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:

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

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

$[x'] = [\;]_{k \times d} [\;]$ k

W $d \times 1$

"unsuperv"

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?

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

Questions? Comments?

$\log \frac{a}{c} + \log \frac{b}{d}$ $\log(p \times q) - \log$ $\log p + \log q$

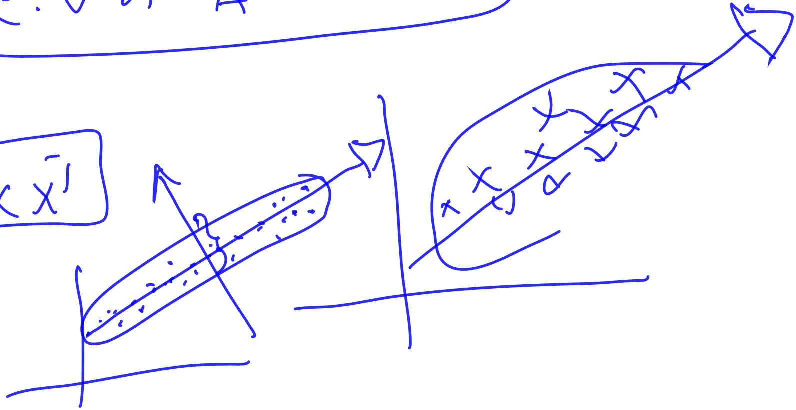
Blank

$w^T A w$ s.t. $w^T w = 1$

e.v of A

$X X^T$

$p(w_1|x) \gtrless p(w_2|x)$



Blank

$$\frac{P(\omega_1|x)}{} \gtrless \frac{P(\omega_2|x)}{}$$

$$\frac{a \cdot b}{\cancel{x}} \lessgtr \frac{d \cdot c}{\cancel{x}}$$

$$\log a + \log b - \log c - \log e \gtrless 0$$

$$a \cdot b \lessgtr d \cdot e$$

$$\log a + \log b \lessgtr \log d + \log e$$

Discussions Point - I

$w_1 x_1 + w_2 x_2 = 10$

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

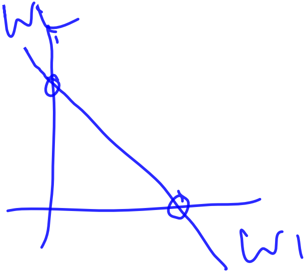
Let us take an example:

$$\min_{\mathbf{w}} (w_1x_1 + w_2x_2 - 10)^2 + \lambda g(\mathbf{w})$$

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

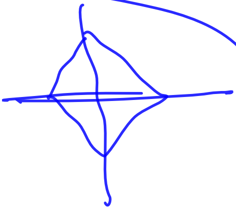
- First term has many solutions and at all these points the term is zero. (locus is a line)
- We need to find a solution (on this line) that minimizes the $g(\mathbf{w})$.

Blank



$$\min_w \ err$$

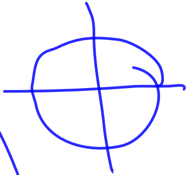
$$\min_w \ err + l_p\,norm$$



$\|w\|_1 = K$

$|w_1| + |w_2| = K$

Lasso

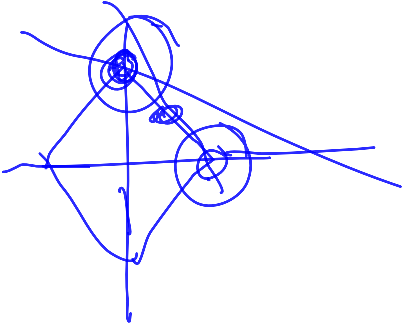


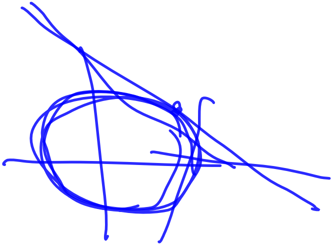
$\|w\|_2 = 0$

$\sqrt{w_1^2 + w_2^2} = K$

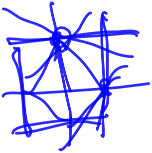
Rige

Blank





L1 norm lead to sparsity.



L2 need not be

Some Plots(from Internet)

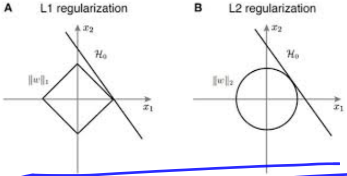
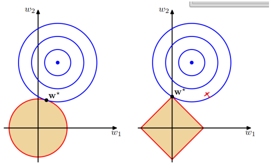
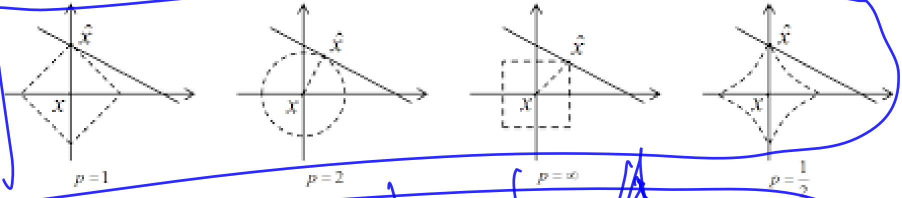
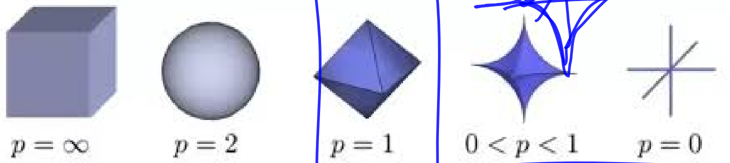


Figure 3.4 Plot of the contours of the unregularized error function (blue) along with the constraint region (3.30) for the quadratic regularizer $q = 2$ on the left and the lasso regularizer $q = 1$ on the right, in which the optimum value for the parameter vector $\mathbf{w}$ is denoted by $\mathbf{w}^\star$.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







Blank

Discussions Point -II

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

$\frac{\partial}{\partial x}(x^T A x)$

$2Ax$

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

$w^T x^T x w$

$$\boxed{\sum_{i=1}^{N}(y_i - \mathbf{w}^T\mathbf{x}_i)^2 + \lambda \mathbf{w}^T\mathbf{w}} - L2$$

Hint:

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

$$\boxed{w = (X^T X + \lambda I)^{-1} X^T Y}$$

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

Blank

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

near singular

$$(X^T X + \lambda I)^{-1}$$

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