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)

$\max x^T A x$

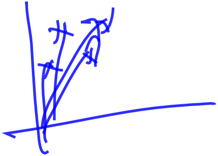
$\text{s.t. } \|x\| = 1$

x^*

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

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

All samples
100

batch

~~stoch~~

Mini batch

one of el

single s

$$w^{k+1} \leftarrow w^{k} - \eta \nabla J$$

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:

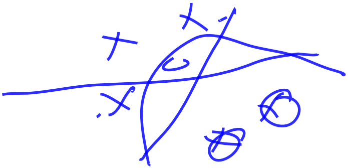
$y_n = +1$

$y_n = -1$

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

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:

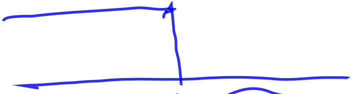
A 0-1 loss function

B Hinge loss

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

Comment on:

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

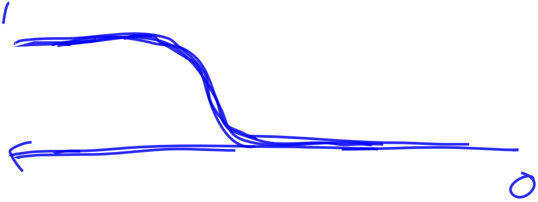


$y_i w^T x_i$

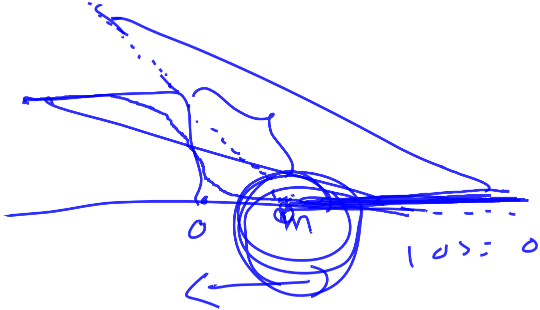
0-1 loss count

y_i	$o_i = \text{sign}(\omega^T x)$
+	+
+	-1
−	+
−	−

Blank



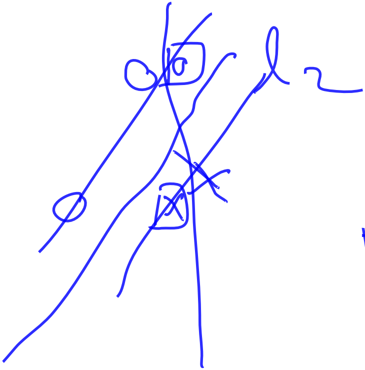
$$\frac{1}{1+e^{c^T x}} \qquad \frac{1}{1+e^{-\alpha w^T x}}$$



$$\boxed{\max(0, m - y_i w^T x)}$$

correct clas small penalty

Blank



$w^T x > m$

$w^T x \le m$

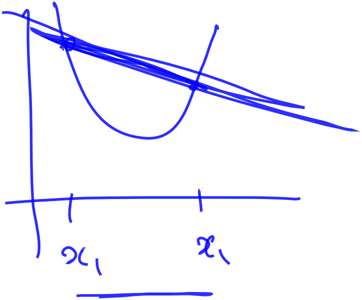
l_1, l_2 are acc for pw

$\mathrm{sing}(w^T x) \gtrless 0$

$w^T x > \underline{m} \cdot 1$

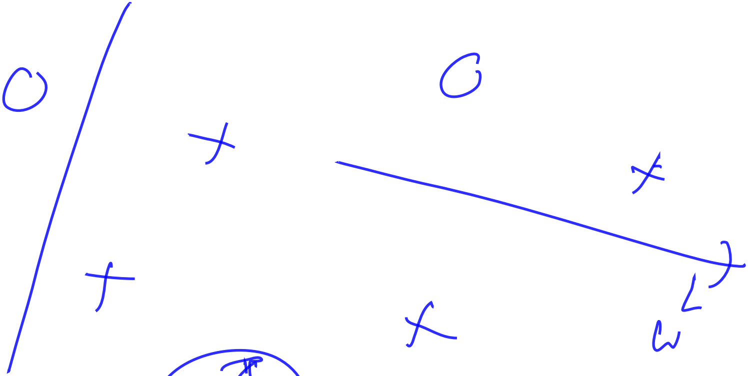
$w^T x < -\underline{m} \cdot 1$

Blank



Convex opt

Blank



$$\frac{\partial E}{\partial w}$$

Discussion Point - II: Are the perceptrons doing GD?

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

$t_i = y_i$

$o_i = sign(w^T x_i)$

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

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

JP

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 $\mathcal{E}$.

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

$t_i = o_i$

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

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:

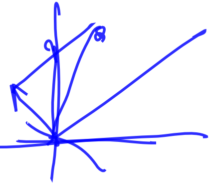
$$\mathbf{w}^{k+1} \leftarrow \mathbf{w}^k - \eta\sum_{i=1}^{N}(t_i - o_i)(-\mathbf{x}_i)$$

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

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

(with changes in scale for learning rate η)

t_i	o_i	$t_i - o_i$
+1	+1	0
−1	−1	0
+1	−1	2
−1	+1	−2



Blank

Blank

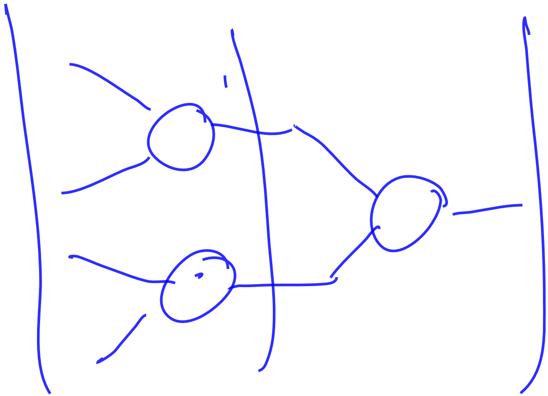
Discussions Point -III: Perceptron Convergence

- **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)