Paul Lintilhac

Foundations of Macine Learning

Homework 1 – For Professor Mohri

Section A: Probability tools

Problem 1:

First, note that in order for $f^{-1}(\delta)$ to be defined, δ must be in the range of f. Given this, for any δ>0 we can define set $t = f^{-1}(\delta)$, and note that t>0, since the domain of f is $(0, +\infty)$. Then we can plug this t into the given equation:

$$\Pr[x > f^{-1}(\delta)] \leq f\left(f^{-1}(\delta)\right) = \delta$$

Therefore,

$$\Pr[x < f^{-1}(\delta)] \leq 1 - \delta$$

Problem 2:

We start by writing an expression for the expectation of X:

$$E[X] = \sum_{n \geq 1} n * P[x = n]$$

Where we have started at n=1, since n=0 does not contribute to the sum. Using the identity

$$P[x = n] = P[x \geq n] - P[x \geq n + 1]$$

We have

$$E[X] = \sum_{n \geq 1} n * P[x \geq n] - \sum_{n \geq 1} n * P[x \geq n + 1]$$

Re-indexing the second integral, we get

$$E[X] = \sum_{n \geq 1} n * P[x \geq n] - \sum_{n \geq 2} (n - 1) * P[x \geq n]$$

$$= P[x \geq 1] + \sum_{n \geq 2} * P[x \geq n]$$

$$= \sum_{n \geq 1} * P[x \geq n],$$

As desired.

Section B: Label Bias

Problem 1:

We can consider p_+ to be the true error rate of the hypothesis that always guesses +1, with $\widehat{p_+}$ the empirical error rate on a sample of size m. We derive a result similar to 2.16 in the text, with the main difference being that the possible values for X are {-1,1} instead of {0,1}. (Because of this, my answer is actually different from that given. At the time of submission, I still could not figure out whether I am doing something wrong or if the homework ignored this distinction). We can start with Hoeffding's inequality, which states:

Let $\{X_1, \dots, X_m\}$ be independent random variables with X_i taking values in $[a_i, b_i]$ for all $i \in [1, m]$. Then for any $\epsilon > 0$, the following inequalities hold for $S_m = \sum_{i=1}^{m} X_i$:

$$\Pr[S_m - E[S_m] \geq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}}$$

$$\Pr[S_m - E[S_m] \leq \epsilon] \leq e^{\frac{-2\epsilon^2}{\sum_{i=1}^{m}(b_i - a_i)^2}}$$

Note that our range for X is [-1,1], so $(b_i - a_i)^2 = 4$ for all $i \in [1, m]$. In addition, we are interested in the empirical and true average error, rather than the sum of the (signed) errors. Dividing by m inside the probability has the desired effect, which is equivalent to replacing ϵ in the equation with $m\epsilon$. Thus the equations become

$$\Pr[p_+ - \widehat{p_+}, \geq \epsilon] \leq e^{\frac{-m\epsilon^2}{2}} < e^{-2m\epsilon^2}$$

$$\Pr[p_+ - \widehat{p_+}, \leq -\epsilon] \leq e^{\frac{-m\epsilon^2}{2}}$$

Using the union bound, we then have that

$$\Pr[|p_+ - \widehat{p_+}| \geq \epsilon] \leq 2e^{\frac{-m\epsilon^2}{2}}$$

Now, since we want to bound our error with confidence $1 - \delta$, we set the above quantity to δ and solve for ϵ:

$$\Pr\left[|p_+ - \widehat{p_+}| \leq \sqrt{\frac{2\log(\frac{2}{\delta})}{m}}\right] \leq 1 - \delta$$

Which proves the desired result.

Section C: Learning in the Presence of Noise

Problem 1:

I actually found the explanation in the textbook (as well as every other explanation I could find online) to be quite unclear, so I am going to briefly re-derive it here using a different approach that I find much clearer than any of the material I could find online (not that the equations in the material I found were necessarily wrong, but just that the logical flow unclear).

We are looking for the sufficient conditions under which the error of our axis-aligned rectangles learning will be less than some $\epsilon > 0$. This condition is as follows: that there exist four rectangles along the sides of R, {r_1, r_2, r_3, r_4} such that $P[r_1 \cup r_2 \cup r_3 \cup r_4] < \epsilon$, and all four sides of our model R_s are within these rectangles. Now since we are only looking for the sufficient condition, we do not need to worry about all possible sets of rectangles {r_1, r_2, r_3, r_4}, but instead can simply pick four whose areas are all equal to $\frac{\epsilon}{4}$. Note that this does not necessarily give the tightest possible bound, and it is very possible that by optimizing the areas of the four side rectangles, we could get a tighter bound.

The probability of a side being inside rectangle r_i is the probability that at least one positive point lie inside that rectangle, or equivalently that not all positive points lie outside that rectangle, i.e. $P[\neg$all outside $r_i]$, and then the probability that this is simultaneously true for all four rectangles is $P[\neg\ all\ outside\ r_1 \cap \neg\ all\ outside\ r_2 \cap \neg\ all\ outside\ r_3 \cap \neg\ all\ outside\ r_4]$. By DeMorgans law, this probability is equal to $P[all\ outside\ r_1 \cup\ all\ outside\ r_2 \cup\ all\ outside\ r_3 \cup all\ outside\ r_4]$. The probability that all m points lie outside rectangle r_i is $(1 - \frac{\epsilon}{4})^m$. Therefore, by the union bound, the sufficient condition for $R(R_s) < \epsilon$ occurs with probability $\leq 4(1 - \frac{\epsilon}{4})^m$.

a)

For this problem, we are considering the additional complication that the positive points are mislabeled with unknown probability η such that $0 < \eta < \eta'$ for some known η'. Let $M \subseteq S$ denote the set of mislabeled points, and let $P \subseteq S$ denote the set of positive observations (though they may also be members of M, i.e. mislabeled). Thus the set of positively labeled points is given by $P - M = P \cap M^c$.

Note that unlike the previous example where $P \subseteq R_s$, now we only know that $P \cap M^c \subseteq R_s$. Notably, whereas in the previous example we implicitly used the fact that $P[P\ \cap\ r_i = \emptyset] = P[R_s\ \cap\ r_i = \emptyset]$, we have only that $P[P \cap M^c\ \cap\ r_i = \emptyset] = P[R_s\ \cap\ r_i = \emptyset]$. Thus the probability that R_s misses $r_i\ is\ P[R_s\ \cap\ r_i = \emptyset] = P[P \cap M^c\ \cap\ r_i = \emptyset]$.

Rewriting the above using the identity $(A \cap B^c) \cup (A \cap B) = A, we\ have$

$$P[R_s\ \cap\ r_i\ \subseteq \emptyset] = P[P \subseteq P \cap (M\ \cup\ {r_i}^c)] = P[P = (M\ \cup\ {r_i}^c)]$$

(since $M\ \cup\ {r_i}^c \subseteq P$ we can cancel P on the RHS and replace containment with equality).

In other words, we are interested in the probability that all points in P are either outside ${r_i}^c$, or in M. By the union bound, for any $p \in P$, $P[p \in (M\ \cup\ {r_i}^c)] \leq P[p \in M] + P[p \in {r_i}^c]$.

From above, we know that $P[p \in M] = \eta \leq \eta'$, and $P[p \in {r_i}^c] = 1 - \frac{\epsilon}{4}$.

Now we can give an upper bound on $P[R_s \cap r_i = \emptyset]$:

$$P[R_s \cap r_i = \emptyset] \leq \left(1 - \frac{\epsilon}{4} + \eta'\right)^m \leq e^{\frac{m(4\eta' - \epsilon)}{4}}$$

b)

Given the above result, we have from the union bound that

$$P\left[\bigcup_{i=1}^{4} \{R_s \cap r_i = \emptyset\}\right] \leq 4e^{\frac{m(4\eta' - \epsilon)}{4}}$$

Problem 2:

a) (see below)

b)

Let $l(x)$ deonte the label received by the learner. Then $P[l(x) = c(x)] = 1 - \eta$

There are two events in which our hypothesis disagrees with the label given to the learner: either the label was correct and our hypothesis was erroneous, or the label was erroneous and our hypothesis was correct. Since these events are disjoint, we can sum their probabilities:

$$P[h(x) \neq l(x)] = P[(h(x) = c(x) \cap l(x) \neq c(x))] + P[h(x) \neq c(x) \cap l(x) = c(x)]$$

We now make the non-trivial assumption that the errors and hypotheses are statistically independent (for example, if the errors are completely random, and they are not correlated to any of the predictive variables that go into our hypothesis). Then the joint probabilities above can be written as

$$P[h(x) \neq l(x)] = P[h(x) = c(x)] * P[l(x) \neq c(x)] + P[h(x) \neq c(x)] * P[l(x) = c(x)]$$

$$(1 - error(h)) * (\eta) + \big(error(h)\big)(1 - \eta) = error(h)(1 - 2\eta) + \eta$$

Clearly, if h^* is our target hypothesis, then the generalization error is 0, so we have $P[h^*(x) \neq l(x)] = \eta$.

c) From the previous question, we know that

$$d(h) - d(h^*) = P[h(x) \neq l(x)] - P[h^*(x) \neq l(x)]$$

$$= error(h)(1 - 2\eta)$$

If $error(h) > \epsilon$ we then have

$$d(h) - d(h^*) \geq \epsilon(1 - 2\eta)$$

d) Let $\epsilon' = \epsilon(1-2\eta)$. Since $error(h) > \epsilon$, we have $d(h) - d(h^*) > \epsilon'$

Using the same argument as we use for bounding the generalization error in the inconsistent case, we have that

$$\Pr\left[\hat{d}(h^*) - d(h^*) > \frac{\epsilon'}{2}\right] \leq e^{-2m\epsilon'^2}$$

Equivalently,

$$\Pr\left[\hat{d}(h^*) - d(h^*) > \epsilon'\right] \leq e^{-\frac{m\epsilon'^2}{2}}$$

Setting the LHS to $\frac{\delta}{2}$ and solving for m, we get

$$m \geq \frac{2}{\epsilon'^2}\log\left(\frac{2}{\delta}\right)$$

e) If instead of fixing our hypothesis h, we would like to give a bound on any hypothesis in our hypothesis space, then we need to consider the union bound on the probability that any of the hypotheses have error>$\frac{\epsilon'}{2}$. Using a similar argument as before, we have that

$$\Pr\left[\bigcup_{h\in H} d(h) - \hat{d}(h) \geq \frac{\epsilon'}{2}\right] \leq |H|e^{-2m\epsilon'^2}$$

Where we have used the symmetry of the distribution of $\hat{d}(h)$ about $d(h)$ in to switch their order. Again setting this to $\frac{\delta}{2}$ and solving for m, we get

$$m \geq \frac{2}{\epsilon'^2}\left(\log|H| + \log\left(\frac{2}{\delta}\right)\right)$$

Note that $\hat{d}(h) - \hat{d}(h^*) = d(h) - d(h^*) + \hat{d}(h) - d(h) + d(h^*) - \hat{d}(h^*)$.

f) Plugging in our value for $\epsilon' = \epsilon(1-2\eta)$, we get

$$m \geq \frac{2}{\epsilon^2(1-2\eta)^2}\left(\log|H| + \log\left(\frac{2}{\delta}\right)\right)$$

Since $\frac{2}{\epsilon^2(1-2\eta)^2}\left(\log|H| + \log\left(\frac{2}{\delta}\right)\right) > \frac{2}{\epsilon'^2}\log\left(\frac{2}{\delta}\right)$,

This implies that with probability at least 1-δ,

$$\hat{d}(h) - d(h) \geq -\frac{\epsilon'}{2}$$

And

$$d(h^*) - \hat{d}(h^*) \geq -\frac{\epsilon'}{2}$$

Combining our results from the previous 2 sections, we get:

$$\hat{d}(h) - \hat{d}(h^*) \geq \epsilon' + \left(-\frac{\epsilon'}{2}\right) + \left(-\frac{\epsilon'}{2}\right) = 0$$