The Perceptron

Cornell CS 3/5780 · Spring 2026

0. The Curse of Dimensionality

- Points drawn from a probability distribution tend to never be close together in high dimensions.
- **Volume Analysis**: For uniform distribution on features, to capture k neighbors in a unit cube $[0,1]^d$, the required edge length $\ell^d \approx k/n$
- **Question**:
  - What happens to ℓ for k/n fixed and d getting big?
  - How big does n need to get to keep ℓ constant?
- **Mitigating the Curse**:
  - Linear Separation: Pairwise distances between points grow with dimensionality, but distances to hyperplanes do not.
  - Low Dimensional Structure: Data often lies on low-dimensional manifolds despite a high-dimensional d.

1. Perceptron Classifier

- **Core Assumption**: Binary classification with $y_i \in \{-1, +1\}$ and data that is *linearly separable*
- **Classification Rule**: Determined by which side of a hyperplane the input $\mathbf{x}$ is on.
- **Formally**: given by direction $\mathbf{w}$ and bias b

$$h(\mathbf{x}_i) = \text{sign}(\mathbf{w}^\top \mathbf{x}_i + b)$$

- Without the bias term, the hyperplane that $\mathbf{w}$ defines would always have to go through the origin.

2. Simplified Formulation

- **Absorbing bias term**: add one additional *constant* dimension

$$\mathbf{x}_i \quad \text{becomes} \quad \begin{bmatrix} \mathbf{x}_i \\ 1 \end{bmatrix}, \qquad \mathbf{w} \quad \text{becomes} \quad \begin{bmatrix} \mathbf{w} \\ b \end{bmatrix}$$

- **New formulation**: under the new definition of $\mathbf{x}$ and weight $\mathbf{w}$,

$$h(\mathbf{x}_i) = \text{sign}(\mathbf{w}^\top \mathbf{x})$$

- **Key Observation**: Note that $\mathbf{x}_i$ is classified correctly (i.e. on the correct side of the hyperplane) if

$$y_i(\mathbf{w}^\top \mathbf{x}_i) > 0$$

3. Perceptron Algorithm

Input: Training data $D = \{(\mathbf{x}_1, y_1), ..., (\mathbf{x}_n, y_n)\}$

Initialize: $\mathbf{w} = \mathbf{0}$

While TRUE:

1. set $m = 0$
2. for $(\mathbf{x}_i, y_i) \in D$
   - if $y_i(\mathbf{w}^\top \mathbf{x}_i) \leq 0$
     - $\mathbf{w} = \mathbf{w} + y_i \mathbf{x}_i$
     - $m = m + 1$
3. if $m = 0$: break

4. Perceptron Convergence

- **Guarantee**: If a data set is linearly separable, Perceptron finds a separating hyperplane in finite steps.
- **Separability**: $\exists \mathbf{w}^*$ such that $y_i(\mathbf{x}^\top \mathbf{w}^*) > 0$, for all $(\mathbf{x}_i, y_i) \in D$.
- **Rescaling**: weights, features such that $\|\mathbf{w}^*\| = 1$, $\|\mathbf{x}_i\| \leq 1 \ \ \forall \, \mathbf{x}_i \in D$
- **Margin**: the distance γ from the hyperplane to the closest data point:

$$\gamma = \min_{(\mathbf{x}_i, y_i) \in D} |\mathbf{x}_i^\top \mathbf{w}^*|$$

- **Key Observation**: For all $\mathbf{x}$ we must have $y(\mathbf{x}^\top \mathbf{w}^*) = |\mathbf{x}^\top \mathbf{w}^*| \geq \gamma$.

5. Convergence Theorem

- **Theorem**: For separable data with margin γ, the Perceptron algorithm makes at most $1/\gamma^2$ mistakes.
- **Question**: What is more desirable, a large margin or a small margin? When will the Perceptron converge quickly?

- **Fact 1**: for misclassified $\mathbf{x}$, we have $y(\mathbf{x}^\top \mathbf{w}) \leq 0$
- **Fact 2**: for any $\mathbf{x}$, we have $y(\mathbf{x}^\top \mathbf{w}^*) > \gamma$ due to margin (previous slide)

6. Convergence Proof part 1

- Consider the effect of an update on $\mathbf{w}^\top \mathbf{w}^*$:

$$\begin{aligned}(\mathbf{w} + y\mathbf{x})^\top \mathbf{w}^* &= \mathbf{w}^\top \mathbf{w}^* + y(\mathbf{x}^\top \mathbf{w}^*) \\ &\geq \mathbf{w}^\top \mathbf{w}^* + \gamma\end{aligned}$$

- Consider the effect of an update on $\mathbf{w}^\top \mathbf{w}$:

$$\begin{aligned}(\mathbf{w} + y\mathbf{x})^\top (\mathbf{w} + y\mathbf{x}) &= \mathbf{w}^\top \mathbf{w} + 2y\mathbf{w}^\top \mathbf{x} + y^2(\mathbf{x}^\top \mathbf{x}) \\ &\leq \mathbf{w}^\top \mathbf{w} + y^2(\mathbf{x}^\top \mathbf{x}) \\ &\leq \mathbf{w}^\top \mathbf{w} + 1\end{aligned}$$

- This means that for each update, $\mathbf{w}^\top \mathbf{w}^*$ grows by at least γ and $\mathbf{w}^\top \mathbf{w}$ grows by at most 1.

7. Convergence Proof part 2

- We initialize $\mathbf{w} = 0$. Hence, initially $\mathbf{w}^\top \mathbf{w} = 0$ and $\mathbf{w}^\top \mathbf{w}^* = 0$.
- After M updates, (1) $\mathbf{w}^\top \mathbf{w}^* \geq M\gamma$ and (2) $\mathbf{w}^\top \mathbf{w} \leq M$
- Starting with (1) and ending with (2)

$$\begin{aligned} M\gamma &\leq \mathbf{w}^\top \mathbf{w}^* \\ &= \|\mathbf{w}\| \|\mathbf{w}^*\| \cos(\theta) \\ &\leq \|\mathbf{w}\| \\ &= \sqrt{\mathbf{w}^\top \mathbf{w}} \leq \sqrt{M} \end{aligned}$$

- Rearranging $M\gamma \leq \sqrt{M}$, we conclude $M \leq 1/\gamma^2$

8. Summary

- The Perceptron is a binary linear classifier
- We absorb the bias term by adding a constant feature dimension
- Guaranteed to converge if data is linearly separable
  - Number of mistakes bounded by $1/\gamma^2$ where γ is the margin
  - Larger margins lead to faster convergence
- Cannot solve non-linearly separable problems (like XOR)