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

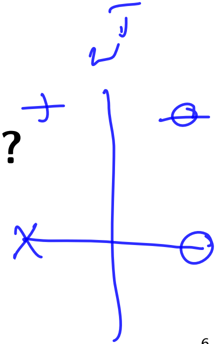
Recap:

- 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 $((x_1, x_2), y)$:

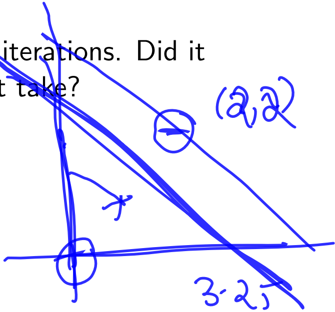
$$\{((1,1),+),\ \ ((2,2),-),\ \ ((0,0),+)\}$$

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

$$\begin{array}{ccc} + & + & - \\ \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} & \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix} & \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} \\ + & + & - \end{array}$$

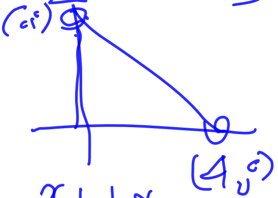
if $w^T x \geq 0$ $+$

else $-$

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 $\mathbf{w}^1$?

$$w^0 = \begin{bmatrix} -1 \\ -1 \\ 4 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 1 \\ 4 \end{bmatrix}$$



$x_1 + x_2 = 4$

$-x_1 - x_2 = -4$

$$w^1 = \begin{bmatrix} -1 \\ -1 \\ 4 \end{bmatrix} + 0.1 \begin{bmatrix} 2 \\ 2 \\ 1 \end{bmatrix} \times (-1)$$

$$= \begin{bmatrix} -1.2 \\ -1.2 \\ 3.9 \end{bmatrix} = \begin{bmatrix} -1 \\ -1 \\ 3.25 \end{bmatrix}$$

Blank

Blank

Blank

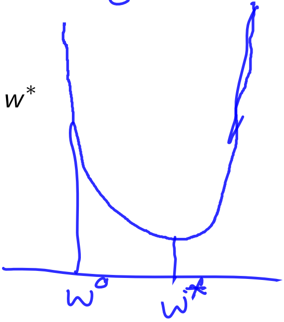
Discussions Point -II

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., $\eta = \frac{w^* - w^0}{\Delta}$) $\leftarrow \eta^*$

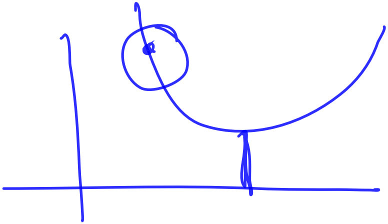
$$w^1 = w^0 - \eta\Delta$$

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

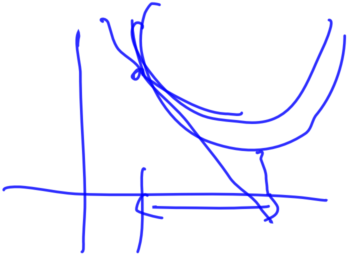


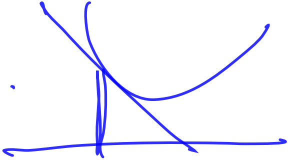
- If No, why? If Yes, Why is this idea not used?

$$\boxed{\eta^* = \frac{\|\nabla J\|^2}{\nabla J^T H \nabla J}}$$



Assume:
Fn is quadr





$+ \quad -$

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

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