ECE595 / STAT598: Machine Learning I
Lecture 17.3: Perceptron 2 - Convergence

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

Outline

Discriminative Approaches

- Lecture 16 Perceptron 1: Definition and Basic Concepts
- Lecture 17 Perceptron 2: Algorithm and Property
- Lecture 18 Multi-Layer Perceptron: Back Propagation

This lecture: Perceptron 2

- Perceptron Algorithm
  - Loss Function
  - Algorithm
- Optimality
  - Uniqueness
  - Batch and Online Mode
- Convergence
  - Main Results
  - Implication

Convergence of Perceptron Algorithm

Theorem. Assume the following things:

- The two classes are linearly separable
- This means: $(\boldsymbol{\theta}^*)^T(y_j\boldsymbol{x}_j) = y_j((\boldsymbol{w}^*)^T\boldsymbol{x}_j + w_0^*) \geq \gamma$ for some $\gamma > 0$
- $\|\boldsymbol{x}_j\|_2 \leq R$ for some constant
- Initialize $\boldsymbol{\theta}^{(0)} = \mathbf{0}$

Then, batch mode perceptron algorithm converges to the true solution $\boldsymbol{\theta}^*$

$$\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\|^2 = 0,$$

when the number of iterations k exceeds

$$k \geq \frac{\|\boldsymbol{\theta}^*\|^2 R^2}{\gamma^2}.$$

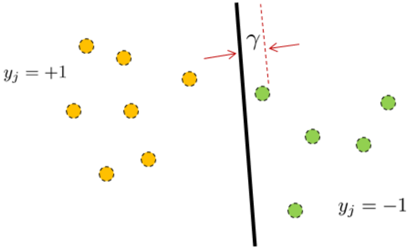
Interpreting the Perceptron Convergence

Theorem. Assume the following things:

- The two classes are linearly separable
- This means: $(\boldsymbol{\theta}^*)^T(y_j\boldsymbol{x}_j) = y_j((\boldsymbol{w}^*)^T\boldsymbol{x}_j + w_0^*) \geq \gamma$ for some $\gamma > 0$
- $\|\boldsymbol{x}_j\|_2 \leq R$ for some constant
- Initialize $\boldsymbol{\theta}^{(0)} = \mathbf{0}$

Comment.

- γ is the margin
- $\boldsymbol{\theta}^*$ is ONE solution such that the margin is at least γ



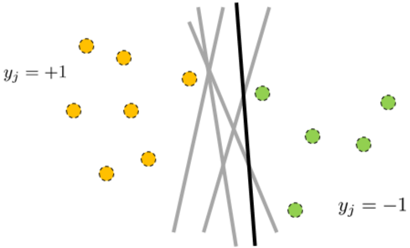
Interpreting the Perceptron Convergence

Theorem. Assume the following things:

- The two classes are linearly separable
- This means: $(\boldsymbol{\theta}^*)^T(y_j\boldsymbol{x}_j) = y_j((\boldsymbol{w}^*)^T\boldsymbol{x}_j + w_0^*) \geq \gamma$ for some $\gamma > 0$
- $\|\boldsymbol{x}_j\|_2 \leq R$ for some constant
- Initialize $\boldsymbol{\theta}^{(0)} = \mathbf{0}$

Comment.

- If you do not initialize at $\mathbf{0}$, still converge.
- The solution $\boldsymbol{\theta}^*$ might be different.



Interpreting the Perceptron Convergence

Then, batch mode perceptron algorithm converges to the true solution $\boldsymbol{\theta}^*$

$$\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\|^2 = 0$$

when the number of iterations k exceeds

$$k \geq \frac{\|\boldsymbol{\theta}^*\|^2 R^2}{\gamma^2}.$$

Comment:

- You can turn batch mode to online mode by picking only one $j \in \mathcal{M}_k$
- You will do slower, but you can still converge
- $\boldsymbol{\theta}^*$ is the converging point of *this* particular sequence $\{\boldsymbol{\theta}^1, \boldsymbol{\theta}^2, \ldots \boldsymbol{\theta}^\infty\}$
- Not an arbitrary separating hyperplane

Interpreting the Perceptron Convergence

Then, batch mode perceptron algorithm converges to the true solution $\boldsymbol{\theta}^*$

$$\|\boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^*\|^2 = 0,$$

when the number of iterations k exceeds

$$k \geq \frac{\|\boldsymbol{\theta}^*\|^2 R^2}{\gamma^2}.$$

Comment:

- R controls the radius of the class.
- Large R: Wide spread. Difficult. Need large k.
- γ controls the margin.
- Large γ: Big margin. Easy. Need small k.

Summary of the Convergence Theorem

- **Algorithm**: You use gradient descent on $J_{\text{soft}}(\boldsymbol{\theta})$
- **Solution**: You get a global minimizer for $J_{\text{hard}}(\boldsymbol{\theta})$
- But this is just one of the many global minimizers
- **Assumption**: Linearly separable
- If not linearly separable, then will oscillate
- **Margin**: At optimal solution there is a margin because separable
- Applications: Not quite; There are many better methods
- Theoretical usage: Good for analyzing linear models. Very simple algorithm.

Reading List

Perceptron Algorithm

- Abu-Mostafa, Learning from Data, Chapter 1.2
- Duda, Hart, Stork, Pattern Classification, Chapter 5.5
- Cornell CS 4780 Lecture https://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote03.html
- UCSD ECE 271B Lecture http://www.svcl.ucsd.edu/courses/ece271B-F09/handouts/perceptron.pdf