Machine Learning Notes 12.23

Notes Group

December 30, 2025

1 Dimensionality Reduction

In high-dimensional data analysis, we often wish to reduce the dimension of the data while preserving its intrinsic structure. We consider a dataset $X = \{x_1, \ldots, x_n\}$ where $x_i \in \mathbb{R}^d$. We seek a mapping $\phi : \mathbb{R}^d \to \mathbb{R}^k$ with $k < d$.

1.1 Approaches

1. **PCA (Principal Component Analysis):** Finds a linear projection that minimizes reconstruction error (or equivalently, maximizes variance).

$$\min_P \sum_x \|x - P(x)\|^2$$

2. **Random Projection (Johnson-Lindenstrauss):** Finds a linear mapping that preserves pairwise Euclidean distances between points.

1.2 The Johnson-Lindenstrauss (JL) Lemma

The JL Lemma provides a guarantee that a simple random linear projection can preserve pairwise distances with high probability, provided the target dimension k is large enough. Notably, k depends logarithmically on the number of samples n, but is independent of the original dimension d.

1.2.1 Problem Statement

Let $x_1, \ldots, x_n \in \mathbb{R}^d$. We want to find a linear mapping $\phi : \mathbb{R}^d \to \mathbb{R}^k$ such that for a given error tolerance $\varepsilon > 0$, the following condition holds for all $i, j \in [n]$:

$$(1-\varepsilon)\|x_i - x_j\|^2 \le \|\phi(x_i) - \phi(x_j)\|^2 \le (1+\varepsilon)\|x_i - x_j\|^2 \tag{1}$$

1.2.2 The Theorem

The central question is: How large must k be to guarantee that such a ϕ exists?

Theorem 1 (Johnson-Lindenstrauss Lemma)**.** *Let $x_1, \ldots, x_n \in \mathbb{R}^d$. For any $\varepsilon > 0$, let*

$$k = O\left(\frac{\ln n}{\varepsilon^2}\right) \quad \textit{(specifically } k \ge \frac{8 \ln n}{\varepsilon^2}\textit{).}$$

Then, there exists a linear mapping $\phi : \mathbb{R}^d \to \mathbb{R}^k$ such that for all $i, j \in [n]$:

$$(1-\varepsilon)\|x_i - x_j\|^2 \le \|\phi(x_i) - \phi(x_j)\|^2 \le (1+\varepsilon)\|x_i - x_j\|^2.$$

Key Insights:

- The required target dimension k is logarithmic in the number of data points n.
- k is independent of the original dimension d. This is powerful for extremely high-dimensional data.
- The mapping ϕ is typically realized using a random matrix (e.g., Gaussian entries).

1.2.3 Proof of JL Lemma

We prove the existence of such a map using the Probabilistic Method with a Gaussian random matrix.

Step 1: Construction of the Map Let k be an integer. We define $\phi : \mathbb{R}^d \to \mathbb{R}^k$ as:

$$\phi(x) = \frac{1}{\sqrt{k}} Ax$$

where A is a $k \times d$ matrix with independent entries $A_{ij} \sim \mathcal{N}(0, 1)$.

Step 2: Distribution of the Projected Norm Let $u = x_i - x_j$. Without loss of generality, assume $\|u\|^2 = 1$. We analyze the random variable $\|\phi(u)\|^2$.

$$\|\phi(u)\|^2 = \left\| \frac{1}{\sqrt{k}} Au \right\|^2 = \frac{1}{k} \sum_{m=1}^{k} (A_m \cdot u)^2$$

where A_m is the m-th row of A. Since $A_{ij} \sim \mathcal{N}(0, 1)$, the dot product $Y_m = A_m \cdot u$ is a sum of independent Gaussians, so $Y_m \sim \mathcal{N}(0, 1)$. Consequently, Y_m^2 follows a Chi-squared distribution with 1 degree of freedom. The scaled norm corresponds to a sum of k such variables:

$$k\|\phi(u)\|^2 = \sum_{m=1}^{k} Y_m^2 = r, \quad \text{where } r \sim \chi_k^2$$

The condition $(1 - \varepsilon) \leq \|\phi(u)\|^2 \leq (1 + \varepsilon)$ becomes:

$$(1 - \varepsilon)k \leq r \leq (1 + \varepsilon)k$$

Step 3: Concentration of Measure We apply the concentration inequality for the Chi-squared distribution.

Lemma 1 (Concentration of χ_k^2)**.** *Let $r \sim \chi_k^2$. For $\varepsilon \in (0, 1)$:*

$$\Pr\left((1 - \varepsilon)k \leq r \leq (1 + \varepsilon)k\right) \geq 1 - 2\exp\left(-\frac{k}{2}\left(\frac{\varepsilon^2}{2} - \frac{\varepsilon^3}{3}\right)\right)$$

Let E_{ij} be the event that the pair (x_i, x_j) is distorted by more than ε. The probability of failure for a single pair is:

$$\Pr(E_{ij}) \leq 2\exp\left(-k\left(\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}\right)\right)$$

Step 4: Union Bound There are $\binom{n}{2} < \frac{n^2}{2}$ pairs. By the Union Bound:

$$\Pr(\exists i, j : E_{ij}) \leq \frac{n^2}{2} \cdot 2\exp\left(-k\left(\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}\right)\right)$$

To guarantee existence, we need this probability < 1:

$$n^2 \exp\left(-k\left(\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}\right)\right) < 1$$

Taking logs and rearranging:

$$k > \frac{2\ln n}{\frac{\varepsilon^2}{4} - \frac{\varepsilon^3}{6}} \approx \frac{8\ln n}{\varepsilon^2}$$

Thus, for $k = O(\frac{\ln n}{\varepsilon^2})$, the random projection succeeds with high probability. □

2 Generalization: An Algorithm-Dependent View

Classic generalization theory (VC Theory) focuses on the complexity of the hypothesis class. However, modern machine learning often deals with over-parameterized models where classic bounds become vacuous. We shift perspective to **Algorithmic Stability**.

2.1 Review: VC Theory vs. Modern Regime

- **VC Theory (Algorithm-Independent):** Focuses on the capacity of the model class $\mathcal{H}$.

$$\text{Gen.Gap} \leq \tilde{O}\left(\sqrt{\frac{\text{VC}(\mathcal{H})}{n}}\right)$$

  This assumes *under-parametrization* where $n \gg \text{VC}(\mathcal{H})$.

- **Modern Deep Learning (Over-parametrization):** Often $\text{VC}(\mathcal{H}) \gg n$ or #params $\gg n$. In this regime, the VC bound might suggest a gap > 1 (vacuous), yet models generalize well in practice. We need a theory that depends on the algorithm itself (e.g., how SGD selects a specific solution).

2.2 Algorithmic Stability

Stability measures how much the output of a learning algorithm changes if we perturb the training dataset slightly (e.g., by changing one example).

2.2.1 Notation

- Algorithm A.
- Training set $S = (z_1, \ldots, z_n)$.
- Neighboring dataset $S^i = (z_1, \ldots, z_{i-1}, z_i', z_{i+1}, \ldots, z_n)$, where the i-th example z_i is replaced by an independent sample z_i'.
- Loss function $\ell(\cdot, \cdot)$.
- $A(S)$ denotes the hypothesis (classifier) learned by algorithm A on set S.

Definition 1 (Uniform Stability)**.** *An algorithm A is said to have **uniform stability** β with respect to a loss function ℓ if for all training sets S, all neighboring sets S^i, and all data points z:*

$$|\ell(A(S), z) - \ell(A(S^i), z)| \leq \beta(n)$$

Desired Behavior:

- **Stable:** $\beta(n) = O\left(\frac{1}{n}\right)$ ($\checkmark$)
- **Unstable:** $\beta(n) = \Omega(1)$ ($\times$)

2.3 Stability Implies Generalization

If an algorithm is stable, its empirical risk is a good proxy for its true risk.

2.3.1 Definitions

- **True Risk:** $R(A(S)) = \mathbb{E}_z[\ell(A(S), z)]$
- **Empirical Risk:** $R_{emp}(A(S)) = \frac{1}{n}\sum_{i=1}^{n} \ell(A(S), z_i)$

Theorem 2. *Assume the loss function is bounded, i.e., $0 \leq \ell(\cdot, \cdot) \leq M$. Assume algorithm A is symmetric (invariant to permutation of S). If A has uniform stability β, then:*

$$\mathbb{E}_S[R(A(S)) - R_{emp}(A(S))] \leq \beta$$

2.3.2 Proof

We aim to bound the expected generalization gap:

$$\mathbb{E}_S[R(A(S)) - R_{emp}(A(S))]$$

1. Decompose the expectation:

$$\mathbb{E}_S[R_{emp}(A(S))] = \mathbb{E}_S\left[\frac{1}{n}\sum_{i=1}^{n}\ell(A(S), z_i)\right]$$

By symmetry of the algorithm and i.i.d. data, the expectation is the same for any index i. Thus:

$$\mathbb{E}_S[R_{emp}(A(S))] = \mathbb{E}_S[\ell(A(S), z_i)]$$

2. Introduce Ghost Sample: Let S^i be the dataset where z_i is replaced by z'_i. Since z_i and z'_i are i.i.d., and A is symmetric:

$$\mathbb{E}_S[\ell(A(S), z_i)] = \mathbb{E}_{S,z'_i}[\ell(A(S^i), z'_i)]$$

(Essentially, evaluating the model trained on S against point z_i is statistically identical to evaluating the model trained on S^i against point z'_i).

3. Expand the True Risk: The true risk is the expected loss on a fresh point z'_i:

$$\mathbb{E}_S[R(A(S))] = \mathbb{E}_{S,z'_i}[\ell(A(S), z'_i)]$$

4. Combine and Bound: Substituting these back into the generalization gap expression:

$$\begin{aligned}\mathbb{E}_S[R(A(S)) - R_{emp}(A(S))] &= \mathbb{E}_{S,z'_i}[\ell(A(S), z'_i)] - \mathbb{E}_{S,z'_i}[\ell(A(S^i), z'_i)] \\ &= \mathbb{E}_{S,z'_i}\left[\ell(A(S), z'_i) - \ell(A(S^i), z'_i)\right]\end{aligned}$$

By the definition of uniform stability, for any S, S^i and test point z'_i:

$$\left|\ell(A(S), z'_i) - \ell(A(S^i), z'_i)\right| \leq \beta$$

Therefore:

$$\mathbb{E}_S[R(A(S)) - R_{emp}(A(S))] \leq \beta$$

This proves that for a stable algorithm, the generalization gap vanishes as $\beta \to 0$ (typically as $1/n$).