Foundations of Machine Learning
Courant Institute of Mathematical Sciences
Homework assignment 1
Due: January 31, 2006

Problem 1: Probability Review

Professor Mamoru teaches at a university whose computer science and math building has $F = 30$ floors.

(1) Assume that the floors are independent and that they have the same probability to be selected by someone taking the elevator. How many people should take the elevator in order to make it likely (probability more than half) that two of them go to the same floor? [*Hint*: use the Taylor series expansion of $e^{-x} = 1 - x + \ldots$ and give an approximate general expression of the solution.]

(2) Professor Mamoru is popular and his floor is in fact more likely to be selected than others. Assuming that all other floors are equiprobable, derive the general expression of the probability that two persons go to the same floor, using the same approximation as before. How many people should take the elevator in order to make it likely that two of them go to the same floor when the probability of Professor Mamoru's floor is .25, .35, or .5? When $q = .5$, would the answer change if the number of floors were instead $F = 1000$?

(3) The probability models assumed in (1) and (2) are both naive. If you had access to the data collected by the elevator guard, how would you define a more faithful model.

Problem 2: PAC Learning

(1) Concentric Circles

Let $X = \mathbb{R}^2$ and consider the set of concepts of the form $c = \{(x, y) : x^2 + y^2 \leq r^2\}$ for some real number r. Show that this class can be (ϵ, δ)-PAC-learned from training data of size $m \geq (1/\epsilon) \log(1/\delta)$.

(2) Non-Concentric Circles

Let $X = \mathbb{R}^2$ and consider the set of concepts of the form $c = \{x \in \mathbb{R}^2 : ||x - x_0|| \leq r\}$ for some point $x_0 \in \mathbb{R}^2$ and real number r.

Gertrude, an aspiring machine learning researcher, attempts to show that this class of concepts may be (ϵ, δ) PAC-learned with sample complexity $m \geq (3/\epsilon)\log(3/\delta)$, but is having trouble with her proof. Her idea is that the learning algorithm would select the smallest circle consistent with the training data. She has drawn 3 regions r_1, r_2, r_3 around the edge of concept c, each of size $\epsilon/3$ in probability measure P (see the sketch below). She wants to argue that if the generalization error is $\geq \epsilon$, then one of these regions must have been missed by the training data, and hence this event will occur with probability at most δ. Can you tell Gertrude if her approach works?