MCS 590 – Foundations of Data Science

Fall 2017

Problem Set 2

Lev Reyzin

Due: 11/3/17 at the beginning of class

Instructions: Atop your problem set, please write your name and list your collaborators.

Problems

1. Given the set of integers $\{1, 2, \ldots, n\}$, what is the expected number of draws d with replacement until the integer 1 is drawn? What is the expected number of draws until every integer from the set is drawn? (This is needed for the expected cover time of K_n.)

2. What is the hitting time h_{uv} for two adjacent vertices on a cycle of length n? What is the hitting time if edge (u, v) is removed?

3. From Radon's theorem, we know that the VC dimension of a d-dimensional half-space is bounded from above by $d+1$. Suppose that the m data points, which live in d-dimensional space, when implicitly projected to a much (perhaps exponentially) higher dimensional space via a Kernel, are linearly separable by an SVM with a margin of $\rho > 0$. Using Occam's Razor for VC dimension, as well as the random projection result from Problem Set 1, what is the generalization bound this SVM achieves? (Do not use explicit margin bounds that were presented in class.)

4. Give an algorithm for efficiently PAC-learning 3-CNF formulae (conjunctions of clauses, each of which are disjunctions of three literals[1]), and argue that your algorithm is correct.

5. What is the VC-dimension of a) circles in the plane and b) triangles in the plane? (In both cases, the region contained inside the circle/triangle, resp. is labeled positive, and the rest is labeled negative.) Prove your answer.

[1] A literal is a variable or its negation