1 Generalization

Goal of This Chapter. Bounding risk. Risk is sum of empirical risk and generalization gap. ER is minimized using ERM and we upper bound generalization gap using VCDim, Rademacher complexity etc.
Risk $\mathcal{R}[\hat{Y}] = \mathbb{E}[loss(\hat{Y}(X), Y)]$
Optimal Predictor $\hat{Y}(x) = \mathbb{1}\left\{P(Y=1|X=x) \geq \frac{l(1,0)-l(0,0)}{l(0,1)-l(1,1)}P(Y=0|X=x)\right\}$, obtained by comparing $\mathbb{E}[loss(1,Y)|X], \mathbb{E}[loss(0,$
LRT $\hat{Y}(x) = \mathbb{1}\{\mathcal{L}(x) \geq \eta\}, \mathcal{L}(x) = \frac{P(x|Y=1)}{P(x|Y=0)}$. Optimal predictor is LRT by Bayes.
MLE is LRT with $\eta = 1$, as $\hat{Y}(x) = argmax_{y\in\{0,1\}}P(X=x|Y=y)$
MAP is LRT if has uniform prior and $\frac{l(1,0)-l(0,0)}{l(0,1)-l(1,1)} = 1$. In this case, MAP is equivalent to MLE.
Empirical Loss of f on $S = \{(x_i, y_i)\}$ is $\mathcal{R}_S[f] = \frac{1}{n}\sum_i l(f(x_i), y_i)$
Empirical Loss minimizer in given function class $\mathcal{F}$ is $argmin_{f\in\mathcal{F}}\mathcal{R}_S[f]$

1.1 Perceptron

Perceptron Algorithm for linearly separable data. $w_0 = 0$, Take i random at each iteration. $w_{t+1} = w_t + y_i x_i$ if margin mistake($y_i\langle w_t, x_i\rangle < 1$) don't update otherwise.
Theorem (Mistake Bound) Perceptron algorithm makes at most $\frac{2+D(S)^2}{\gamma(S)^2}$ margin mistakes for any linearly separable data S. γ is max-min margin, D is diameter. (proof) For optimal predictor w^*, $||w_t|| \leq m_t(2 + D(S)^2)$ and $||w_t|| \geq \langle w^*.w_t\rangle = \sum_{k=1}^{t} < w^*.w_k - w_{k=1}\rangle \geq m_t\gamma(S, w^*) = m_t\gamma(s)$ This guarantees convergence to a perfect classifier(w.r.t. train data). - Why?
Theorem (Generalization Bound) $P(Yw(S_n)^TX < 1 \leq \frac{1}{n+1}\mathbb{E}_{S_{n+1}}\left[\frac{2+D(S_{n+1})^2}{\gamma(S_{n+1})^2}\right]$ (proof) Based on leave one out set from $n+1$ data and use previous theorem. 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$(1-\delta)$, $\Delta_{gen}(f) \leq \sqrt{\frac{\log(1/\delta)}{2n}}$
Average Stability. $\Delta(\mathcal{A}) = \mathbb{E}_{S,S'}\left[\frac{1}{n}\sum_{i=1}^{n}(loss(\mathcal{A}(S), Z_i) - loss(\mathcal{A}(S^{(i)}), Z_i'))\right]$
Proposition. Average stability is expected generalization gap. i.e. $\mathbb{E}[\Delta_{gen}(\mathcal{A}(S))] = \Delta(\mathcal{A})$ (proof) Use S'

Uniform Stability. $\Delta_{sup}(\mathcal{A}) = \sup_{S,S',d_H(S,S')=1}\sup_z |loss(\mathcal{A}(S), z) - loss(\mathcal{A}(S'), z)|$ upper bounds average stability.
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. with probability $1-\delta$, $\Delta_{gen} \leq \sqrt{\frac{\log(|\mathcal{F}|)+\log(1/\delta)}{2n}}$
VC Dimension. is the size of largest set shattered by the function class. With probability $1-\delta$, $\Delta_{gen} \leq \sqrt{\frac{VCDim(\mathcal{F})\log(n)+\log(1/\delta)}{n}}$.
(Empirical) Radamacher Complexity. ERC: $\hat{\mathcal{R}}_n(\mathcal{L}) = \mathbb{E}\left[\sum_h \in \mathcal{L}\frac{1}{n}\sum_i \sigma_i h(z_i)\right]$, RC: $\mathbb{E}_S[\hat{\mathcal{R}}_n(\mathcal{L})]$

2 Dimension Reduction with PCA

Goal of This Chapter. Characterizing the PCA by two equivalent formulation: Variance maximization and Error Minimization. Formulating PPCA using MLE.
$\mathcal{X} = \{x_1, \cdots, x_N\}, x_n \in \mathbb{R}^D$, mean 0 data. Covariance matrix $S = \frac{1}{n}\sum x_n x_n^T$
Maximum variance Perspective Low dimensional projection to column orthonormal $B = [b_1, \cdots, b_M] \in \mathbb{R}^{D\times M} z_n = B^T x_n$. Then reconstructed $\tilde{x_n} = Bz_n = BB^T x_n$. For $M = 1$, variance of projected data $= b_1^T S b_1$. Thus, b_1 should be the eigenvector of S corresponding to the largest eigenvalue. Extension to higher M
Minimum Error Perspective $\tilde{x_n} = Bz_n$. Minimize $J_M = \frac{1}{N}\sum ||x_n - \tilde{x_n}||^2$. Gradient w.r.t. z gives $z = B^T x$ so $\tilde{x_n} = Bz_n = BB^T x_n$. Then $J_M = \frac{1}{N}\sum || \sum_{m=M+1}^{D} b_m z_m n||^2 = \frac{1}{N}\sum_n \sum_{m=M+1}^{D}(b_m^T x_n)^2 = \sum_{m=N+1}^{D} b_m S b_m$.
PPCA $x = Bz + \mu + \epsilon \in \mathbb{R}^D$. $p(x|B, \mu, \sigma^2)$ has mean μ, variance $BB^T + \sigma^2 I$. For T containing eigenvectors of data covariance matrix, Λ with eigenvalues on diagonal, any orthogonal R, MLE is given as

$$\mu_{ML} = \frac{1}{N}\sum x_n, \quad B_{ML} = T(\Lambda - \sigma^2 I)^{1/2}R, \quad \sigma^2_{ML} = \frac{1}{D-M}\sum_{j=M+1}^{D}\lambda_i$$

3 Clustering

Goal of This Chapter. Let c_j denote the mean of the cluster C_j. Given data points, we want to cluster data points that minimize one of three following measures, especially focusing at the k-means.

k-center clustering $\Phi_{kcenter}(\mathcal{C}) = \max_{j=1}^{k} \max_{a_i \in C_j} d(a_i, c_j)$
k-median clustering $\Phi_{kmedian}(\mathcal{C}) = \sum_{j=1}^{k} \sum_{a_i \in C_j} d(a_i, c_j)$
k-means clustering $\Phi_{kmeans}(\mathcal{C}) = \sum_{j=1}^{k} \sum_{a_i \in C_j} d^2(a_i, c_j)$
Lemma n points a_i with centroid c satisfies $\sum_i ||a_i - x||^2 = \sum ||a_i - c||^2 + n||c - x||^2$ for any x. (proof) By definition, expand $||a_i - x||^2 = ||(a_i - c) + (c - x)||^2$
Lloyd's algorithm for Clustering Start with k center. Cluster each point with the center nearest to it. Find the centroid of each cluster and replace the set of old centers with the centroids. Repeat. *Might fall in local minimum.*
Spectral Clustering Cluster on projected points. $C \in \mathbb{R}^{n \times d}$ with row i the center of cluster data i belonging.

$$\Phi_{kmeans} = ||A - C||_F^2$$

Theorem A_k be teh projection of rows of A to the first k right singular vectors of A. Then for any C of rank k, $||A_k - C||_F^2 \leq 8k||A - C||_2^2$. (proof) (i) $rank(A_k - C) \leq 2k$ so that $||A_k - C||_F^2 \leq 2k||A_k - C||_2^2$ (ii) $||A_k - C||_2 \leq ||A_k - A||_2 + ||A - C||_2 \leq 2||A - C||_2$. Combine two inequalities.

4 Density Estimation with GMM

4.1 MLE for GMM

Goal of This Chapter. Model $p(x|\theta) = \sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)$ with $\sum_{k=1}^{K} \pi_k = 1$, the sum of Gaussian. Our objective is given samples from distribution and fixed K, finding MLE $\theta = (\mu, \Sigma_k, \pi)$
Responsibility $r_{nk} = \frac{\pi_k \mathcal{N}(x_n|\mu_k, \Sigma_k)}{\sum_{j=1}^{K} \pi_j \mathcal{N}(x_n|\mu_j, \Sigma_j)}$ measures amount of k-th Gaussian contributes to x_n.
Likelihood $p(\mathcal{X}|\theta) = \prod_{i=1}^{N} (\sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k))$.
Log-likelihood $\mathcal{L} = \log p(\mathcal{X}|\theta) \sum_{i=1}^{N} \log((\sum_{k=1}^{K} \pi_k \mathcal{N}(x|\mu_k, \Sigma_k)))$
Optimality Condition $\frac{\partial \mathcal{L}}{\partial \mu_k} = 0$, $\frac{\partial \mathcal{L}}{\partial \Sigma_k} = 0$, $\frac{\partial \mathcal{L}}{\partial \pi_k} = 0$. Given θ, we use this to find MLE θ^{new}.

4.2 Theorems for EM Updates in GMM

Theorem $\mu_k^{new} = \frac{\sum_{n=1}^{N} r_{nk} x_n}{\sum_{n=1}^{N} r_n k} = \mathbb{E}_{r_{nk}}[x_n]$. (proof) By $\frac{\partial \mathcal{L}}{\partial \mu_k} = 0$.
Theorem $\Sigma_k^{new} = \frac{1}{\sum_{n=1}^{N} r_{nk}} \sum_{n=1}^{N} r_{nk} (x_n - \mu_k)(x_n - \mu_k)^T$. (proof)
Theorem $\pi_k^{new} = \frac{1}{N} \sum_{n=1}^{N} r_{nk}$. (proof)

4.3 Latent Variable Perspective

5 Sampling by MCMC

Goal of this Chapter. Calculating integration(or expectation) using random walks. P with P_{xy} indicating probability of moving from state x to state y. $\sum_y P_{xy} = 1$ for each x. i.e. $p(t+1) = p(t)P$.
stationary Vector of P is prob. vector satisfying $\pi P = \pi$.
Long term Average of random walk $p(t)$ is $a(t) = \frac{1}{t}(p(0) + \cdots + p(t-1))$ **Theorem** Long term average converges to the unique stationary vector of random walk, if the MC is "strongly connected"

5.1 MCMC

γ= average value of f at the states seen in a t step walk. $\mathbb{E}[\gamma] = \sum_i f_i(\frac{1}{t} \sum_{j=1}^{t}$ prob(walk is in state i at time $j = \sum_i f_i a_i(t)$. By theorem, this converges to $\sum f_i \pi_i$, for stationary point π of the walk P. Thus, it remains to construct P that stationary point of P is p.

5.2 MCMC Algorithms

Lemma If $\pi_x p_{xy} = \pi_y p_{yx}$ for all x, y and $\sum_x \pi_x = 1$, then π is stationary distribution of the walk. i.e. $\pi P = \pi$. (proof) $\pi_x = \sum_y \pi_x P_{xy} = \sum_y \pi_y P_{yx} = (\pi P)_x$.
Metropolis-hasting Algorithm r =maximum degree of vertex. $p_{ij} = \frac{1}{r} \min(1, \frac{p_j}{p_i})$, $\quad p_{ii} = 1 - \sum_{j \neq i} p_{ij}$. By lemma, stationary vector of P is p.
Gibbs Sampling For d dimensional variable, make edges between variables that only changes one coordinate. If x, y differs only in the first coordinate, let $p_{xy} = \frac{1}{d} p(y_1|x_2, \cdots, x_d) = \frac{p(y)}{d \times p(x_2, \cdots, x_d)}$. $p(y) = \frac{p(x)}{d \times p(x_2, \cdots, x_d}$ so that $p(x)p_{xy} = p(y)p_{yx}$