

## 58147 Machine Learning (Spring 2005)

### Exercise 2 (Wednesday 9 February)

1. Let  $X = \{1, 2, \dots, k\}$ , and consider the concept class  $C$  that consists of closed intervals  $[a, b]$  where  $a, b \in X$  and  $a \leq b$ . What is the cardinality  $|C|$  of the concept class?

Show how the Halving Algorithm for this  $C$  can be implemented efficiently. (The implementation should do less than  $O(|C|)$  computation per prediction.)

2. We consider two important inequalities that can both be proved by applying Jensen's inequality with  $f(x) = -\ln x$ .

- (a) Show that the arithmetic mean of non-negative numbers  $a_1, \dots, a_n$  is at least their geometric mean, i.e.,

$$\frac{1}{n} \sum_{i=1}^n a_i \geq \left( \prod_{i=1}^n a_i \right)^{1/n}.$$

- (b) Given  $\mathbf{p} \in \mathbf{R}^n$  and  $\mathbf{q} \in \mathbf{R}^n$  such that  $p_i \geq 0$  and  $q_i \geq 0$  for all  $i$  and  $\sum_{i=1}^n p_i = \sum_{i=1}^n q_i = 1$ , we define their *relative entropy* (or *Kullback-Leibler divergence*) as

$$d_{\text{KL}}(\mathbf{p}, \mathbf{q}) = \sum_{i=1}^n p_i \ln \frac{p_i}{q_i}.$$

Show that

$$d_{\text{KL}}(\mathbf{p}, \mathbf{q}) \geq 0$$

holds for all  $\mathbf{p}$  and  $\mathbf{q}$  (that satisfy the conditions above). This is known as the *information inequality*.

3. We consider the Weighted Average algorithm, with  $\eta$  and  $c$  that satisfy the conditions of Theorem 2.9 in the lecture notes. In this problem you are asked to generalise the proof of Theorem 2.9.

- (a) Show that if there are  $k$  different experts  $i$  that all satisfy  $L(\mathcal{E}_i) \leq M$  for some  $M$ , then

$$L(\text{WA}) \leq c\eta M + c \ln \frac{N}{k}.$$

- (b) Change the initialisation of the algorithm so that  $w_{1,i} = p_i$  for all  $i$ , where  $p_i > 0$  for all  $i$  and  $\sum_{i=1}^N p_i = 1$  but otherwise  $\mathbf{p}$  is arbitrary. Show that for the modified algorithm WA' we have

$$L(\text{WA}') \leq \min_{1 \leq i \leq N} \left( c\eta L(\mathcal{E}_i) + c \ln \frac{1}{p_i} \right).$$

(This generalises directly to a countably infinite set of experts, if we ignore the computational issue of actually running such an algorithm.)

4. Show that when  $y_t \in \{-1, 1\}$ , the WA algorithm for log loss with  $\eta = 1$  satisfies

$$L_{\log}(y_t, \mathbf{v}_t \cdot \mathbf{x}_t) = \ln \frac{W_t}{W_{t+1}}.$$

In other words, for this special case the condition of Theorem 2.9 holds as equality for  $c = 1$ . (Notice that a factor  $1/2$  was missing from the definition of log loss in the lecture notes; this has now been fixed.)

5. Calculate the values  $\tilde{c}_L$  for logarithmic and Hellinger loss. (Consider just the case  $y \in \{-1, 1\}$ .)