IE613: Online Machine Learning Jan-Apr 2016

Assignment 1: Feb 6

Instructions: You are free to code in Python or Matlab or C. Your code should ask to input training set and output a classi er upon execution. Discussion among the class participants is highly encouraged. But please write your own code to make sure that you understand the algorithms. Submit the code by 16th February.

Train a classi er using the following datasets

1. Pima Indians Diabetes dataset
2. Ionosphere dataset

For each dataset, use the following algorithms

ERM on half-

spaces Logistic

regression SVM

Ada-Boost (use decision stumps as weak hypotheses)

Each dataset is divided into training data and test data. Report both training error and test error in the following format.

Pima Data

Algorithm	output (hypothesis/es)	training error	test error
ERM on half-spaces	Hypothesis w= [-0.9 0.23333333 0.17914573 -0.12545455 0.0968254 -0.01879433 1.23360656 -0.13046541 0.39333333]	30%	50%
Logistic regression	Hypothesis: w = [-9.05 9.78333333 50.64045226 -6.05909091 -1.21984127 -5.54148936 35.64232489 16.05941503 9.64333333]	31%	47%
SVM	Hypothesis w= [-1.1 -0.00294118 -0.31548223 -0.2147541 0.84545455 0.8212963 0.15259366 0.7047181 0.16020408]	22%	43%
Ada-Boost	Hypothesis [2, 0, 6, 1, 7, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2, 2]	27%	43%

Note: The Hypothesis for Adaboost represent the individual decision stumps i.e the dimension of the feature vector along which the decision tree or the threshold classifier has been based on. Dimension 2 for example represents the 3rd dimension of the feature vector,so on and so forth.

Ionosphere Data

Algorithm	output (hypothesis/es)	training error	test error
ERM on half-spaces	Hypothesis w= [3.7 -1.45 -0.32535 0.3728 0.1179 -0.465825 -1.796125 -1.98075 -0.64755 0.765875 2.9026 0.75625 0.1696 -0.24785 1.142875 -1.166475 -1.4858 1.054725 -0.229875 -0.993775 0.521425 -0.404375 -0.24975 0.1153 0.079075 -1.225025 1.1628 -0.156525 -2.646225 -0.45535 0.013375 0.105425 -0.831875 1.611125]	16%	12%
Logistic regression	Hypothesis: w = [175.95 -89.15 -6.1459 1.08765 -27.1029 -15.735125 -20.1828 -39.60285 -16.6774 -6.874 22.2626 22.93315 -0.71535 15.746425 -18.266225 5.9579 8.40525 -11.7154 35.8666 -2.2729 -14.384825 23.592225 -14.49995 -13.6275 -15.030925 -0.57795 51.3684 -16.078275 -26.239 -26.199325 -22.23775 -12.039625 7.305475 46.005825]	5%	7%
SVM	Hypothesis w= [0.55 -0.2 -1.36815 -0.01765 -0.9008 -0.325375 -0.2093 0.30555 -0.36975 0.18165 -0.257025 0.040925 0.107075 -1.21775 0.046425 1.0263 -0.196275 -0.0498 0.263125 0.0398 -0.0923 0.16675 0.1311 0.04245 -0.377225 0.760075 1.08045 0.6772 -0.385 -0.429175 0.341225 0.603575 0.211375 0.331075]	14%	11%
Ada-Boost	Hypothesis [1, 4, 0, 3, 14, 7, 1, 26, 32, 26, 21, 0, 30, 9, 27, 2, 26, 13, 7, 24, 26, 15, 7, 25, 22, 18, 30, 26, 17, 0, 1, 0, 1, 7, 0, 1, 9, 20, 7, 18, 13, 0, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26,	13%	11%

	26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26, 26]		