Deep Learning - Theory and Practice

IE 643
Lecture 3

August 5, 2022.

1. Recap
   - Perceptron and Learning

2. Perceptron Convergence

Recap: Perceptron and Learning

Perceptron

- Input is of the form $x = (x_1, x_2, \ldots, x_d) \in \mathbb{R}^d$.
- We associate weights $w = (w_1, w_2, \ldots, w_d) \in \mathbb{R}^d$ to the connections.
- **Prediction Rule:**
  - $\langle w, x \rangle \geq \theta \implies$ predict 1.
  - $\langle w, x \rangle < \theta \implies$ predict -1.

Perceptron - Geometric Idea

Prediction Rule

- $\langle w, x \rangle \geq \theta \implies$ predict 1.
- $\langle w, x \rangle < \theta \implies$ predict -1.

Geometric Idea: To find a separating hyperplane.

Perceptron - Geometric Idea

Hyperplane in $\mathbb{R}^d$

$\mathcal{H} = \{x \in \mathbb{R}^d : \exists w \neq 0 \text{ and } b \in \mathbb{R} \text{ s.t. } \langle w, x \rangle = b\}$.

Perceptron - Geometric Idea

Open Halfspaces associated with a hyperplane $\langle w, x \rangle = b$

$\mathcal{Q}_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle > b\}$ and $\mathcal{Q}_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle < b\}$.

Perceptron - Geometric Idea

Open Halfspaces associated with a hyperplane $\langle w, x \rangle = b$

$\mathcal{Q}_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle > b\}$ and $\mathcal{Q}_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle < b\}$.

Perceptron - Geometric Idea

Closed Halfspaces associated with a hyperplane $\langle w, x \rangle = b$

$\mathcal{S}_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle \geq b\}$ and $\mathcal{S}_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle \leq b\}$.

Perceptron - Geometric Idea

Closed Halfspaces associated with a hyperplane $\langle w, x \rangle = b$

$\mathcal{S}_1 = \{x \in \mathbb{R}^d : \langle w, x \rangle \geq b\}$ and $\mathcal{S}_2 = \{x \in \mathbb{R}^d : \langle w, x \rangle \leq b\}$.

Perceptron - Geometric Idea

Prediction Rule

- $\langle w, x \rangle \geq \theta \implies$ predict 1.
- $\langle w, x \rangle < \theta \implies$ predict -1.

Geometric Idea: To find a separating hyperplane (w, θ) such that samples with class labels 1 and -1 lie on alternate sides of the hyperplane.

Perceptron - Geometric Idea

Prediction Rule

- $\langle w, x \rangle \geq \theta \implies$ predict 1.
- $\langle w, x \rangle < \theta \implies$ predict -1.

Geometric Idea: To find a separating hyperplane (w, θ) such that samples with class labels 1 and -1 lie on alternate sides of the hyperplane.

Perceptron

- Input is of the form $\tilde{x} = (x, 1) = (x_1, x_2, \ldots, x_d, 1) \in \mathbb{R}^{d+1}$.
- We associate weights $\tilde{w} = (w, -\theta) = (w_1, w_2, \ldots, w_d, -\theta) \in \mathbb{R}^{d+1}$ to the connections.
- **Prediction Rule:**
  - $\langle \tilde{w}, \tilde{x} \rangle = \langle w, x \rangle - \theta \geq 0 \implies$ predict 1.
  - $\langle \tilde{w}, \tilde{x} \rangle = \langle w, x \rangle - \theta < 0 \implies$ predict -1.

Perceptron - Data Perspective

- **Input: data point** $x = (x_1, x_2, \ldots, x_d)$, **label** $y \in \{+1, -1\}$.

Perceptron - Training

Perceptron Training Procedure

$\tilde{w}^1 = 0$
for $t \leftarrow 1, 2, 3, \ldots$ **do**
    receive (x^t, y^t), $x^t \in \mathbb{R}^d$, $y^t \in \{+1, -1\}$.
    Transform x^t into $\tilde{x}^t = (x^t, 1) \in \mathbb{R}^{d+1}$.
    $\widehat{y} = Perceptron(\tilde{x}^t; \tilde{w}^t)$
    if $\widehat{y} \neq y^t$ **then**
        $\tilde{w}^{t+1} = \tilde{w}^t + y^t \tilde{x}^t$
    else
        $\tilde{w}^{t+1} = \tilde{w}^t$

Perceptron Update Rule - Geometric Idea

- The current weights are $\tilde{w}^t = (1, 1, -4)$.

Perceptron Update Rule - Geometric Idea

- Suppose a new point $x^t = (2, 1)$, $y^t = 1$ arrives.

Perceptron Update Rule - Geometric Idea

- With the current weights, perceptron outputs $\hat{y}^t = -1$ for the point $x^t = (2, 1)$.
- An error occurs and $\tilde{w}^t$ gets updated to $\tilde{w}^{t+1} = \tilde{w}^t + y^t \tilde{x}^t$, .
- After update, the new weights become $\tilde{w}^{t+1} = (3, 2, -3)$.

Convergence of Perceptron Training

Perceptron Convergence - Geometric Intuition

- What are some natural assumptions to expect the perceptron training to converge?
- Let us first motivate such assumptions through geometric intuition.

Perceptron Convergence - Geometric Intuition

- Can the data be separated by a hyperplane?

Perceptron Convergence - Geometric Intuition

- **First assumption:** At least the data should be such that the samples with label 1 can to be separated by a hyperplane from samples with label -1.

Perceptron Convergence - Geometric Intuition

- **First assumption:** At least the data should be such that the samples with label 1 can to be separated by a hyperplane from samples with label -1.
- Is this assumption sufficient?

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

Perceptron Convergence - Geometric Intuition

- **Refined assumption:** We not only want the data to be separated but the separation should be good enough!

Perceptron Convergence - Separability Assumption

Linear Separability Assumption

Let $D = \{(x^t, y^t)\}_{t=1}^{\infty}$ denote the training data where $x^t \in \mathbb{R}^d$, $y^t \in \{+1, -1\}$, $\forall t = 1, 2, \ldots$. Then there exist $\mathbb{R}^d \ni w^* \neq 0,\ \gamma > 0$, such that:

$$\langle w^*, x^t \rangle > \gamma \text{ where } y^t = 1,$$
$$\langle w^*, x^t \rangle < -\gamma \text{ where } y^t = -1.$$

Perceptron Convergence - Separability Assumption

Linear Separability Assumption

Let $D = \{(x^t, y^t)\}_{t=1}^{\infty}$ denote the training data where $x^t \in \mathbb{R}^d$, $y^t \in \{+1, -1\}$, $\forall t = 1, 2, \ldots$. Then there exist $\mathbb{R}^d \ni w^* \neq 0,\ \gamma > 0$, such that:

$$y^t \langle w^*, x^t \rangle > \gamma.$$

Perceptron Convergence - Mistake Bound

- We will try to derive useful bounds on the number of mistakes that a perceptron can commit during its training.
- **Assumption on data:** Linear Separability
- Assume that the T rounds of training have been completed in perceptron training. Assume T to be some large number.
- Assume that M mistakes are made by the perceptron in these T rounds. (Obviously, $M \leq T$.)
- We ask if the number of mistakes M can be bounded by some suitable quantity.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- **First step:** To bound the difference $\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle$.

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\begin{aligned}\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &= \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle \\ &= \langle w^*, w^t \rangle + \langle w^*, y^t x^t \rangle - \langle w^*, w^t \rangle \text{ (how?)}\end{aligned}$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\begin{aligned}
\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &= \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, w^t \rangle + \langle w^*, y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, y^t x^t \rangle
\end{aligned}$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\begin{aligned}
\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &= \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, w^t \rangle + \langle w^*, y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, y^t x^t \rangle \\
&= y^t \langle w^*, x^t \rangle \text{ (how?)}
\end{aligned}$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\begin{aligned}
\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle &= \langle w^*, w^t + y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, w^t \rangle + \langle w^*, y^t x^t \rangle - \langle w^*, w^t \rangle \\
&= \langle w^*, y^t x^t \rangle \\
&= y^t \langle w^*, x^t \rangle > \gamma
\end{aligned}$$

Perceptron Convergence - Mistake Bound

- We begin the analysis by considering an arbitrary round $t \in \{1, 2, \ldots, T\}$ where a mistake is made by the perceptron.
- **Recall:** During this t-th round:
  - Perceptron computes $\hat{y}^t = \text{sign}(\langle w^t, x^t \rangle)$.
  - $\hat{y}^t \neq y^t$.
  - Perceptron update: $w^{t+1} = w^t + y^t x^t$.
- Now from linear separability assumption, we have $w^* \neq 0$ such that $y^t \langle w^*, x^t \rangle > \gamma$.
- Now we can write

$$\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle > \gamma.$$

Perceptron Convergence - Mistake Bound

- Now when no mistake is made in round t, we have $w^{t+1} = w^t$.
- Hence $\langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = 0$.

Perceptron Convergence - Mistake Bound

Recall our assumptions:

- Assume that the T rounds of training have been completed in perceptron training. Assume T to be some large number.
- Assume that M mistakes are made by the perceptron in these T rounds. (Obviously, $M \leq T$.)

Perceptron Convergence - Mistake Bound

$$\sum_{t=1}^{T} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \sum_{\substack{t \in \{1,\ldots,T\}, \\ t\text{:mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle + \sum_{\substack{t \in \{1,\ldots,T\}, \\ t\text{:no mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle$$

Perceptron Convergence - Mistake Bound

$$\sum_{t=1}^{T} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \sum_{\substack{t \in \{1,\ldots,T\}, \\ t\text{:mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle +$$

$$\sum_{\substack{t \in \{1,\ldots,T\}, \\ t\text{:no mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle$$

$$= \sum_{\substack{t \in \{1,\ldots,T\}, \\ t\text{:mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle + 0 \text{ (how?)}$$

Perceptron Convergence - Mistake Bound

$$\sum_{t=1}^{T} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle = \sum_{\substack{t \in \{1,\ldots,T\}, \\ t\text{:mistake is made} \\ \text{at round } t}} \langle w^*, w^{t+1} \rangle - \langle w^*, w^t \rangle$$

$$> M\gamma \text{ (how?)}$$