathbf{x}|| \quad \text{Subject to: } ||\mathbf{x}|| = 1$$

$x^T A^T A x$

Solution to this problem is:

- (a) Eigen Vector corresponding to the largerst eigen value of $\mathbf{A}\mathbf{A}^T$.
- (b) Eigen Vector corresponding to the largerst eigen value of $\mathbf{A}^T\mathbf{A}$.
- (c) Eigen Vector corresponding to the smallest eigen value of $\mathbf{A}\mathbf{A}^T$.
- (d) Eigen Vector corresponding to the smallest eigen value of $\mathbf{A}^T\mathbf{A}$.

Read later[1]

[1] https://cw.fel.cvut.cz/old/_media/courses/a3m33iro/overdetermined_homogenous_linear_equations.pdf

Blank

Discussion Point - III

- We know the $d \times N$ data matrix $\mathbf{X}$ with every column as data elements. Show that $\mathbf{A} = \mathbf{X}\mathbf{X}^T$ is Symmetric. Show that $\mathbf{A}$ is PSD. [2]
- If $\mathbf{A}$ is real symmetric PSD matrix, we know $\mathbf{A} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\mathbf{T}$, then what is $\mathbf{V}^\mathbf{T}\mathbf{A}\mathbf{V}$? What is this process called?

[2] A matrix A is PSD if $z^T A z \geq 0 \ \forall z \in R^d$.

Blank

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

Consider a square matrix $\mathbf{A}$ with eigen values λ_i and eigen vectors $\mathbf{v}_i$. Then for $\mathbf{A}^T$,

- (a) Eigen values and eigen vectors are the same as that of $\mathbf{A}$.
- (b) Eigen values are the same. Eigen vectors are $\mathbf{v}^T$.
- (c) Eigen values are $\frac{1}{\lambda_i}$
- (d) We can comment for symmetric matrix $\mathbf{A}$. But not for other square matrices.
- (e) None of the above.

Blank

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

Consider the problem of finding the minima of $f(x) = x^2 - 2x + 4$

$$\min_x x^2 - 2x + 4$$

The optimum value of the function f^* is:

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

Blank

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

Consider a data matrix $\mathbf{X}$ where every column is mean subtracted samples. We compute $\mathbf{A} = \mathbf{X}\mathbf{X}^{\mathbf{T}}$. Then $\mathbf{A}$

- is a $d \times d$ matrix
- is a $N \times N$ matrix
- is a symmetric matrix
- is a scaled version of the covariance matrix
- none of the above

Blank

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

The inverse of an upper triangular matrix is:

- (a) Upper triangular
- (b) Lower triangular
- (c) Need not be triangular.
- (c) Will not exist
- (d) None of the above

Blank

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

The the eigen values of $\mathbf{A}$ are λ_i, what are the eigen values of $\alpha\mathbf{A}$, where α is a scalet.

- (a) λ_i itself.
- (b) $\frac{\alpha}{\lambda_i}$
- (c) $\alpha\lambda_i$
- (d) $\frac{1}{\alpha}\lambda_i$
- (e) None of the above.

Blank

What Next: (two sessions?)

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

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

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:

- 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.

Blank

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:

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

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

Blank

Discussion Point - II

We know that:

- Eigen Decomposition of Symmetric Matrix $\mathbf{S}$

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

- SVD of $\mathbf{A}$

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

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

Blank

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:

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

- Write down the objective in matrix form. (Hint: use $\Gamma = Diag(\gamma_i)$). What is the final closed form expression for the $\mathbf{w}$?
- (Advanced) Consider a two step Itertive algorithm:
  - 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)
  - 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

What Next:? (next three)

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

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

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

SMAI-M20-L10: LSI etc.

C. V. Jawahar

IIIT Hyderabad

August 31, 2020

Class Review Questions

1. If $A = UDV^T$, then $A^T A$ is
   - (i) always a rank deficient matrix (ii) VD^2V^T (iii) UD^2U^T (iv) A square matrix (v) is always full rank (vi) can some times be full rank (vi) none of the above etc.
2. Consider a set of general vectors $\mathbf{a}_i \in R^d$. (assume all elements are some random numbers in the range of $[0, 1]$) $\mathbf{b}$ is another such vector. Consider the matrix:

$$\mathbf{A} = \sum_{i=1}^{k} \alpha_i \mathbf{a}_i \mathbf{a}_i^T + \sum_{i=k+1}^{d} \beta_i \mathbf{b}\mathbf{b}^T$$

Comment on the effective rank of $\mathbf{A}$ for various values of α_i and β_i

Announcements:

Post Your queries about:

- Chapters 2, 3 5 in the book.
- Class Reviews in L01-L08
- Micro-Lecture Videos L01-L08

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
  - Occam's razor and role of model complexity
  - Estimating error on validation set.
- 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

This Lecture:

Micro-Lecture Videos

1. Rank of Term-Document Matrix and LSI
   - Q: How do we compare documents? words?
   - Appreciate "one-hot" representation of words.
2. Bias and Variance
   - Yet another balancing act to do while designing ML solution.
3. Loss Function
   - Not all intuitive loss functions are in use due to practical difficulties.

Blank

Questions? Comments?

Blank

Discussions Point -I

Matrix completion: We know completion of rank 1 matrix as:

$$A = \begin{bmatrix} 7 & ? & ? \\ ? & 8 & ? \\ ? & 12 & 6 \\ ? & ? & 2 \\ 21 & 6 & ? \end{bmatrix} \Rightarrow \begin{bmatrix} 7 & 2 & 1 \\ 28 & 8 & 4 \\ 42 & 12 & 6 \\ 14 & 4 & 2 \\ 21 & 6 & 3 \end{bmatrix}$$

Example Problem Formulation:

$$\min \sum_i \sum_j (A_{ij} - B_{ij})^2 \ s.t \ rank(B) = 1$$

Q: What is this summation over? What is the min over?

The rank constraint is not easy to enforce. See alternate formulations based on Nuclear norm[1]. i.e., L1 norm over Diag(D) instead of L0 norm.

[1] Nuclear norm of a matrix is just the sum of the singular values of the matrix

Blank

Discussion Point - II

LSI has many applications [2]

- Compare the documents in the low-dimensional space (data clustering, document classification).
- Find similar documents across languages, after analyzing a base set of translated documents (cross-language information retrieval).
- Find relations between terms (synonymy and polysemy).
- Given a query of terms, translate it into the low-dimensional space, and find matching documents (information retrieval).
- Find the best similarity between small groups of terms, in a semantic way.
- Expand the feature space of machine learning / text mining systems
- Analyze word association in text corpus

Q: **How does it help in Synonymy** ?

Q: **How does it help in Clustering documents** ?

[2] Read: https://en.wikipedia.org/wiki/Latent_semantic_analysis

Blank

Discussion Point - III

Appreciate the fact that there is an important balancing act to be done to get good solutions. Why?

1. overfitting and underfitting
2. bias and variance
3. empirical error and model complexity

Blank

What Next:? (next three)

- Applications and insights into of SVD and Eigen Decomposition
- Supervised Learning: Regularization and Loss Functions
- Bayesian View and Optimal Classification
- PCA and Dimensionality Reduction

SMAI-M20-L11: Semantic Representation for all: "Words", "Documents" "Consumers" and "Products"

C. V. Jawahar

IIIT Hyderabad

September 2, 2020

[0 0 0 1 0 0 0]

Class Review Questions

- Consider a vocabulary of size d. One hot representation of a word i is "1" at the location (index) corresponding to that word and zero else where.
  Given a document that contains P words, $\mathbf{w_1}, \ldots, \mathbf{w_P}$, we compute

$$\mathbf{x} = \sum_{i=1}^{P} \mathbf{w}_i$$

- Consider a document is represented by a histogram of the words in the document. $\mathbf{h}$ i.e., h_i is the number of occurrence of the i th word in the document.
  If the document is paraphrased in certain manner, does the representation change ? (invariant = does not vary/change)
- Can we address paraphrasing if we had pre-multiplied by a certain similarity matrix.

Blank

Recap:

- Problem Space:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
  - Learn useful features
- Supervised Learning:
  - Notion of Training, Validation and Testing
  - Loss Function and Optimization
  - Need of Generalization and Worry of Overfitting
  - Occam's razor and role of model complexity
  - Balancing between Bias and Variance
  - Estimating error using validation set.
- 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
  - LSI,

Representation: Bag of Words and One-Hot

The Bag of Words Representation

I love this movie! It's sweet, but with satirical humor. The dialogue is great and the adventure scenes are fun... It manages to be whimsical and romantic while laughing at the conventions of the fairy tale genre. I would recommend it to just about anyone. I've seen it several times, and I'm always happy to see it again whenever I have a friend who hasn't seen it yet!

it	6
I	5
the	4
to	3
and	3
seen	2
yet	1
would	1
whimsical	1
times	1
sweet	1
satirical	1
adventure	1
genre	1
fairy	1
humor	1
have	1
great	1
...	...

Representation: Bag of Words and One-Hot

Representation: Bag of Words and One-Hot

Representation: Bag of Words and One-Hot

This Lecture:

Micro-Lecture Videos

1. Rank and Recommendation Systems
   - Appreciate why it is low-rank and also incomplete.
   - How do we compare two different "users" or two different "products"?
   - How do we formulate the matrix completion problem?
2. Regularization in Regression
   - Ridge and Lasso. (L2 and L1)
   - Appreciate connection between regularization and overfitting.
3. Orthogonal line fitting
   - Another Eigen vector problem
   - Lead to PCA (to connect to the next lecture).

Questions? Comments?

Blank

Blank

Discussion Point - I

How do we represent words?

1. As one-hot vector
2. As columns of TD Matrix of size $N \times d$
3. After multiplying with U_k^T
4. Word2Vec

Blank

Blank

Blank

Discussions Point -II

Consider the problem of finding solution to an ill-posed problem: Find the solution to $w_1x_1 + w_2x_2 = 10$ There are many and we prefer one that also minimize an Lp norm.

Regularization with Lp norm is popular. Why does L1 norm induce sparsity?

Blank

Blank

Some Plots(from Internet)

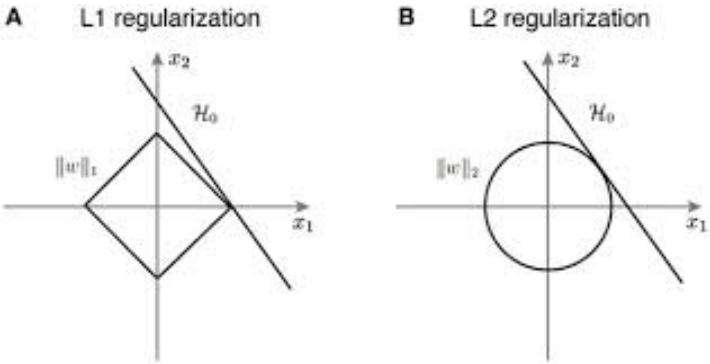
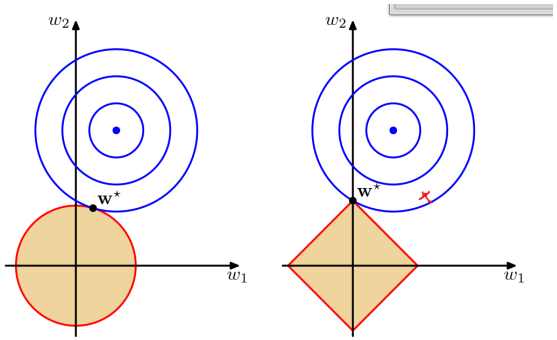
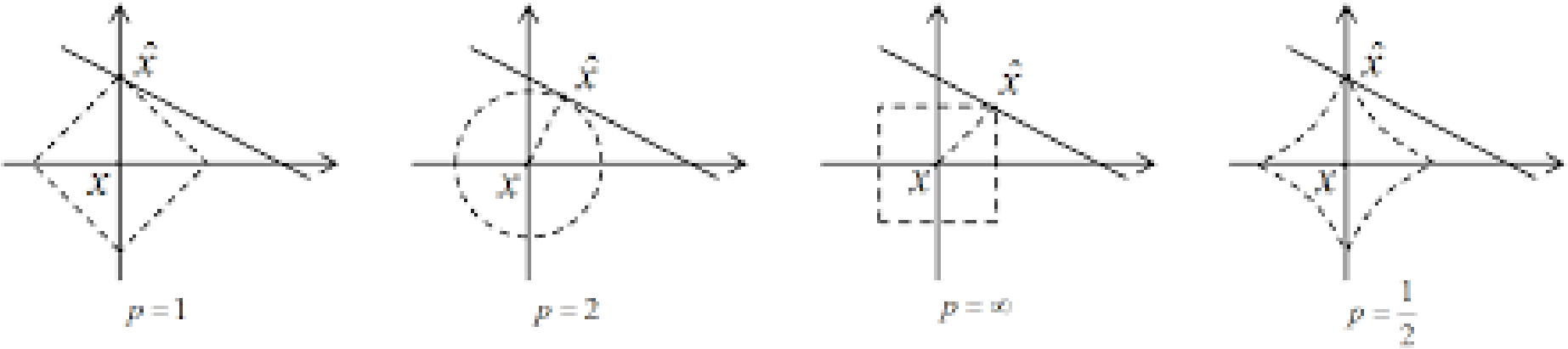
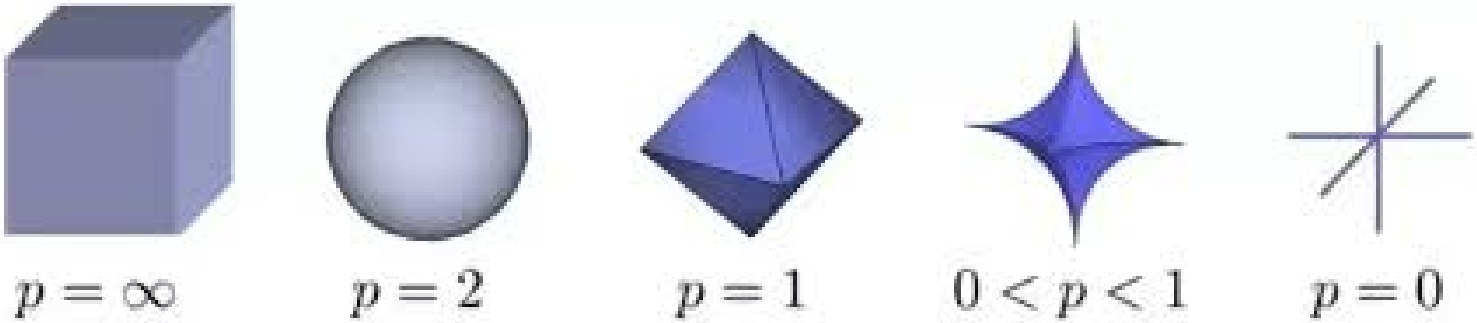


Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$. The lasso gives a sparse solution in which $w_1^\star = 0$.







Blank

What Next:?

1. PCA and Dimensionality Reduction
2. Bayesian Optimal (Cont.)
3. What are good features?

SMAI-M20-L12:PCA

C. V. Jawahar

IIIT Hyderabad

September 4, 2020

Announcements

Topics for Quiz 1 (10 days from now):

- Upto (including PCA) (yet to see some more details)
  - Mathematical Foundations (LA, Prob, Opt.)
  - Supervised Learning (formulation, Simple Algorithms, Bayesian Optimal, Related Concepts, Performance Metrics and Evaluation)
  - Matrix Factorization and Applications
  - Linear Regression
  - PCA
- New topics (Not for Q1): Perceptrons and Gradient Descent. (even if you see these in the class before the Q1)

Class Review

(use notations and conventions from the class) Consider the problem of linear regression where we minimize the loss

$$\mathcal{L}_1 = \frac{1}{N}\sum_{i=1}^{N} \alpha_i (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda_1 g(\mathbf{w})$$

where $g()$ is a regularization term. We also write the loss in matrix form as

$$\mathcal{L}_2 = \frac{1}{N}[Y - \mathbf{X}\mathbf{w}]^T A [Y - \mathbf{X}\mathbf{w}] + \lambda_2 g(\mathbf{w}).$$

- What can we say about the matrix A? size? properties? elements?
- What can we say about the optima, equivalence?
- What can we say when the objective is regularized?
- When regularizing what happens to the objective/optimal solution?

Blank

Recap:

- Problem Space:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
  - Learn useful features
- Supervised Learning:
  - Notion of Training, Validation and Testing
  - Loss Function and Optimization
  - Performance Metrics, Estimating error using validation set.
  - Need of Generalization, overfitting, Occam's razor, model complexity, Bias and Variance, Regularization.
- 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
- Mathematical Foundations: Linear Algebra, Probability, Optimization
  - SVD, Eigen Decomposition, MLE
  - Matrix Factorization in LSI, Recommendation Systems

This Lecture:

$\log a + \log b - \log c - \log d$

$\log \frac{a}{c} + \log \frac{b}{d}$

Micro-Lecture Videos

1. PCA: Dimensions that preserve maximum variance
   - $\mathbf{x}' = \mathbf{W}\mathbf{x}$
   - How to find $\mathbf{W}$?
2. Decision boundaries for Multivariate Gaussians.
   - Analysis under ideal situations.
3. Deep/Neural Embeddings.
   - What are good features?
   - How data can help in discovering features?

$\begin{bmatrix} x' \end{bmatrix} = \begin{bmatrix} \\ \\ \\ \end{bmatrix}$

$k \times d$

$\underline{W}$ $d \times 1$

"unsupervised"

$\log a + \log b - \log c - \log d$

$\log \frac{a}{c} + \log \frac{b}{d}$

$\log(p \times q) - \log \quad \log p + \log q$

Questions? Comments?

Blank

Blank

Discussions Point - I

Regularization with Lp norm is popular. Why does L1 norm induce sparsity?

Let us take an example:

$$\min_{\mathbf{w}} \ (w_1 x_1 + w_2 x_2 - 10)^2 + \lambda g(\mathbf{w})$$

Let $g(\mathbf{w})$ is the Lp norm of $\mathbf{w}$

- First term has many solutions and at all these points the term is zero. (locus is a line)
- We need to find a solution (on this line) that minimizes the $g(\mathbf{w})$.

Blank

Blank

Some Plots(from Internet)

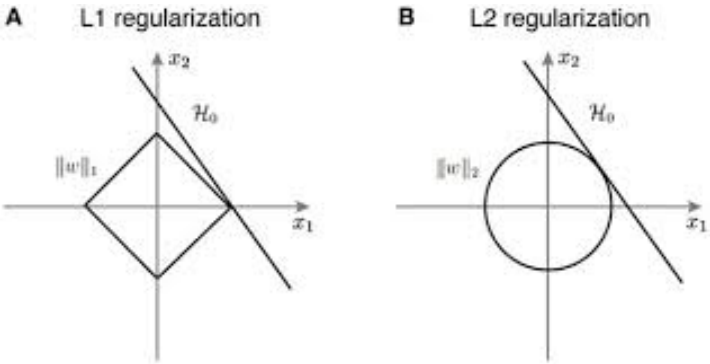
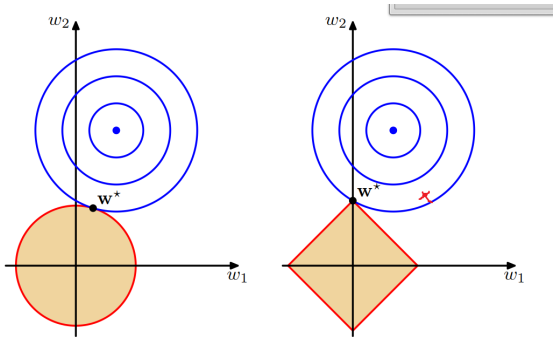
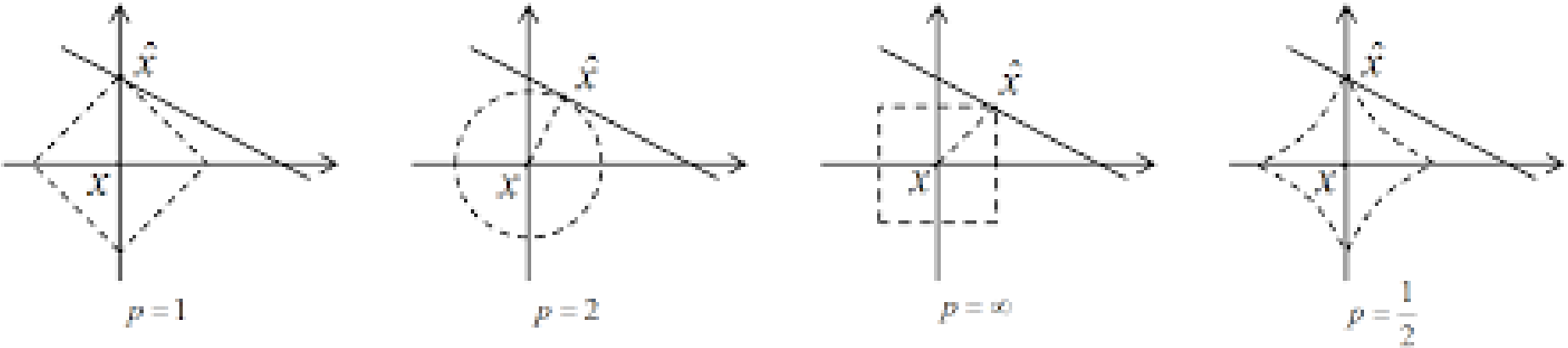
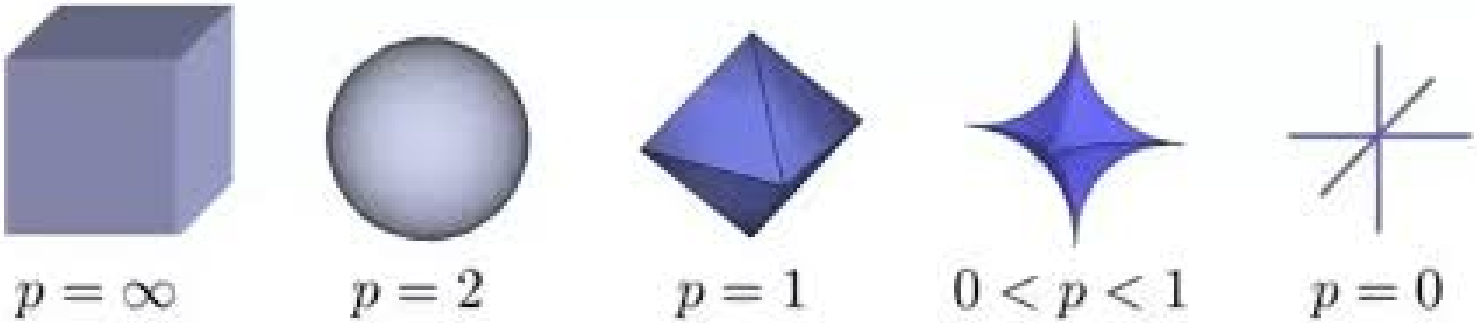


Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$. The lasso gives a sparse solution in which $w_1^\star = 0$.







Blank

Discussions Point -II

Can we have a closed form expression for ridge regression just like the simple linear regression?

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

Hint:

$$[Y - Xw]^T[Y - Xw] + \lambda w^T w = [Y - Xw]^T[Y - Xw] + \lambda w^T(I^T I)w$$

Blank

Blank

What Next:? (next three)

1. PCA and Dimensionality Reduction (more)
2. Bayesian Optimal (Cont.)
3. What are good features?
4. Perceptron Algorithm
5. Gradient Descent Optimization

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:

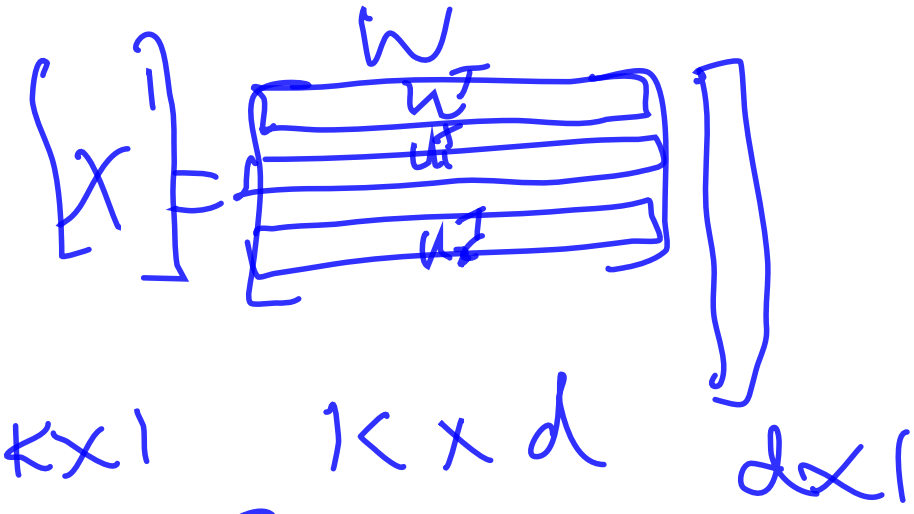
- Problem Space:
  - Learn a function $y = f(\mathbf{W}, \mathbf{x})$ from the data.
  - Dimesnionality Reduction and Representation (Feature Selection, PCA, Neural Embeddings)
  - Matrix Factorization for Data Matrices: (LSI, Matrix Completion, Recommendation Systems)
- Supervised Learning:
  - Notions 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

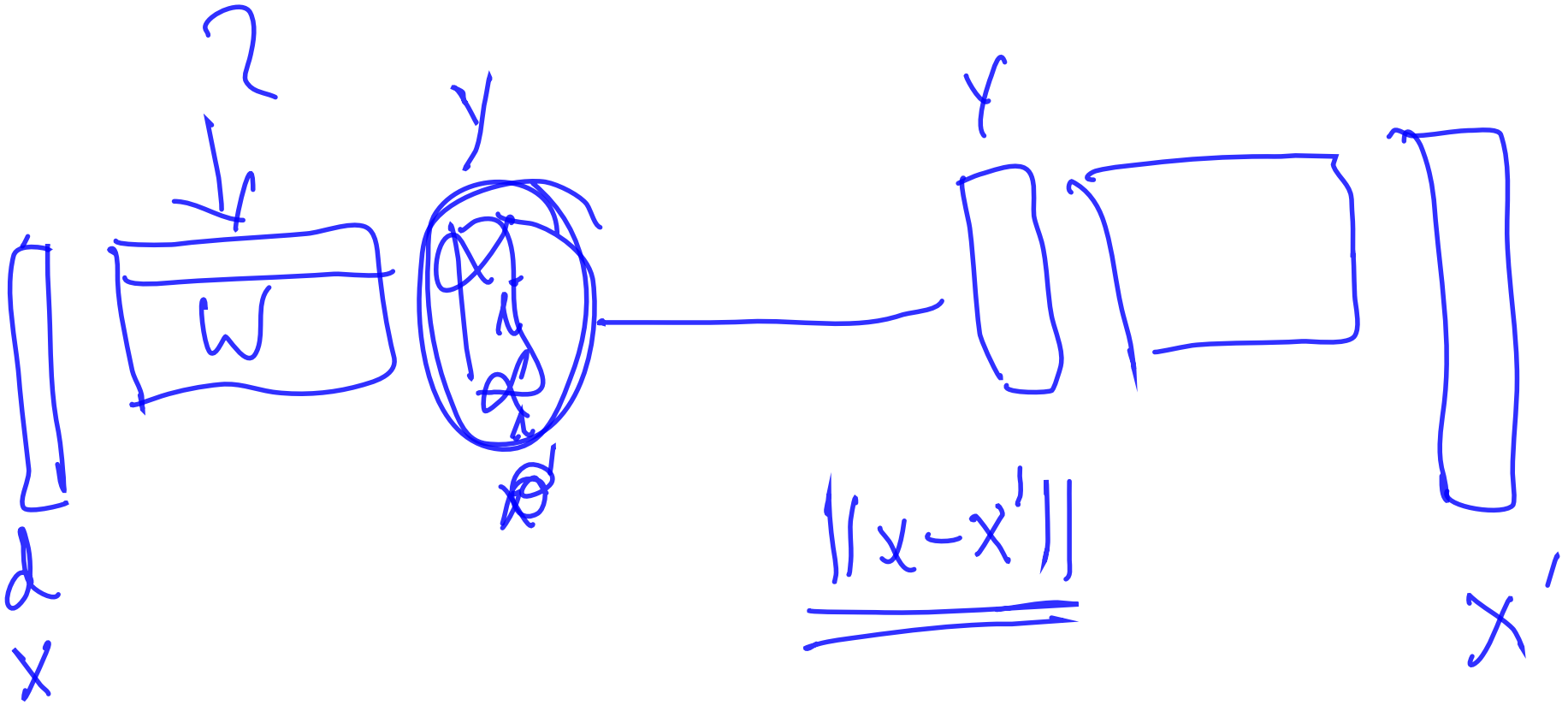


$x \quad d \times 1$

$\Sigma \quad d \times d$

$u \quad d \times 1$

$a^T b = b^T a$



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?

Blank

Blank

Discussions Point - II

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

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)

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

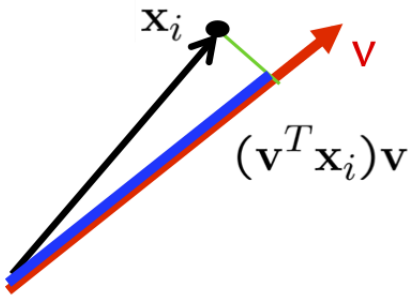
Minimum Reconstruction Error: 1st PC a vector v such that projection on to this vector yields minimum MSE reconstruction

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

blue^2 + green^2 = black^2

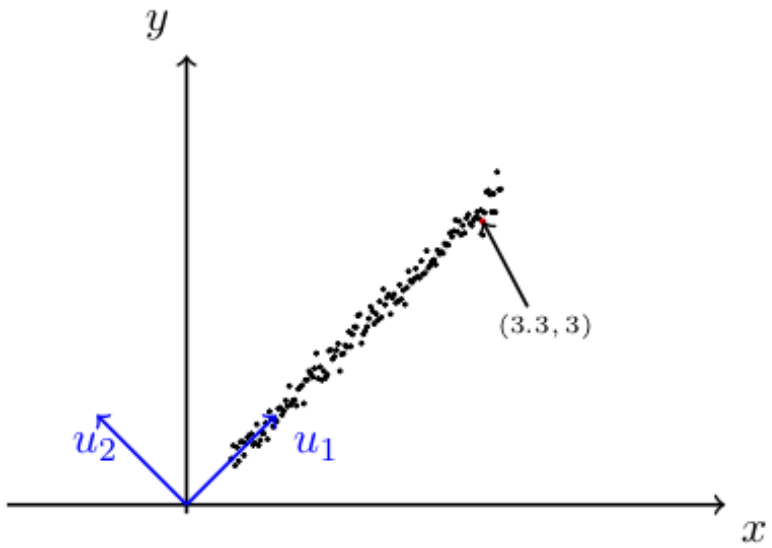
black^2 is fixed (it's just the data)

So, maximizing blue^2 is equivalent to minimizing green^2



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)

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

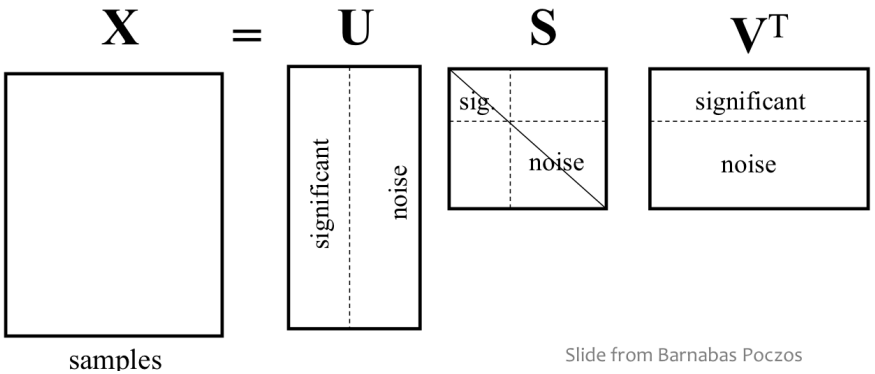
- But we are going to reconstruct it using fewer (only $k = 1 < n$ dimensions, ignoring the low variance u_2 dimension)

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

(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 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 $\mathbf{A} = \mathbf{X}^\mathbf{T}\mathbf{X}$ and $\Sigma = \mathbf{X}\mathbf{X}^\mathbf{T}$
- 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

SMAI-M20-L15: Perceptron Algorithm

C. V. Jawahar

IIIT Hyderabad

September 11, 2020

Class Review

Propeties of SVD

- What are the properties of U, D, V?
- How SVD and EV are related?
- What do we say about SVD of square matrix?
- What can we say about rank deficiency and SVD?

Announcements

1. **Serious Observation**
   - **Unethical collaborations and Practices**:
     - **Open** collaboration. If you want to create any unofficial channels/models/groups make sure that a TA or official representative is there. Talk to the instructor. Take permission.
     - Reposting/sharing class videos/slides/notes without permission
   - Please inform these/similar and anything else that is going on and take corrective actions in next 48 Hrs.

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, Eigen Face
  - Gradient Descent Optimization

This Lecture:

1. MSE using GD (Delta Rule)
   - $\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta y_i \mathbf{x}_i$
   - Appreciate that perceptrons and this are different.
2. Variations in GD (Stochastic and Mini Batch)
   - Single sample, mini batch, batch
   - Stochastic versions
3. Neuron Models
   - $y = \phi(\mathbf{w}^T \mathbf{x})$
   - Step, sigmoid, tanh etc.

Questions? Comments?

Blank

Discussion Point - I

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

Discussions Point - II

We know the AND and OR logic in $\{-1, +1\}$ as

x_1	x_2	y
-1	-1	-1
-1	1	-1
1	-1	-1
1	1	1

x_1	x_2	y
-1	-1	-1
-1	1	1
1	-1	1
1	1	1

1. Which is AND? which is OR?
2. Design a two input neuron with $\phi(z)$ as $sign(z)$ for both AND and OR. Draw pictorially.
3. Can we do NAND and NOR similarly? (Try later)
4. Can we do ExOR? (draw and see). Is it Linearly seperable? Ans: NO

Blank

Blank

Discussions Point - III

Consider the following three samples and their labels $((x_1, x_2), y)$:

$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

Look at the perceptron update rule with $\eta = 0.1$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta \sum_{\mathbf{x}_i \in \mathcal{E}} y_i \mathbf{x}_i$$

Start with line equations given below and do two iterations. Did it converge? If not, how many more iterations will it take?

- line $x_1 = x_2$
- line that pass through (0,2) and (2,0)
- line that pass through (0,4) and (4,0)

Blank

Blank

Blank

Blank

What Next:?

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

SMAI-M20-L16: Perceptrons

C. V. Jawahar

IIIT Hyderabad

September 16, 2020

Announcements

1. **Quiz 1**
   - On 23 Sep. (Either in the class slot or in the Tutorial Slot)
   - All topics except (Gradient Descent, Perceptrons)
   - More instructions will be posted.

Class Review

Consider a perceptron algorithm (batch mode) implementation with initialization $\mathbf{w}^0$ as random initialization learning rate η as 0.1 and termination criteria as "if $||\mathbf{w}^{k+1} - \mathbf{w}^k||_2^2 < 10^{-6}$, terminate".

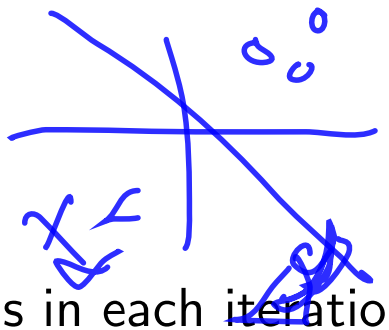
- What happens when the training data is separable and non-separable?
- What can we say about convergence?
- What can we say about error in the training and test data?
- How does the final solution depend on the initialization?
- How does the final solution depend on the learning rate?

Blank

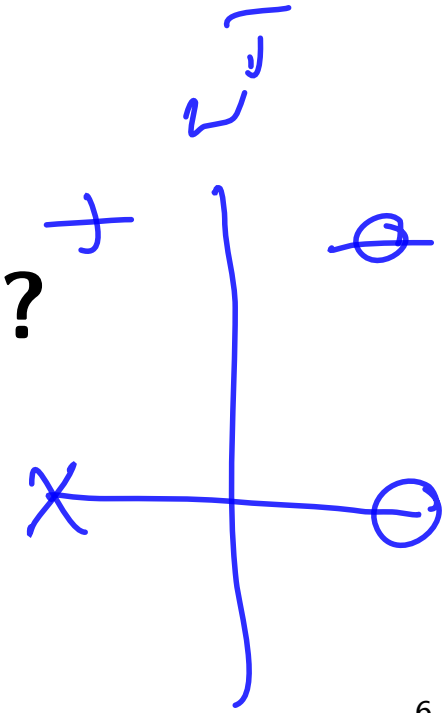
Recap:

- 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.
- Approaches:
  - Optimal Decision as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2
  - MLE
  - Dimesnionality Reduction and Representation (Feature Selection, PCA, Neural Embeddings)
  - Application of PCA: Eigen Face
  - Matrix Factorization for Data Matrices (SVD, Eigen Docomposition)
  - Application of Matrix Factorization: LSI, Matrix Completion, Recommendation Systems)
  - Nearest Neighbour, Linear Discriminants
  - Gradient Descent
  - Linear Regression: Closed form, GD, RegularizationOptimization
  - Perceptron Algorithm and Neuron Model

This Lecture:



1. **Perceptron -II**
   - Appreciate geometrically what happens in each iteration.
2. **Naive Bayes Classifier**
   - An algorithm that makes assumptions; very useful in certain domains.
3. **Three Different Views of Classification**
   - Discriminative.
   - Bayesian under Gaussian Assumptions
   - Nearest Neighbour (Distance based)



Questions? Comments?

Blank

Discussions Point - I

Consider the following three samples and their labels $((x_1, x_2), y)$:

$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

Look at the perceptron update rule with $\eta = 0.1$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta \sum_{\mathbf{x}_i \in \mathcal{E}} y_i \mathbf{x}_i$$

Start with line equations given below and do two iterations. Did it converge? If not, how many more iterations will it take?

- line $x_1 = x_2$
- line that pass through (0,2) and (2,0)
- line that pass through (0,4) and (4,0)

Blank

Classify as + ve if $\mathbf{w}^T\mathbf{x} \geq 0$ else - ve.

1. Start $\mathbf{w}^0 = [-1, -1, 4]^T$. What is $\mathbf{w}^1$?
2. Start $\mathbf{w}^0 = [-1, -1, 2]^T$. What is $\mathbf{w}^1$?
3. Start $\mathbf{w}^0 = [-1, -1, 1.9]^T$. What is $\mathbf{w}^1$?
4. Start $\mathbf{w}^0 = [1, -1, 0]^T$. What is $\mathbf{w}^1$?

Blank

Blank

Blank

Discussions Point -II

Consider a parabolic loss function. We are at $\mathbf{w}^0$.

- Gradient at $\mathbf{w}^0$ only tells us that we need to increase.
- Why don't we find an η that takes us to the optimal solution $\mathbf{w}^*$ in single step? Is it possible at all? (i.e., $\eta = \frac{w^* - w^0}{\Delta}$)

$$w^1 = w^0 - \eta\Delta$$

$$w^1 = w^0 - \frac{w^* - w^0}{\Delta}(-\Delta) = w^*$$

- If No, why? If Yes, Why is this idea not used?

Blank

Blank

Discussion Point - III

Consider a nearest neighbour algorithm (say binary classification) with 1 M training data. Though the KNN is effective, it is computationally not very attractive. (Why?).

A good strategy could be to "prune" the training data with "no loss in accuracy" or sometimes "better generalization". Further read: [1] [2]

1. Is pruning possible? Can a pruned algorithm be as effective to the original (at least on a small toy data that you can think of)?
2. Can we formulate the problem as "selection" of a small set or "computing a small set" (new samples may be different from original)?
3. Should we remove or retain central points or border points?
4. Should we formulate the problem as incremental or decremental selection?

[1] Fast Condensed Nearest Neighbor Rule https://icml.cc/Conferences/2005/proceedings/papers/004_Fast_Angiulli.pdf

[2] Instance Pruning Techniques, http://axon.cs.byu.edu/papers/wilson.icml97.prune.pdf

What Next:?

1. Analysis of Perceptron Algorithm
2. More on Loss Functions
3. Logistic Regression

SMAI-M20-L17: Perceptron and Loss Functions

C. V. Jawahar

IIIT Hyderabad

September 18, 2020

Announcements

- **Quiz 1** Already announced.
- **Some Observations:**
  - Number of views of the video before the lecture session is low. (some times even after the session :-)). May be most of you are very comfortable.
  - Number of attempts for CR is high. Good enough. Performance is OK. No concerns.
  - Number of people who ask queries to TAs for office hours is low. Increased a bit.
  - Ask: (i) Are you putting effort (ii) Are you asking questions at the right forums (iii) Are you coming prepared?
  - Higher Education and Online education expects you to put effort beyond the class rooms.
- **Round up Session on Monday 21**. (no new content)
  - See all videos (and lecture sessions?), if required more than once.
  - Ask specific queries (a form will be shared)
  - Do post by Saturday mid night. (latest by Sunday noon)

Class Review

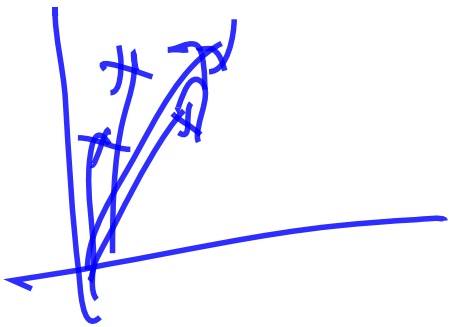
$$\max_{x} \; x^T A x$$
$$\text{s.t.} \quad \|x\| = 1$$

x^*

$$\|w^{k+1} - w^k\| < \epsilon$$

Appreciating PCA

- Is PCA good for compression?
- How does it compare with linear regression?
- Can we have a GD based computation for PCA? (there is a small extra step of norm being unity)
- How are eigen values, vectors related to the covariances?



∇ over a subset

Blank

Recap:

- 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.
- Approaches:
  - Optimal Decision as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2
  - MLE
  - Dimesnionality Reduction and Representation (Feature Selection, PCA, Neural Embeddings)
  - Application of PCA: Eigen Face
  - Matrix Factorization for Data Matrices (SVD, Eigen Docomposition)
  - Application of Matrix Factorization: LSI, Matrix Completion, Recommendation Systems)
  - Nearest Neighbour, Linear Discriminants
  - Gradient Descent
  - Linear Regression: Closed form, GD, RegularizationOptimization
  - Perceptron Algorithm and Neuron Model

This Lecture:

1. Perceptron -II
   - Analysis of Perceptron Algorithm
2. Logistic Regression
   - Probabilitsic view in defining loss/goal. (MLE next)
3. Loss Functions
   - MSE, MAE, Hinge Loss, Cross Entropy

Questions? Comments?

Blank

Blank

Discussions Point - I

Look at three popular functions we know:

A 0-1 loss function

B Hinge loss

C $P(\omega|x)$ in Logistic regression

Comment on:

1. Smoothness
2. Convexity
3. Continuity
4. Differentiability

Blank

Blank

Blank

Blank

Discussion Point - II: Are the perceptrons doing GD?

$$J' = \sum_{i=1}^{N} (t_i - o_i)^2$$

with $o_i = sign(\mathbf{w}^T \mathbf{x}_i)$? There is a problem. With small changes in the $\mathbf{w}$ or the line, the J is not changing. Let us now define:

$$J = \sum_{i=1}^{N} (t_i - o_i)^2 (-\mathbf{w}^T \mathbf{x}_i)$$

This additional terms pulls (or pushes) *proportionally*. Let us now rewrite this J as sum of two parts one over $\mathcal{E}$ and the other on not in $\mathcal{E}$.

$$J = J_1 + J_2 = \sum_{\mathbf{x_i} \in \mathcal{E}} (t_i - o_i)^2 (-\mathbf{w}^T \mathbf{x}_i) + \sum_{\mathbf{x}_i not \in \mathcal{E}} (t_i - o_i)^2 (-\mathbf{w}^T \mathbf{x}_i)$$

$$J = J_1 + J_2 = \sum_{\mathbf{x_i} \in \mathcal{E}} (2 \cdot t_i)^2 (-\mathbf{w}^T \mathbf{x}_i) + \sum_{\mathbf{x}_i not \in \mathcal{E}} 0 \times (\mathbf{w}^T \mathbf{x}_i)$$

We know that J_2 is zero. When $\mathbf{x}_i \in \mathcal{E}$, $(t_i - o_i)$ is $2t_i$ i.e., $2y_i$.

Discussion Point - II: Are the perceptrons doing GD?

- We know either $(t_i - o_i)$ is zero Or we know that $(t_i - o_i)$ is $2t_i$
- And $\frac{\partial(-\mathbf{w}^T\mathbf{x}_i)}{\partial \mathbf{w}} = -\mathbf{x_i}$.

$$\nabla J = -2\sum_{i=1}^{N}(t_i - o_i)\mathbf{x}_i$$

This leads to: $\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \eta\nabla J$ as:

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \eta\sum_{i=1}^{N}(t_i - o_i)(-\mathbf{x}_i)$$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta\sum_{\mathbf{x}\in\mathcal{E}} 2.t_i\mathbf{x}_i$$

$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k + \eta'\sum_{\mathbf{x}_i\in\mathcal{E}} y_i\mathbf{x}_i$$

(with changes in scale for learning rate η)

Blank

Blank

Discussions Point -III: Perceptron Convergence

- **Claim 1** If there exist a set of weights that are consistent with the data, the perceptron algorithm will converge.
- **Claim 2**If the training data is not Linearly Separable, the perceptron algorithm will eventually repeat the same set of weights and thereby enter an infinite loop.
- **Claim 3** If the training data is linearly separable, algorithm will converge in a maximum of N steps. Find N.
- **Claim 4** Every boolean function can be represented by some network of perceptrons only two levels deep.

Read and understand the proofs for the above claims[1]

[1] Page 229 of Duda, Hart and Stork: https://cds.cern.ch/record/683166/files/0471056693_TOC.pdf

Blank

Blank

What Next:? (next three)

1. Logistic Regression
2. Multi Class Classification (beyond binary)
3. More Dimensionality Reductioin Schemes (eg. LDA/Fisher)

SMAI-M20-L18:Roundup Session

C. V. Jawahar

IIIT Hyderabad

September 21, 2020

Class Review

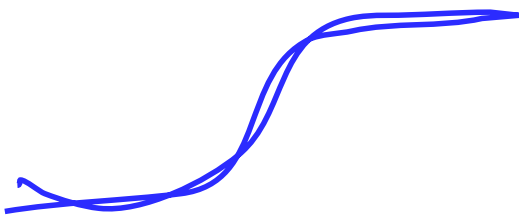
$[0,1]$ $[-1,1]$

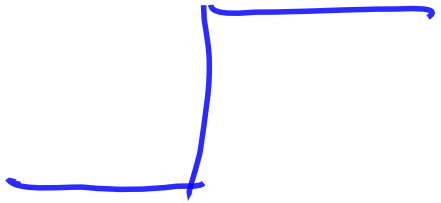
$\frac{1}{1+e^{-z}} \times \frac{e^{z}}{e^{z}}$

Consider the sigmoid function $g(\alpha z) = \frac{1}{1+e^{-\alpha z}}$

- When α varies what happens to this function?
- Where is its range, what is its value at zero, max value, min value,
- How is this related to tanh? $\tanh(z) = \frac{e^{z}-e^{-z}}{e^{z}+e^{-z}}$
- What is the derivative of sigmoid?
- What is $1 - g(z)$?

Sigm
tanh





Blank

Recap:

- 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.
- Approaches:
  - Optimal Decision as ω_1 if $P(\omega_1|\mathbf{x}) \geq P(\omega_2|\mathbf{x})$ else ω_2, MLE
  - Dimesnionality Reduction and Representation (Feature Selection, PCA, Neural Embeddings)
  - Application of PCA: Eigen Face
  - Matrix Factorization for Data Matrices (SVD, Eigen Docomposition)
  - Application of Matrix Factorization: LSI, Matrix Completion, Recommendation Systems)
  - Nearest Neighbour, Linear Discriminants
  - Gradient Descent
  - Linear Regression: Closed form, GD, RegularizationOptimization
  - Perceptron Algorithm and Neuron Model
  - Logistic Regression

This Lecture:

1. Plans and Preparation for the Quiz
2. Topic 1:
   - Feature vector, Data Matrix, Bayesian Optimal Classifier,MLE etc.
   - Eigen Decomposition, SVD, LSI etc.
3. Topic 2: Supervised Learning
   - Training, Testing, Validation, Performance Metrics, Overfitting, Regularization etc.
   - Loss Functions, Optimization, Bias and Variance
4. Topic 3: Algorithms
   - PCA, Linear Regression, Nearest Neighbour, Naive Bayes etc.
   - Perceptrons, Gradient Descent, Logistic Regression (not for Quiz)

Questions? Comments?

Quiz Preparation

- Go through the class material well enough
  - Micro-Lecture Videos
  - Lecture Session (slides/recordings) and some references there in.
  - Class Reviews
  - Home Works
- Read the text book related chapters, Solve problems and practice
- Not expected for Quiz1:
  - Additional references/material list available (moodle)
  - Lecture notes and additional references
  - (https://www.dropbox.com/sh/h91lhc0xpmh2ekw/AAC-FuNgqOO0-txb3FvJqns-a?dl=0)
- Use computer; Keep phone next to you.
- Keep pen, paper and a calculator for your use.
- Strictly avoid any communication with classmates during this period.

Quiz Questions

Example 5 Questions; Need not be the same

1. Test whether you are a student of SMAI 2020 (like CAPTCHA) (may be 10 questions that you can solve fast if you are a student in SMAI-2020) [1]
2. Test whether you not only enrol but also attempt to follow[2] SMAI-2020.
3. Numerical Problems
4. A question set very similar to the regular class review one.
5. A bit more involved. Keep couple of A4 sheets and a paper ready.

[1] CAPTCHA stands for Completely Automated Public Turing test to tell Computers and Humans Apart.

[2] Follow Class Material

Topic - I

- Feature vector, Data Matrix, Bayesian Optimal Classifier,MLE etc.
- Eigen Decomposition, SVD, LSI etc.
- Q: In posterior probability P(w1/X), does X denote class label or feature in the vector x?
- Yes;
- In SVD we do $A = UDV^T$. The dimensions of U are given as m x r, but we know that U is the matrix of eigen vectors of AA^T, and AA^T is symmetric so it is supposed to have independent eigen vectors. If we assume a to be M x N matrix, the dimension of you must be M X M not M X R, sinec all the eigen vectors are linearly independent.

Blank

Blank

Blank

Blank

Topic - II

- Training, Testing, Validation, Performance Metrics, Overfitting, Regularization etc.
- Loss Functions, Optimization, Bias and Variance
- Q: Given a regularization problem, how do we minimise the loss?

Blank

Blank

Blank

Blank

Topic - III

- PCA, Linear Regression, Nearest Neighbour, Naive Bayes etc.
- Perceptrons, Gradient Descent, Logistic Regression (not for Quiz)

Blank

- Q: Steps of minimizing loss functions to obtain $\min||X - X'||$

Blank

Let $\mathbf{u_1}, \ldots \mathbf{u_d}$ be d orthonormal vectors. We can represent the vectors $\mathbf{x}$ as $\sum_{i=1}^{d} \alpha_i \mathbf{u}_i$. Where the scalar α_i is $\mathbf{x}^T \mathbf{u_i}$. However, if we use smaller than d basis vectors, there could be some loss or reconstruction error. Let us consider the loss when we use only one $\mathbf{u}$. i.e.,

$$\mathbf{x} - \mathbf{u}\mathbf{u}^T\mathbf{x}$$

Sum of the reconstruction loss for all the N data samples is now:

$$\sum_{i=1}^{N} ||\mathbf{x_i} - \mathbf{u}\mathbf{u}^T\mathbf{x}_i||^2$$

$$= \sum_{i=1}^{N} \left(\mathbf{x}_i^T\mathbf{x}_i + (\mathbf{u}\mathbf{u}^T\mathbf{x}_i)^T(\mathbf{u}\mathbf{u}^T\mathbf{x}_i) - 2\mathbf{x}_i^T\mathbf{u}\mathbf{u}^T\mathbf{x}_i \right)$$

We would like to minimize this. First term is positive (non negative). It is independnet of $\mathbf{u}$. Therefore, we would like to minimize:

$$\sum_{i=1}^{N} (\mathbf{x_i}^T\mathbf{u}\mathbf{u}^T\mathbf{u}\mathbf{u}^T\mathbf{x_i} - 2\mathbf{x}_i^T\mathbf{u}\mathbf{u}^T\mathbf{x}_i)$$

Cont.

We also know that $\mathbf{u}^T\mathbf{u} = 1$.

$$= \sum_{i=1}^{N} -\mathbf{x}_i^T\mathbf{u}\mathbf{u}^T\mathbf{x}_i = \sum_{i=1}^{N} -\mathbf{u}^T\mathbf{x}_i\mathbf{x}_i^T\mathbf{u} = -\mathbf{u}^T\Sigma\mathbf{u}$$

Minimizing the reconstruction error now becomes that of Maximizing

$$\mathbf{u}^T\Sigma\mathbf{u}$$

with our familiar constraint of $\mathbf{u}^T\mathbf{u} = 1$. This reduces the solution as the eigen vectors corresponding to the largest eigen values.

Blank

Blank

Blank

Blank

What Next:? (next three)

1. Logistic Regression
2. Multi Class Classification (beyond binary)
3. More Dimensionality Reductioin Schemes (eg. LDA/Fisher)