MAS473 Introduction to DL - Final Summary

Yujun Kim

November 2023

1 Generalization

- Risk is expected loss of a predictor.
- Optimal predictor of binary classification need conditional distribution of Y given x, which is not known in general.
- Likelihood ratio test(LRT)
- MLE is LRT. MAP under uniform priori is LRT.

1.1 Perceptron

- **Theorem (Mistake Bound)** Perceptron algorithm makes at most $\frac{2+D(S)^2}{\gamma(S)^2}$ margin mistakes for any linearly separable data S.
  This guarantees convergence to a perfect classifier(w.r.t. train data). - Why?
- **Theorem (Generalization Bound)**

$$P(Yw(S_n)^T X < 1 \leq \frac{1}{n+1}\mathbb{E}_{S_{n+1}}\left[\frac{2+D(S_{n+1})^2}{\gamma(S_{n+1})^2}\right]$$

  This implies a good generalization if trained with many samples. Why?

1.2 Generalization Gap

Generalization Gap. $\Delta_{gen}(f) = \mathcal{R}[f] = \mathcal{R}_S[f]$
Basic analysis using Hoeffding's inequality. For a single function f. With high probability,

$$\Delta_{gen}(f) \leq \sqrt{\frac{\log(1/\delta)}{2n}}$$

- **Algorithmic Stability.** $\Delta(\mathcal{A}) = \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^{n}(loss(\mathcal{A}(S), Z_i) - loss(\mathcal{A}(S^{(i)}), Z_i'))\right]$
  Average stability is expected generalization gap.

- **Uniform Stability.** $\Delta_{sup}(\mathcal{A}) = \sup_{S,S',d_H(S,S')=1} \sup_z |loss(\mathcal{A}(S), z) - loss(\mathcal{A}(S'), z)|$
  Theorem (ERM is uniformly stable) If loss is strongly convex, $L-$Lipschitz with respect to w in the domain,
$$\Delta_{sup}(ERM) \leq \frac{4L^2}{\mu n}$$
- **Finite hypothesis.**
$$\Delta_{gen} \leq \sqrt{\frac{\log(|\mathcal{F}|) + \log(1/\delta)}{2n}}$$
- **VC Dimension.** Measure of maximum number of arbitrary classifiable points. Size of largest set shattered.
$$\Delta_{gen} \leq \sqrt{\frac{VCDim(\mathcal{F})\log(n) + \log(1/\delta)}{n}}$$
  If $||x|| \leq D$
- **(Empirical) Radamacher Complexity.** Measure degree of complexity to fit arbitrary label.

2 Dimension Reduction with PCA

3 Clustering