Lecture 11

Huang Ziheng

November 28, 2017

1 Deep Learning

① Neural Networks(CNN,RNN)
② Optimization
③ Generalization (Test error $\approx$ Training error)
For example, why CNN + SGD $\mapsto$ Image has a small training error.
Loss surface, $l(w, S) \sim$ empirical loss and $l(w, P_D) \sim$ true loss
If random draw w then by Chernoff Bound they are similar. But the dimension of parameters is far more then the sample size so there exist many w that will make them far different.
No-bad local minimize: Suppose all the local minimization are global minimization.

2 Online Learning

Online Learning with Expert Advice Alg: Weighted Majority Vote:
For t=1,2...T, every expert i$\in [N]$ makes a prediction $\tilde{y}_{t,i}$ Adversarial gives the ground truth y_t where $\tilde{y}_{t,i}, y_t \in 0,1$ Loss $|\tilde{y}_{t,i} - y_t|$ for each expert i at time t.
Now at time t, alg makes a prediction $\tilde{y}_t$, Loss for the alg:$\sum_{t=1}^{T} |\tilde{y}_t - y_t| \approx \min_{i\in[N]} \sum_{t=1}^{T} |\tilde{y}_{t,i} - y_t|$
Alg(WM) Parameter $\beta \in (0, 1)$
Initialize $w_{1,i} = 1, i \in [N]$
For t=1,2...T
① Majority Vote $\tilde{y}_t = 1$ if $\sum_{y_{i,t}=0} w_{i,t} < \sum_{y_{i,t}=1} w_{i,t}$
② If $\tilde{y}_t = y_t$ then $w_{t+1,i} \leftarrow w_{t,i}$ Else $w_{t+1,i} \leftarrow \beta w_{t,i}$ for all i such that $\tilde{y}_{t,i} \neq y_t$
Thm: Let $L_T = \sum_{t=1}^{T} |\tilde{(y)}_t - y_t|$, $m_T^{(i)} = \sum_{t=1}^{T} |\tilde{y}_{t,i} - y_t|$, $m_T^* = \min_{i\in[N]} m_T^{(i)}$
Then $L_T \leq \frac{\ln \frac{1}{\beta}}{\ln \frac{2}{1+\beta}} m_T^* + \frac{\ln N}{\ln \frac{2}{1+\beta}}$
Proof: Potential Function Method
$W_t := \sum_{i=1}^{N} w_{t,i}$ Now when $\tilde{y}_t \neq y_t$ (Alg makes wrong prediction) then $W_{t+1} \leq (\frac{1+\beta}{2}) W_t$ then $W_T \leq \frac{1+\beta}{2})^{L_T} N$ And $W_T \geq w_{T,i} = \beta^{m_T^{(i)}}$ $\forall i$

Randomized Weighted Majority Vote Alg Parameter $beta \in (\frac{1}{2}, 1)$
Initialize $w_{1,i} = 1$ For t=1,2...T
① Randomized majority vote according to $\frac{w_{t,i}}{\sum_i w_{t,i}}$
② The same, $w_{t+1,i} \leftarrow \beta w_{t,i}\ \forall i\ s.t.\ \tilde{y}_{t,i} \neq y_t$
Define expected loss $L_T = \sum_{t=1}^{T} \sum_{i=1}^{N} \frac{w_{t,i}}{\sum_i w_{t,i'}} |\tilde{y}_{t,i} - y_t|$
Thm: For $\beta \in (\frac{1}{2}, 1)$ we have: $L_T \leq (2-\beta)m_T^* + \frac{\ln N}{1-\beta}$
Assume T is known, let $\beta = 1 - \sqrt{\frac{\ln N}{T}}$ then we have $L_T \leq m_T^* + 2\sqrt{T * \ln N}$
and then $\frac{L_T}{T} \leq \frac{m_T^*}{T} + O(\sqrt{\frac{\ln N}{T}})$ (Homework1: This thm)
Now for online learning T is usually unknown so we use Doubling Trick to solve it: we can guess a T first, then if we want to continue then we double T. It is easy to prove that with this trick we can get a similar result.

3 Von-Neurnann Minmax Thm

$\min_p \max_q p^T M q = \max_q \min_p p^T M q$
Proof of Von-Neurnann Minmax Thm:
Repeated Game, zero-sum matrix game.
Each row is an expert, row player combines experts and chooses p_t. Column player is the adversarial, chooses q_t. Now at time t, expert i suffers loss $(Mq_t)_i$ and row player loss $p_t^T M q_t$.
Alg: Initialize $p_1 = (\frac{1}{N}, ..., \frac{1}{N})$, $\beta \in (\frac{1}{2}, 1)$
For t=1,2,...,T we want to make $q_t = max_q p_t^T M q$
① Row player chooses p_t
② Column player chooses q_t (q_t can depend on p_t)
③ Row player observes the loss of each row (Mq_t)
④ $p_{t+1}(i) = p_t(i)\beta_i^{Mq_t}/z_t$ where z_t is a normalization factor $s.t.$ $\sum p_{t+1}(i) = 1$
Assume $M_{ij} \in [0, 1]$

$$\sum_{t=1}^{T} p_t^T M q_t \leq (2-\beta)\min_i (\sum_{t=1}^{T} Mq_t)_i + \frac{\ln N}{1-\beta}$$
$$\frac{1}{T}\sum_{t=1}^{T} p_t^T M q_t \leq \frac{1}{T}\min_i (\sum_{t=1}^{T} Mq_t)_i + O(\sqrt{\frac{\ln N}{T}})$$

Where the left side is $\min_p \max_q p^T M q$ while the one of the right side is another so...

4 Unsupervised Learning

Clustering. K-cluster.