Lecture 6

Huang Ziheng

November 18, 2017

1 Bayesion and Frequentist

Bayesion: We learn a distribution of classifier so it is a stochastic classifier. Then the function of it can be valued by its expectation.
Voting classifier vs stochastic classifier: (skip)
As VC theory is some kind of Frequentist theory, so now by the view of Bayesion we want a uniform convergence, but it is a stochastic classifier, we should uniform all the distribution of the stochastic classifiers.
For fixed prior distribution $\mathcal{P}$ (w.h.p over random draw of training data) for all distribution $\mathcal{Q}$
Recall that for VC-theory or Margin-theory we want to get:

$$\text{For all classifier } f \in \mathcal{H} \text{ , } err_D(f) \leq err_S(f) + Complexity$$

So for this case it should be:

$$\text{For all distribution } Q \text{ , } err_D(Q) \leq err_S(Q) + D(Q||P)$$

2 PAC Bayes Thm:

For any fixed prior distribution $\mathcal{P}$, w.p $1-\delta$

$$err_D(Q) \leq err_S(Q) + \sqrt{\frac{D(Q||P)+log(\frac{3}{\delta})}{n}}$$

Holds for all α simultaneously.

Lemma 1 For any functional f of classifier $\hbar$

$$E_{\hbar\sim Q}[f(\hbar)] \leq lnE_{\hbar\sim P}[e^{f(\hbar)}] + D(Q||P)$$

Homework1

Lemma 2 Let $f(\hbar) = n * [err_D(\hbar) - err_S(\hbar)]^2$
Then $Pr[E_{\hbar\sim P}exp(f(\hbar)) \geq \frac{3}{\delta}] \leq \delta$
Proof of it:
$\forall fixed\ \hbar$, $Pr(|err_D(\hbar) - err_S(\hbar)| \geq \epsilon) \leq 2e^{-2n\epsilon^2}$
Homework2

Lemma 3 Improved PAC Bayes Thm:

$$Pr(D_B(err_S(Q)||err_D(Q)) \geq \delta) \geq \frac{D(Q||P)+log\frac{n+1}{\delta}}{n}$$

3 Term Project 3

By hardware human brain is far stronger than any computer but now or in the near future the size will be comparable, at least the speed of computer is faster. At the same time the performance of our computer is far below the human brain performance. So may be human brain is not work by concurrent algorithm but by a global control system and worked like distributed computation, so we should explore it.
① Decoupled Neural Interfaces using Synthetic Gradients
② Understanding Synthetic Gradient and Decoupled Neural Interfaces.

4 PAC-Bayes implies Margin theory for SVM

We have a distribution Q and it derives a voting classifier and a stochastic classifier, but the error of the first one cannot exceed the 2-times of the second one. Because if the voting classifier is wrong then the stochastic is only half right.
Assume the linear classifier goes from the origin (because we can add 1 more dimension to make it), then the unit normal vector of the classifier is uniformly distributed on the surface of unit ball centering at origin, this is our prior distribution.
But it is hard to calculate so we use $P \sim \mathcal{N}(0, I)$ instead.???
And gaussian distribution for posteriors distribution is easy to calculate the KL-distance, thus we suppose $Q \sim \mathcal{N}(\mu, I)$
Actually $Q \sim \mathcal{N}(u * \overrightarrow{w}, I)$, where $u \in \mathcal{R}$
PAC-Bayes

$$err_D(Q) \leq err_S(Q) + \sqrt{\frac{D(Q||P)+log\frac{3}{\delta}}{n}}$$

1) $err_S(D)$
2) $D(Q||P) = D(\mathcal{N}(u * \overrightarrow{W}, I)||\mathcal{N}(0, I)) = \frac{u^2}{2}$
Proof: Only 1-dimension is needed to integral and all others equal
Now consider $err_S(Q)$ where $Q \sim \mathcal{N}(u * \overrightarrow{W}, I)$
For any given (x, y) the probability for wrong classified is:

$$\frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-t^2} dt \text{ where } t = u * y * \frac{w*x}{||x||}$$

So we have: $err_S(Q) = \frac{1}{n} \sum \Phi(u * y_i * \frac{w*x_i}{||x_i||})$ where $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{t^2}{2}} dt$
Then when $u \to 0$ obviously the margin is small.