T-61.5040 Oppivat mallit ja menetelmät
T-61.5040 Learning Models and Methods
Pajunen, Viitaniemi

Solutions to exercise 5, 16.2.2007

Problem 1.

i) Consider a set of $n = v$ observations Z_v. Then, by definition, the first of the formulas defining v hold, $G(v) = v \log 2 = \log 2^v$, i.e. $\max N(Z_v) = 2^v$. That is, every possible dichotomy can be obtained using the given set of functions.
We can pick a subset of size n from the above set of observations Z_v, and again all possible dichotomies can be obtained for this set of observations. Thus $G(n) = n \log 2$ when $n \le v$.

In other words, linearity must hold for all n in $[1, v]$, there cannot be n for which the linearity does not hold. We can rewrite the definition of $G(n)$:

$$G(n) = \begin{cases} = n \log 2, & n \le v \\ \le v(\log(\frac{n}{v}) + 1), & n > v \end{cases}$$

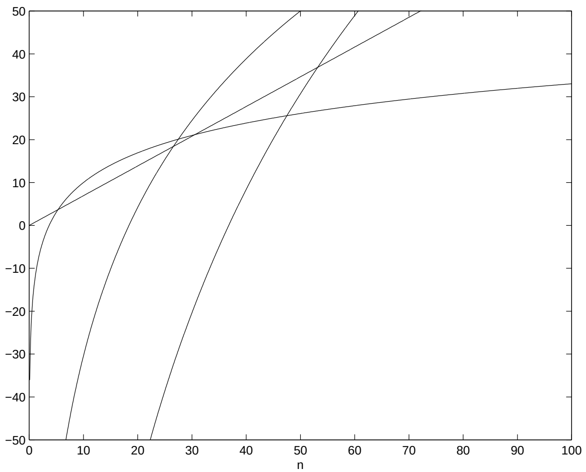

Figure 1: Straight line: $n \log 2$. Curves: $v(\log \frac{n}{v} + 1)$ where $v = 10, 50, 100$ (from left to right).

ii)

$$\begin{aligned} \lim_{n\to\infty} \frac{G(n)}{n} &\le \lim_{n\to\infty} \frac{v(\log(\frac{n}{v}) + 1)}{n} \\ &= \lim_{n\to\infty} \frac{v}{n}(\log n - \log v + 1) \\ &= \lim_{n\to\infty} \frac{v \log(n)}{n} - \frac{v \log v}{n} + \frac{v}{n} \\ &= \lim_{n\to\infty} \frac{v \log(n)}{n} \end{aligned}$$

Now use L'Hospital's rule: differentiate numerator and denominator separately with respect to n. We get

$$\lim_{n\to\infty} \frac{v \log(n)}{n} = \lim_{n\to\infty} \frac{v}{n} = 0$$

and thus

$$\lim_{n\to\infty} \frac{G(n)}{n} = 0.$$

Problem 2.

i) There are 2^{n-1} different vectors and $2^{n-1} - m$ are outside the training set. Since we don't know the last bits for the off-training set vectors, the expected contribution to the average error is $1/2 \left[\frac{2^{n-1}-m}{2^{n-1}}\right]$. Since the training error is s, its contribution to the average error is $ms/2^{n-1}$. So the average error is their sum

$$c = [2^{n-2} + (s - 1/2)m]/2^{n-1}.$$

If m is small and/or s is close to $1/2$, the average error is close to $1/2$. The only reason that the average error is not $1/2$ is that the average error considers also the training set. If training data is well classified ($s < 1/2$), then the average error is also slightly less than $1/2$. Off-training set error is always $1/2$.

Comments: note that this result seems to contradict SLT bounds. But considering the situation carefully, we note that we are not averaging over all possible training sets of size m. We have a given training set and conditional to that training set, the function h cannot generalize outside the training set. This is not surprising when considering the No Free Lunch theorems. Off-training set error is independent of the training error if we average uniformly over all problems.

ii) Denote by N the number of binary vectors incorrectly classified by h. Then $c = N/2^{n-1}$. The training error s is the number of such vectors picked in the training set, divided by m.

The inequality $|c - s| \le \epsilon$ is equivalent to $|c - z/m| \le \epsilon$ where $z = sm$.

We are going to use the Hoeffding inequality:

If $x_1, \dots, x_n$ are iid random variables for which $x_i - E(x_i) \in [a_i, b_i]$ and $X = \sum_i x_i$, then

$$p(X - E(X) \geq \epsilon) \leq \exp(-2\epsilon^2 / \sum_i (b_i - a_i)^2).$$

Write $c - s = \sum_{i=1}^m (c - z_i)/m$ where z_i denotes the classification error at point x_i. Then z_i is Bernoulli distributed with parameter c. Now look at $p(c - s \geq \epsilon)$. Hoeffding inequality applies for $(c - s)$ with $E(c - s) = \sum E(c - z_i)/m = 0$. As $(c - z_i)/m \in [(c-1)/m, c/m]$, $(b_i - a_i)^2 = 1/m^2$ and we get

$$p(c - s \geq \epsilon) \leq \exp(-2m\epsilon^2).$$

Changing signs we get

$$p(c - s \leq -\epsilon) \leq \exp(-2m\epsilon^2),$$

so

$$p(|c - s| \geq \epsilon) \leq 2\exp(-2m\epsilon^2)$$

Comments: this result explicitly shows what the SLT bound means. It examines the distribution of training errors s around their mean c. There is a binomial distribution behind this: randomly picking training points result in random errors and the sum of errors is tightly clustered around c. That is why it is unlikely that c and s are very far from each other, taken over all training sets.

Problem 3.

i) A reasonable prior would be such that nc has a binomial distribution $\text{Bin}(n, 1/2)$. If we choose h independent of the problem, each prediction it makes is equally likely to be or right or wrong. We furthermore assume that the prediction errors are independent, which leads to a binomial distribution.

ii) Probability of winning at least seven times is $p(c \leq 0.3)$. For a constant prior this is $4/11$. The expected winnings are then $4/11 * 10000 - 2500 \approx 1136 > 0$. Relying on a constant prior, you should take the deal.

The casino manager does not use a constant prior since he is still in business. He assumes that each outcome is independent of the others, and red/black are equally probable. Then $p(c \leq 0.3) = p(c = 0) + p(c = 0.1) + p(c = 0.2) + p(c = 0.3) = 2^{-10}(1 + 10 + 90/2 + 720/6) = 176/1024 \approx 0.17$.

The expected winnings for the casino are $\text{E}(-C) = 2500 - 10000 * 0.17 = 300$

Problem 4.

We consider the conditional density $p(c|s, h, m)$. We will compare the values of this for $c = 0.1$ and $c = 0.5$. Use Bayes' Theorem to write

$$p(c|s, h, m) \propto p(s|c, h, m)p(c|h, m).$$

On the right-hand side, the likelihood $p(s|c, h, m)$ is obtained by considering m random points when the average error is c. The product sm has a binomial distribution $\text{Bin}(m, c)$.

The new probability $p(c|h, m)$ is literally "what do we know about c when all we assume is h and m?" c is determined by the true function f, which we don't know, and is not affected by m. All we can assume is that our function h guesses correctly as often as not. We also assume that the guessing errors are independent. Therefore nc has a binomial distribution $\text{Bin}(n, 1/2)$.

Suppose $s = 0.1$, $n = 1000$, and $m = 100$. What is the posterior when $c = 0.1 = s$ and $c = 0.5$? These points are selected since they maximize the first and the second probability in the posterior correspondingly. The first probability $p(s|c, h, m)$ is

$$\binom{m}{sm} c^{sm}(1 - c)^{m - sm} \tag{1}$$

and the second probability $p(c|h, m)$ is

$$\binom{n}{cn} 2^{-n}. \tag{2}$$

Setting $c = 0.1 = s$, (1) gives $\binom{100}{10}(0.1)^{10}(0.9)^{90}$ and (2) gives $\binom{1000}{100}2^{-1000}$.

When $c = 0.5$, (1) gives $\binom{100}{10}2^{-100}$ and (2) gives $\binom{1000}{500}2^{-1000}$.

Compute the ratio $p(c = 0.1|s, h, m)/p(c = 0.5|s, h, m)$ to obtain

$$(0.1)^{10}(0.9)^{90}2^{100}\frac{\prod_{i=101}^{500} i}{\prod_{j=501}^{900} j}.$$

The dominating term is the last ratio of products which is very small. Numerically, the whole ratio is about $2 * 10^{-144}$. This shows that the posterior probability is significantly higher at $c = 0.5$.

Comments: Despite SLT bounds, we can't conclude that small s implies small c unless we are willing to make strong assumptions about the true model f (and therefore about $p(c|h)$).