Project 1 - Random projections

In this project we will experiment with random hyperplanes for classification. Your program will take a dataset as input and produce new features following the procedure below. The input is in the same format as for previous assignments.

Inputs:

Data matrix X: n rows, m columns

Training labels: Y

Level: k

Let Z and Z1 be an empty list initially.

For i = 0 to k do:

(a) Create random vector w where each wj is uniformly sampled between -1 and 1.

(b) Let xj be our training data rows from X. Determine the largest and smallest wT xj across all training rowsxj in X. Select w0 randomly in the range [minj wT x; maxj wT xj ].

(c) Project training data X (each row is datapoint xj) onto w. Let projection vector zi be Xw + w0 (here X is n \_ m matrix and w are m \_ 1 vector).

Append (1 + sign(zi))=2 as new column to the right end of Z. Remember that zi is a n \_ 1 vector and so for each row zki of zi, (1 + sign(zki))=2 is 0 if zki < 0 and 1 otherwise.

(d) Project test data X1 (each row is datapoint x1j ) onto w. Let projection vector z1 i be X1w. Append zi i as new column to the right end of Z1.

1. Run linear SVM on Z and predict on Z1.

2. Do for values of k=10, 100, 1000, and 10000.

3. How does the error compare to liblinear on original data X and X1 for

each k?

Submit a document containing the error of linear SVM (cross-validated C) on the \_rst split of each of the six datasets on the course website. Do this on the original data representation and the new representation for all values of k. Submit your program that creates features and run LinearSVC (in Python scikit) on the new training data and predicts on the new test data. In Lin-

earSVC set the max iter parameter to 10000 so that we do a deep search.