Lecture 8

Lai Zehua 2014012668

2017 11 21

Top Conference: NIPS Late May COLT ICML Feb ICLR Oct

Algorithmic Stability and Generalization

SVM and Boosting try to improve the margin. Now we try to analyse the property of a algorithm and its error. For Boosting: $l = \frac{1}{n}\sum exp(-y_i \sum \alpha_t h_t(x))$. For SVM: $l = \frac{1}{n}\sum[1 - y_i(w^T x_i + b)] + \frac{\lambda}{2}||w||^2$.

Let $\mathcal{A}$ be a learning algorithm. $S = \{(x_i, y_i)\}$ be the training data set. Let $l(\mathcal{A}(S), z)$ be the loss function, $\mathcal{A}(S)$ be the result of the learning algorithm, z be the test data. Risk function is $R(\mathcal{A}(S)) = E_z[l(\mathcal{A}(S), z)]$. Empirical risk is $R_{emp}(\mathcal{A}(S)) = \frac{1}{n}\sum l(\mathcal{A}(S), z_i)$

Definition. A learning algorithm $\mathcal{A}$ is said to have **uniform stability $\boldsymbol{\beta}$** with respect to loss l, if for $\forall S = (z_1, ..., z_n), S^i = (z_{-i}, z_i')$, $|l(\mathcal{A}(S), z) - l(\mathcal{A}(S^i), z)| \leq \beta$

Theorem. Suppose $\mathcal{A}$ has uniform stability β with respect to loss l and $l \leq M$, then with probability $1 - \delta$,

$$R(\mathcal{A}(S)) \leq R_{emp}(\mathcal{A}(S)) + \beta + (n\beta + M)\sqrt{\frac{2log\frac{1}{\delta}}{n}}$$

Proof. The theorem is equivalent to $\mathbb{P}[R(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S)) \geq \beta + \epsilon] \leq exp(-\frac{n\epsilon^2}{2(n\beta+M)^2})$ (Chernoff bound).

Let $f(S) = R(\mathcal{A}(S)) - R_{emp}(\mathcal{A}(S))$, then $E_S[f(S)] = E_{S,z_i'}[l(\mathcal{A}(S), z_i') - l(\mathcal{A}(S^i), z_i')] \leq \beta$ and $|f(S) - f(S^i)| \leq 2(\beta + \frac{M}{n})$.

Combine the two inequality and Mcdiarmid lemma, the result follows. (Details as homework) □

In (kernel) SVM, loss function is $l = \frac{1}{n}\sum[1 - y_i(w^T x_i + b)] + \frac{\lambda}{2}||w||^2$. The learning algorithm is to minimize the loss function.

Suppose $||x|| \leq 1$, for example, if the kernel is Gaussian kernel, $||x|| = 1$. Then SVM has uniform stability $\beta(n) = O(\frac{1}{\lambda n})$. *Stability and Generalization, Olivier Bousquet, André Elisseeff.*

Deep Learning

1. Architecture.

2. Learning Algorithm, SGD.

2.1 Optimization. (non-convex optimization)

2.2 Generalization. *Understanding deep learning requires rethinking generalization. Stanford, Deep learning theory. Generalization Bounds of SGLD for non-convex learning: two theoretical viewpoints.*