EE6550 Machine Learning, Spring 2016

Homework Assignment #3 Report

102061210 王尊玄

1. Base hypothesis set :

   Base hypothesis set, from which we pick a base classifier in every iteration of adaboost algorithm, is set to be the collection of hypotheses derived from each individual dimension of input data. Every base classifier follows either sign or negated sign of a particular dimension of input data. For adult training data, whose sample dimension is 109, we have base hypothesis set size equal to 109. Since the sign and negated sign of a particular dimension are difference set of each other, it is for sure either of the two but never both of them provides correct prediction with probability more than 50%, which is qualified to following inequality,

   $$\mathrm{P_m}(R(h_s) \leq 0.5 - \gamma) \geq 1 - \delta$$

   , and is a weak classifier.

   Besides, in case of raw input data not being binary data, we have to first convert them to binary by simply setting a threshold. In my code, I compute mean of all input data in one dimension as threshold of such dimension.

2. Result :

   Cross validation is performed to determined best choice of T, i.e. in our case, number of dimension of input data we care about. Cross validation error against different T of 5-fold and 10-fold cross validation is given as following table,

5-fold cross validation error vs T								
T	5	6	7	8	9	10	11	12
CrossErr	0.1936	0.1891	0.1833	0.1868	0.1874	0.188	0.1917	0.1946
T	13	14	15	16	17	18	19	20
CrossErr	0.1931	0.1962	0.1942	0.1974	0.1988	0.1997	0.1985	0.201

10-fold cross validation error vs T								
T	5	6	7	8	9	10	11	12
CrossErr	0.195	0.1891	0.1833	0.1878	0.1873	0.1876	0.1914	0.194
T	13	14	15	16	17	18	19	20
CrossErr	0.1932	0.1962	0.1948	0.1968	0.1979	0.1977	0.1997	0.2005

Both 5-fold and 10-fold cross validation, by choosing T with minimum cross validation error, give best choice of T as 7. Then, the only

parameter of my adaboost algorithm is set to best T, which is 7, and is run over entire training data, giving following hypothesis,

hypothesis with T = 7 over entire training data set							
dimension	1	2	13	28	29	34	67
negate	0	0	0	0	0	0	0
weight	0.1579	0.2105	0.5764	0.2036	0.2804	0.5685	0.1975
Accuracy over testing data			81.63%		(1 means do negate)		

The above hypothesis is saved in n5_entireDataSet.mat or n10_entireDataSet.mat files.

Additionally, I use the hypothesis with best performance over validation data set among $(\#\text{ of choice of T}) \times n$ hypotheses in n-fold cross validation process. Maximum accuracy of prediction (best performance) over validation data is shown as follows,

accuracy over validation set		
n	5	10
maxAcc	82.28%	82.96%

Giving 2 hypothesis and their performance over testing data is in the following,

hypothesis with T = 11 from 3th fold excluded training data set, 5-fold						
dimension	1	2	13	28	29	34
negate	0	0	0	0	0	0
weight	0.0868	0.2094	0.5747	0.2042	0.274	0.5702
dimension	40	42	47	54	62	
negate	0	0	0	1	0	
weight	0.1012	0.1602	0.1965	0.1235	0.0799	
Accuracy over testing data			81.91%	(1 means do negate)		

hypothesis with T = 8 from 3th fold excluded training data set, 10-fold								
dimension	1	2	13	28	29	34	42	47
negate	0	0	0	0	0	0	0	0
weight	0.905	0.2084	0.5717	0.2042	0.278	0.5685	0.1612	0.1981
Accuracy over testing data			82.17%			(1 means do negate)		

The above 2 hypotheses can be obtained from n5.mat and n10.mat respectively.