

## 58147 Machine Learning (Spring 2005)

### Exercise 7 (Wednesday 16 March)

**Revised version**, 15 March 2005: there was an assumption missing from 3 (b).

1. Let  $\mathcal{F}$  be a class of functions from  $Z$  to  $\mathbf{R}$ . Given a sequence of functions  $\mathbf{f} = (f_1, \dots, f_n) \in \mathcal{F}^n$  and a vector  $\mathbf{v} \in \mathbf{R}^n$ , define the function  $\mathbf{v} \cdot \mathbf{f}$  by  $(\mathbf{v} \cdot \mathbf{f})(z) = \sum_{i=1}^n v_i f_i(z)$ . We can then define the *convex hull* of  $\mathcal{F}$  as

$$\text{conv}(\mathcal{F}) = \left\{ \mathbf{v} \cdot \mathbf{f} \mid \mathbf{f} \in \mathcal{F}^n, \mathbf{v} \in \mathbf{R}^n, \sum_{i=1}^n v_i = 1, |v_i| \geq 0 \text{ for all } i \right\}$$

and the *symmetric convex hull* of  $\mathcal{F}$  as

$$\text{absconv}(\mathcal{F}) = \left\{ \mathbf{v} \cdot \mathbf{f} \mid \mathbf{f} \in \mathcal{F}^n, \mathbf{v} \in \mathbf{R}^n, \sum_{i=1}^n v_i = 1 \right\}.$$

Show that  $R_m(\mathcal{F}) = R_m(\text{conv}(\mathcal{F})) = R_m(\text{absconv}(\mathcal{F}))$ .

*Hint:* The “difficult” part is showing  $R_m(\text{absconv}(\mathcal{F})) \leq R_m(\mathcal{F})$ . Notice that  $f \in \text{absconv}(\mathcal{F})$  iff  $-f \in \text{absconv}(\mathcal{F})$ , so you can omit the absolute values when evaluating  $R_m(\text{absconv}(\mathcal{F}))$ .

2. In this and next problem, you may assume that all probability spaces are finite if you find that simpler. Our discussion of martingales is mainly based on R. Motwani and P. Raghavan: *Randomized Algorithms*, in case you wish a more coherent presentation.
  - (a) Consider throwing three symmetrical 6-sided dice. Let  $A$ ,  $B$  and  $C$  be the results of the individual dice and  $X = A + B + C$ . Show that

$$\mathbb{E}[\mathbb{E}[X \mid A, B] \mid A] = \mathbb{E}[X \mid A].$$

- (b) Generalise your argument from part (a) to show that the above equation holds for any random variables  $X$ ,  $A$  and  $B$ .

*Hint:* Define

$$Y(a, b) = \mathbb{E}[X \mid A = a, B = b] = \frac{\sum_x x \Pr(X = x, A = a, B = b)}{\sum_x \Pr(X = x, A = a, B = b)}.$$

Then write out  $\mathbb{E}[Y \mid A = a]$  in a similar form, and notice that it is equal to  $\mathbb{E}[X \mid A = a]$ .

3. Given a  $\sigma$ -algebra  $\mathcal{F}$ , define an equivalence relation  $\sim$  where  $a \sim b$  if  $a$  and  $b$  belong to exactly the same sets  $A \in \mathcal{F}$ . Formally,

$$a \sim b \quad \Leftrightarrow \quad \forall A \in \mathcal{F}: ((A \cap \{a, b\} = \{a, b\}) \vee (A \cap \{a, b\} = \emptyset)).$$

- (a) Define

$$\mathbb{E}[X \mid \mathcal{F}] = \mathbb{E}[X \mid Y]$$

where  $Y$  is any random variable such that  $Y(a) = Y(b)$  if  $a \sim b$  but  $Y(a) \neq Y(b)$  otherwise. Show that this is a valid definition, in the sense that  $\mathbb{E}[X \mid Y_1] = \mathbb{E}[X \mid Y_2]$  for any  $Y_1$  and  $Y_2$  that both satisfy the condition.

- (b) Assume that  $(\mathcal{F}_0, \dots, \mathcal{F}_n)$  is a filter and  $(X_0, X_1, \dots, X_n)$  a sequence of random variables such that  $X_i$  is  $\mathcal{F}_i$ -measurable. Prove that if

$$\mathbb{E}[X_{i+1} \mid \mathcal{F}_i] = X_i,$$

then

$$\mathbb{E}[X_{i+1} \mid Z_0, Z_1, \dots, Z_i] = X_i$$

for all  $(Z_i)$  such that  $Z_i$  is  $\mathcal{F}_i$ -measurable and  $X_i$  is a function of  $Z_0, \dots, Z_i$ .

(See next page.)

- 4.-5. Consider the car image data set from Exercise 5. This time, the task is to run the Marginalised Perceptron (Algorithm 2.41) with a few different values for the margin parameter  $\rho$ , compute for each resulting weight vector an error bound using Theorem 3.27, and choose the “best” one following the ideology presented on pages 184–185 for model selection.

Keep one third of the data as a test, and in the end check how well the theoretical bounds match actual performance (probably not very).

Some details:

- The largest margin value that even theoretically could make sense is  $\rho_0 = \max_t \|\mathbf{x}_t\|_2$ . To keep things simple, consider a set of 15 margin values  $\rho_i$  with  $\rho_i = 2^{-i}\rho_0$ .
- For large values of  $\rho$ , the algorithm will not converge, so you need to just stop when it looks like no significant progress is being made. Anyway, you need to use a sufficiently small learning rate, say  $\eta = 0.001$ .
- For applying the bound, pick  $\delta = 0.01$  and  $\mu = 2\rho$ . In the calculation you then need to replace  $\delta$  by  $\delta/15$  since you are making 15 simultaneous estimates.
- After picking the final hypothesis, compute the bound for it using several different values of  $\mu$ . Notice that picking the value that minimises the bound is *not* a valid procedure. (Neither is getting the value from your instructor, of course, but we want to keep things simple here.)