Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 1 – Solution
January 31, 2006

Problem 1: Probability Review [50 points]

(1) [25 points]

The main result of this section is known as the standard *birthday paradox*. All floors have the same probability $\frac{1}{F}$ of being selected. Let $k \leq F$ be the number of people in the elevator and let $\bar{p}$ denote the probability that they all select a different floor. Fix the floor for the first person, then a distinct floor can be chosen for the second person with probability $(F-1)/F$. Fixing the first two distinct floors, the floor for the third person can be chosen with probability $(F-2)/F$ and so on. Thus, $\bar{p}$ is given by:

$$\bar{p} = 1 \cdot \frac{F-1}{F} \cdot \frac{F-2}{F} \cdots \frac{F-(k-1)}{F} = \prod_{i=0}^{k-1} \left(1 - \frac{i}{F}\right). \tag{1}$$

Using the approximation $e^{-x} \approx 1 - x$ which here holds for $k \ll F$,

$$\bar{p} \approx \prod_{i=0}^{k-1} e^{-\frac{i}{F}} = e^{-\sum_{i=0}^{k-1} \frac{i}{F}} = e^{-\frac{(k-1)k}{2F}}. \tag{2}$$

The right-hand side is smaller than 1/2 when:

$$k^2 - k - 2\log(2)F > 0. \tag{3}$$

that is for $k \geq \left\lceil \frac{1+\sqrt{1+8\log(2)F}}{2} \right\rceil$. For $F = 30$, it suffices that $k = 7$ for the probability that two people go the same floor to be more than half.

(2) [25 points]

Let q denote the probability of Professor Mamoru's floor. The other floors have the same probability r, thus $q + r(F-1) = 1$. Let $\bar{p}$ be defined as before. $\bar{p} = \bar{p}_1 + \bar{p}_2$ where $\bar{p}_1$ is the probability that all k floors be distinct and different from that of Professor Mamoru, and $\bar{p}_2$ the probability that they all be distinct with one of them Professor

Mamoru's floor. $\bar{p}_1$ can be computed using the same reasoning as in the previous section:

$$\bar{p}_1 = r(F-1) \cdot r(F-2) \cdots r(F-(k-1)) = (1-q)^k \prod_{i=0}^{k-1} \left(1 - \frac{i}{F-1}\right). \quad (4)$$

To compute $\bar{p}_2$, observe that Professor Mamoru's floor can be chosen in k different ways among the k floors. The other $k-1$ floors must be chosen as before among the equiprobable floors, thus:

$$\bar{p}_2 = kq \cdot r(F-1) \cdots r(F-(k-2)) = kq(1-q)^{k-1} \prod_{i=0}^{k-2} \left(1 - \frac{i}{F-1}\right). \quad (5)$$

Thus, the general expression of the probability that two persons go to the same floor is $1 - \bar{p}$ with:

$$\begin{aligned} \bar{p} &= (1-q)^{k-1} \prod_{i=0}^{k-1} \left(1 - \frac{i}{F-1}\right) \left(k(q - \frac{1-q}{F-1}) + \frac{F}{F-1}(1-q)\right) \\ &\approx (1-q)^{k-1} e^{-\frac{(k-2)(k-1)}{2F}} \left(k(q - \frac{1-q}{F-1}) + \frac{F}{F-1}(1-q)\right) \quad (6) \end{aligned}$$

For $q = .25$, it suffices that $k = 5$ persons take the elevator, for $q = .35$, $k = 4$ persons, and for $q = .5$ only $k = 3$. For $q = .5$, the answer is the same for $F = 1000$.

Problem 2: PAC Learning [50 points]

[Thanks to Tyler Neylon for writing both the problem and the solution.]

In both solutions, our training data is the set T and our learned concept $L(T)$ is the tightest circle (with minimal radius) which is consistent with the data.

(1) [25 points] Concentric Circles

Suppose our target concept c is the circle around the origin with radius r. We will choose a slightly smaller radius s by

$$s := \inf\{s' : P(s' \leq ||x|| \leq r) < \epsilon\}.$$

Let A denote the annulus between radii s and r; that is, $A := \{x : s \leq ||x|| \leq r\}$. By definition of s,

$$P(A) \geq \epsilon. \tag{7}$$

In addition, our generalization error, $P(c \,\Delta\, L(T))$ must be small if T intersects A. We can state this as

$$P(c \,\Delta\, L(T)) > \epsilon \Rightarrow T \cap A = \emptyset. \tag{8}$$

Using (7), we know that any point in T chosen according to P will "miss" region A with probability at most $1 - \epsilon$. Defining $error := P(c \,\Delta\, L(T))$, we can combine this with (8) to see that

$$P(error > \epsilon) \leq P(T \cap A = \emptyset) \leq (1 - \epsilon)^m \leq e^{-m\epsilon}.$$

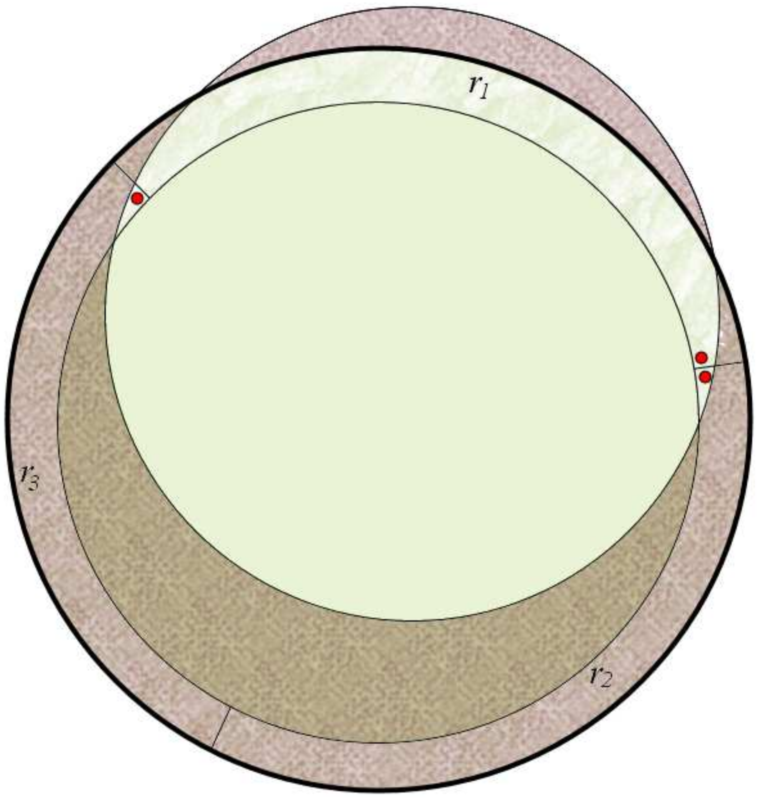
Setting δ to be greater than or equal to the right-hand side leads to $m \geq \frac{1}{\epsilon} \log(\frac{1}{\delta})$.

(2) [25 points] Non-Concentric Circles

As in the previous example, it is natural to assume the learning algorithm operates by returning the smallest circle which is consistent with the data. Gertrude is relying on the logical implication

$$error > \epsilon \Rightarrow T \cap r_i = \emptyset \text{ for some } i, \tag{9}$$

which is not necessarily true here. The figure illustrates a counterexample. In the figure, we have one training point in each region r_i. The points in r_1 and r_2 are very close together, and the point in r_3 is very close to region r_1. On this training data (some other points may be included outside the three regions r_i), our learned circle is the "tightest" circle including these points, and hence one diameter approximately traverses the corners of r_1. In the figure, the darkened areas are the error of this learned hypotheses versus the target circle which has a thick border. Clearly, the error may be greater than ϵ even while $T \cap r_i \neq \emptyset$ for any i; this contradicts (9) and invalidates poor Gertrude's proof.


r_1
r_3
r_2