Exercises
Data Mining: Learning from Large Data Sets
FS 2016

LAS Group, Institute for Machine Learning
Dept. of Computer Science, ETH Zürich
Prof. Dr. Andreas Krause
Web: `http://las.ethz.ch/teaching/dm-f16/`
Email questions to:
Josip Djolonga, josipd@inf.ethz.ch

Series 25, Nov 25th, 2016 (Active Learning)

It is not mandatory to submit solutions and sample solutions will be published after one week. If you choose to submit your solution, please send an e-mail from your `ethz.ch` address with subject `Exercise5` containing a PDF (LaTeX or scan) to `josipd@inf.ethz.ch` until Thursday, 1 Dec 2nd 2016.

Problem 1 (Actively learning a union of intervals):

Suppose you are given a pool $X = \{x_1, \dots, x_n\}$ of n unlabeled examples where each $x_i \in [0, 1]$. Further suppose there are *unknown* constants $0 \leq a < b < c < d \leq 1$ such that all $x_i \in [a, b] \cup [c, d]$ are labeled with 1, whereas all remaining points are labeled with -1. We would like to develop a pool-based active learning scheme that infers the labels of all unlabeled examples. The algorithm sequentially selects one of the n examples and obtains its true label (i.e., there is no noise).

1. Show that in general, n labels are needed to infer the labels of all unlabeled examples.
2. For $x < x'$ define $E(x, x') = \left|X \cap [x, x']\right|$, i.e, the number of examples contained in the interval $[x, x']$. Suppose $E(a, b) \geq m$, $E(b, c) \geq m$ and $E(c, d) \geq m$ for some known constant $m \geq 1$.
   (a) Define an active learning scheme that selects examples given knowledge of m. How many samples are needed as a function of m and n?
   (b) Can you come up with an algorithm that works even without knowledge of m? That is, develop an algorithm that uses (approximately) the same number of labels as the algorithm in (a) with $m = \min\{E(a, b), E(b, c), E(c, d)\}$? You're allowed to use a randomized algorithm and bound the expected number of labels requested.