MID-SEMESTER EXAMINATION (CS-729)

13-Sep-2014 (Saturday)

Important Instructions:

- Duration: 2hrs; Max. Marks: 20.
- Though this question paper is written in English (mainly), your answers should use the **language of Mathematics** (mainly).
- Marks will be awarded for **concise, correct and relevant mathematical arguments**. Marks may not be awarded for arguments based on intuition etc.
- You may employ McDiarmid's inequality or Contraction lemma **without** repeating their proofs. But otherwise, proofs of theorems/results **must be** repeated in your answers even if they were done in lectures or in your textbooks.
- Your handwriting must be **legible**. I will give marks for what I see and understand; and not necessarily for what you thought or what you wrote.
- I have employed the same **notation as in our lectures** and you should also do the same.

Here is a conversation between Chiru and Saketh:

Chiru: ..... Hence, for the loss class induced by the hinge-loss and the function class $\mathcal{F}_W \equiv \{f \mid \exists\ w \in \mathbb{R}^n, \|w\| \leq W \ \ni \ f(x) = w^\top x\ \forall\ x \in \mathcal{X}\}$, the following is true with atleast probability $1-\delta$:

$$R[f] \leq \hat{R}_m[f] + 2\frac{WR}{\sqrt{m}} + \Phi(\delta, m)\ \forall\ f \in \mathcal{F}_W,$$

where $R, \hat{R}_m$ denote the true, empirical risks (computed with m examples), $\max_{x\in\mathcal{X}} \|x\| = R$ and $\Phi(\delta, m) = 3\sqrt{\frac{\log\frac{2}{\delta}}{2m}}$.

Saketh: Sir, this result is interesting. Now, we not only have a proof for statistical consistency for this case but also a new algorithm for model selection!

Chiru (smiling :) OK! what is it?

Saketh (impetuous :) Solving the following will give both, the "best" $\hat{W}_m$ and the "best" $\hat{w}_m$ in it:

$$(\hat{W}_m, \hat{w}_m) \equiv \text{argmin}_{W>0, \|w\|\leq W} \hat{R}_m[f] + 2\frac{WR}{\sqrt{m}}.$$

Basically I am trying to minimize the upper bound (guaranteed risk) wrt. to both the parameters w and hyper-parameter W.

Chiru: This is not guaranteed to work.

Saketh (showing off :) Why? This is a nice well posed optimization problem with non-trivial solutions; infact, it is a convex conic-quadratic program that has efficient solvers. Atleast this is not non-convex as in [1].

Chiru: Is computational efficiency the only concern? My apprehension is for a more basic reason.

Saketh: Consistency? Is'nt it obvious...I mean what you proved just now proves consistency.

Chiru: Is it?

Saketh: Ofcourse.. .. well.. .. seems like ... . does'nt your result prove that ERM is consistent in $\mathcal{F}_{\hat{W}_m}$?

Chiru: I thought you said you had a new algorithm for model selection (which is not ERM).

Saketh: Ok! I got it. $\hat{W}_m$ is itself random, so the guarantees may not remain the same as with a fixed W. The situation is analogous to where we started: with fixed w, the law of large numbers is enough; but with w chosen by an algorithm (ERM) that depends on a random variable (training set), the guarantee weakens to the above. In summary, we need to show that the new algorithm itself is (statistically) consistent. i.e., we need to show that

$$\{R[\hat{w}_m]\} \to R[w^*], \text{in prob. as } m \to \infty,$$

where w^* is the true risk minimizer in $\mathcal{F} \equiv \cup_{W>0}\mathcal{F}_W = \{f \mid \exists\ w \in \mathbb{R}^n \ \ni \ f(x) = w^\top x\ \forall\ x \in \mathcal{X}\}$.

Chiru: Right! But, intuitively, do you think you will be able to show this?

Saketh: One thing is clear: In the related case of ERM on $\mathcal{F}$, the Rademacher complexity $\mathcal{R}(\mathcal{F})$ seems to be ∞, as I can always choose a w such that all points lie on the negative side of the corresponding hyperplane and $\|w\| \to \infty$. Basically, I can get arbitrarily high losses on any labeling of my training set! So, things do'nt work if we stick with uniform convergence criteria alone.

Chiru: Bingo! Do you think uniform convergence criteria is even sufficient for this case?

Saketh: May not be. We only showed that for ERM it is. I will need careful analysis. Let me start with some easy cases:

1. Set of $W > 0$ to choose from is finite: say $W_1, \ldots, W_k$, where each $W_i > 0$. i.e., solve the following:

$$(\hat{W}_m, \hat{w}_m) \equiv \text{argmin}_{W \in \{W_1,\ldots,W_k\}, \|w\| \leq W} \hat{R}_m[f] + 2\frac{WR}{\sqrt{m}}.$$

2. Set of $W > 0$ to choose from is a finite interval: say $(0, B]$. i.e., solve the following:

$$(\hat{W}_m, \hat{w}_m) \equiv \text{argmin}_{W \in (0,B], \|w\| \leq W} \hat{R}_m[f] + 2\frac{WR}{\sqrt{m}}.$$

Chiru: Kudos! Can you prove consistency in the above two cases? More importantly, can you derive learning bounds[1] for these two cases.

Saketh: My students are smarter than me and they will.

[5+15Marks=20Marks]

Chiru: Good. Now that your students proved consistency and derived learning bounds, do you think this apparatus is enough to show Bayes consistency.

Saketh: Well.... No.... Because, norm of the Bayes optimal's parameter, could be greater than B. It seems like Bayes consistency cannot be shown unless W reaches infinity. But infinite interval seems difficult to analyze. Perhaps statistical consistency is easier to show with the following infinite increasing sequence $\{W_1, W_2, \ldots, W_k, \ldots\}$. I will try proving it in the lecture on Wed, 17-Sep-2014 with the help of my students :)

Chiru: I am sure you will.

References

[1] O. Chapelle, V. N. Vapnik, O. Bousquet, and S. Mukherjee. Choosing Multiple Parameters for Support Vector Machines. 46(1–3):131–159, 2002.

[1] Note that learning bounds have true risk on the LHS and quantities computable from training set or constants on the RHS. Do not give bounds that contain unknowns on the RHS.

CS729 Mid Sem Exam Solution

16-Sep-2014

TST $\{R[\hat{\omega}_m]\} \xrightarrow{P} R[\omega^*]$ as $m \to \infty$.

Proof: Let Ω be the set of values W may take. For (i), it is $\Omega = \{W_1, \ldots W_k\}$. (ii) it is $\Omega = [0, B]$.

We have,

$$0 \leq R[\hat{\omega}_m] - R[\omega^*] \quad \text{(defn. of } \omega^*\text{)}$$

$$= \left(R[\hat{\omega}_m] - \hat{R}_m[\hat{\omega}_m]\right) + \left(\hat{R}_m[\hat{\omega}_m] - \hat{R}_m[\omega^*]\right) + \left(\hat{R}_m[\omega^*] - R[\omega^*]\right)$$

(for ERM this was ≤ 0, but here it need not be!)

$$= \left(R[\hat{\omega}_m] - \hat{R}_m[\hat{\omega}_m]\right) + \left(\hat{R}_m[\hat{\omega}_m] + \frac{2\hat{W}_m R}{\sqrt{m}} - \hat{R}_m[\omega^*] - \frac{2\|\omega^*\| R}{\sqrt{m}}\right) + \left(\hat{R}_m[\omega^*] - R[\omega^*]\right)$$

$$+ \left(\frac{2\|\omega^*\| R}{\sqrt{m}} - \frac{2\hat{W}_m R}{\sqrt{m}}\right)$$

($\hookrightarrow \leq$ obj. defn. of $\hat{\omega}_m, \hat{W}_m$.)

$$\leq \max_{\substack{\|\omega\| \leq W \\ W \in \Omega}} \left(R[\omega] - \hat{R}_m[\omega]\right) + \left(\hat{R}_m[\omega^*] - R[\omega^*]\right) + \frac{2\|\omega^*\| R - 2\hat{W}_m R}{\sqrt{m}}$$

- $\max \left(R[\omega] - \hat{R}_m[\omega]\right)$: we know how to analyze this term. Simply repeat the derivation of upper bounding this using Rademacher average with $\|\omega\| \leq W_k$ for (i) & $\|\omega\| \leq B$ for (ii)
- $\left(\hat{R}_m[\omega^*] - R[\omega^*]\right)$: special case of previous term
- $\frac{2\|\omega^*\| R - 2\hat{W}_m R}{\sqrt{m}} \leq$ (ii) $\frac{2BR}{\sqrt{m}}$; (i) $\frac{2W_k R - 2WR}{\sqrt{m}}$

3/δ probability; 3/δ because of union bound.

we have with prob. at least $1-\delta$: Let $\bar{W} = \begin{cases} W_k & \text{for (i)} \\ B & \text{for (ii)} \end{cases}$

In summary,

$$R[\hat{\omega}_m] - R[\omega^*] \leq \frac{2\bar{W}R}{\sqrt{m}} + 4\sqrt{\frac{\log 3/\delta}{2m}} + \frac{2\bar{W}R}{\sqrt{m}}$$

$$= \frac{4\bar{W}R}{\sqrt{m}} + 4\sqrt{\frac{\log 3/\delta}{2m}}$$

Hence proved. (I)

Learning bound remains the same as on Ω!, i.e.

$$R[f] \leq \hat{R}_m[f] + \frac{2\bar{W}R}{\sqrt{m}} + 3\sqrt{\frac{\log 2/\delta}{2m}} \quad \forall f \in \mathcal{F}. \qquad \text{(II)}$$