ELL880: Computational Learning Theory and the Mind

Major Test, Maximum marks: 30

May 4, 2019

1. Consider binary classification problems on the learning domain $\mathbb{X} \times \{\pm 1\}$. Prove that, for all *finite* $\mathbb{X}$, there exists a single-hidden-layer neural network architecture $(V, E, sign)$ such that $\mathcal{H}_{V,E,sign}$ shatters the whole of $\mathbb{X}$. What does the minimal width of the hidden layer need to be to allow for this? Show all steps and working in your proof clearly. (Hint: Assume that you can choose any suitable input representation for $\mathbb{X}$.) **[5]**

2. Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two hypothesis classes with VC-dimension d_1 and d_2 respectively. Recall the growth function bound given by Sauer's Lemma: For any hypothesis class $\mathcal{H}$ with VC-dimension d, if $N > d+1$ then $\tau_{\mathcal{H}}(N) \leq (eN/d)^d$. Here
$$\tau_{\mathcal{H}}(N) = \max_{C \subset \mathbb{X}:|C|=N} |\mathcal{H}_C|.$$
Suppose $\mathcal{H}_1$ and $\mathcal{H}_2$ are *composable*; and consider the hypothesis class $\mathcal{H} = \mathcal{H}_2 \circ \mathcal{H}_1$, which consists of the compositions of hypotheses from $\mathcal{H}_1$ and $\mathcal{H}_2$. More precisely, if the input space corresponding to $\mathcal{H}_1$ is $\mathbb{X}$, this can be stated as $\mathcal{H} = \{h : \exists h_1 \in \mathcal{H}_1, h_2 \in \mathcal{H}_2 \text{ s.t. } \forall \mathbf{x} \in \mathbb{X}, h(\mathbf{x}) = h_2(h_1(\mathbf{x}))\}$.

   Obtain an upper bound, in terms of d_1 and d_2, on $\tau_{\mathcal{H}}(N)$ for $N > \max(d_1, d_2) + 1$, using Sauer's Lemma. Clearly show all steps and working. **[5]**

3. Consider the following bound on the expected true loss of a soft SVM we saw in class:
$$\mathop{\mathbb{E}}_{S \sim \mathcal{D}^N} \left[L_{\mathcal{D}}^{0-1}(sSVM(S))\right] \leq \min_{\mathbf{w}:||\mathbf{w}|| \leq B} L_{\mathcal{D}}^{hinge}(\mathbf{w}) + \sqrt{\frac{8\rho^2 B^2}{N}}.$$
(a) What is the key quantity which is *missing* from this bound? Explain the importance of it not being part of the bound. **[1.5]**

   (b) In training a soft SVM in practice, how does the value of B get optimised to keep this bound as low as possible? Explain the relationship between the actual training/tuning process for an SVM, and obtaining a good value for this generalisation bound. **[2]**

   (c) Does the choice of B here (effected via the process described in the previous part) correspond to some kind of prior knowledge? Explain what kind. Does the prior become stronger or weaker as B increases? **[1.5]**

4. The Minimum Description Length (MDL) paradigm seeks to pick a hypothesis h from a countable class $\mathcal{H}$ which minimises the following (probabilistic) upper bound on the true loss (for binary classification with zero-one loss):
$$L_{\mathcal{D}}(h) \leq L_S(h) + \sqrt{\frac{|h| + \ln(2/\delta)}{2N}},$$
where $|h|$ is the length of $d(h)$ under a description language $d : \mathcal{H} \to \{0, 1\}^*$.

   (a) Prove that the above bound indeed holds with probability greater than $1 - \delta$ over the choice of training set $S \sim \mathcal{D}^N$, for any prefix-free description language, and for all N, $\delta > 0$, $\mathcal{D}$, and $h \in \mathcal{H}$.

Hint: Make use of the Kraft Inequality (if $S \subseteq \{0,1\}^*$ is prefix-free, $\sum_{\sigma \in S} (1/2^{|\sigma|}) \leq 1$), Theorem 7.4 (provided in the Appendix to this paper), and of the Hoeffding inequality bound on UC sample complexity for finite classes:

$$N_{\mathcal{H}}^{UC}(\epsilon, \delta) \leq \frac{log(2|\mathcal{H}|/\delta)b^2}{2\epsilon^2},$$

where b is a bound on the value of the loss function. **[7]**

(b) In interpreting MDL as a formalisation of Occam's Razor, the implicit assumption is that 'simpler' hypotheses (as per some pre-existing notion of simplicity) have shorter description length. However, the above bound holds for *any* prefix-free description language! So there is nothing to stop us from picking a description language which gives longer descriptions to 'simpler' hypotheses, and then the result of applying MDL would seemingly be the inverse of Occam's Razor: we would be favouring less simple hypotheses *a priori*. So the MDL paradigm alone cannot give us Occam's Razor. What more is needed to justify Occam's Razor as an inductive principle? Can you connect this to Hume's Problem of Induction? **[3]**

5. Let $\mathcal{H}$ be the class of signed intervals over $\mathbb{R}$, *i.e.*, $\mathcal{H} = \{h_{a,b,s} : a \leq b, s \in \{-1, 1\}\}$ where $\forall x \in \mathbb{R}$:

$$h_{a,b,s}(x) = \begin{cases} s & \text{if } x \in [a,b] \\ -s & \text{if } x \notin [a,b] \end{cases}$$

Compute the VC-dimension of $\mathcal{H}$, clearly showing all steps and reasoning. **[5]**

Appendix

Theorem 7.4 (SRM bound)

Let $w : \mathbb{N} \to (0,1)$ be such that $\sum_{n=1}^{\infty} w(n) \leq 1$. Let $\mathcal{H} = \bigcup_{n \in \mathbb{N}} \mathcal{H}_n$, where each $\mathcal{H}_n$ enjoys uniform convergence with sample complexity $N_{\mathcal{H}_n}^{UC}(\epsilon, \delta)$. Then, $\forall \delta \in (0,1)$, $\forall \mathcal{D}$, $\forall n \in \mathbb{N}$, $\forall h \in \mathcal{H}_n$:

$$|L_{\mathcal{D}}(h) - L_S(h)| \leq \epsilon_n(N, w(n)\delta)$$

with probability at least $1 - \delta$ over the choice of $S \sim \mathcal{D}^N$.

Here

$$\epsilon_n(N, \delta) = \min\{\epsilon \in (0,1) : N_{\mathcal{H}_n}^{UC}(\epsilon, \delta) \leq N\}.$$