CS 5350/6350: Machine Learining Fall 2017

Homework 5

Handed out: Thursday November 16th, 2017
Due date: Saturday December 2nd, 2017

General Instructions

- You are welcome to talk to other members of the class about the homework. I am more concerned that you understand the underlying concepts. However, you should write down your own solution. Please keep the class collaboration policy in mind.
- Feel free discuss the homework with the instructor or the TAs.
- Your written solutions should be brief and clear. You need to show your work, not just the final answer, but you do *not* need to write it in gory detail. Your assignment should be **no more than 10 pages**. Every extra page will cost a point.
- Handwritten solutions will not be accepted.
- The homework is due by midnight of the due date. Please submit the homework on Canvas.

1 Logistic Regression

Solution:

1.
$$\frac{\mathrm{d}g(w)}{\mathrm{d}w} = \frac{\exp\big(-y_i\mathbf{w}^T\mathbf{x}_i\big) * (-y_i\mathbf{x}_i)}{(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i))} = \frac{(-y_i\mathbf{x}_i)}{(1+\exp(y_i\mathbf{w}^T\mathbf{x}_i))}$$

2.
$$\nabla(J) = \frac{\exp\big(-y_i\mathbf{w}^T\mathbf{x}_i\big) * (-y_i\mathbf{x}_i)}{(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i))} + \frac{\mathbf{2w}}{\sigma^2} = \frac{(-y_i\mathbf{x}_i)}{(1+\exp(y_i\mathbf{w}^T\mathbf{x}_i))} + \frac{\mathbf{2w}}{\sigma^2} \tag{1}$$

3.
Given a training set $S = \{(x_i, y_i)\}, x \in R^n, y \in \{-1, 1\}$
Initialize $w = 0 \in R^n$
For epoch = 1...T:
Shuffle the training set
For each training example$(x_i, y_i) \in S$:

$$w \Leftarrow w - \gamma_t * (\frac{\exp\big(-y_i\mathbf{w}^T\mathbf{x}_i\big) * (-y_i\mathbf{x}_i)}{(1+\exp(-y_i\mathbf{w}^T\mathbf{x}_i))} + \frac{2w}{\sigma^2})$$

Return w

2 Experiments

2.1 The task and data

2.2 Implementation Notes

2.3 Algorithms to Compare

Extra Credit

IDS is a recursive function. The deeper the tree gets, the more memories will be needed and the growth is exponential (2^{depth})). This will not only fill the stack really fast but also need a lot of time to iterate.

2.4 What to report

1. For each algorithm above, briefly describe the design decisions that you have made in your implementation. (E.g, what programming language, how do you represent the vectors, trees, etc.)

   In this assignment, I am using MATLAB. For reading data, I am using file open to read file one line at a time and close it after I am down. I assume that there are 70000 features since the maximum feature index I found is above 67000.
   Support Vector Machine Logistic regression: are basically the same expect loss function are difference. It is much like perceptron in assignment 2 so I used the old code and change its loss function.
   Naive Bayes: I created a 4 by n matrix to store all label count in my function. Row 1 is for y = 1 and label = 1; Row 2 is for y = 1 and label = 0; Row 3 is for y = 0 and label = 1; Row 4 is for y = 0 and label = 0. After obtaining all the data, I then calculate the probabilities.
   Bagged Forests: I changed some aspects of my ID3 in assignment 1 to make it about to take my reformed data.
   SVM over trees: I used a matrix to store every trees' prediction about every data set and use it as input feature-label data into my SVM. I also write a new error report function to check accuracy since data dimension changed.
   Logistic regression over trees: The same method applied for Logistic regression.

2. Report the best hyper-parameters and accuracy on training set and test set. Please fill Table 1.

Experiment Submission Guidelines

1. The report should detail your experiments. For each step, explain in no more than a paragraph or so how your implementation works.

2. *Your code should run on the CADE machines.* You should include a shell script, `run.sh`, that will execute your code in the CADE environment. Your code should produce similar output to what you include in your report.

	Best hyper-parameters	Average cross-validation accuracy	Training accuracy	Test Accuracy
SVM	$\gamma = 1, C = 10$	88.4316%	91.5543%	86.0638%
Logistic regression	$\gamma = 1, \sigma = 10$	81.7251%	81.1214%	77.3404%
Naive Bayes	all the same	50%	50%	50%
Bagged Forests	not applicable	75.8339%	75.4255%	75.8340%
SVM over trees	$\gamma = 1, C = 10$	83.32%	90.383%	82.128%
Logistic regression over trees	$\gamma = 1, \sigma = 1$	80.73%	81.263%	77.872%

Table 1: Result table

You are responsible for ensuring that the grader can execute the code using only the included script. If you are using an esoteric programming language, you should make sure that its runtime is available on CADE.

3. Please do not hand in binary files! We will *not* grade binary submissions.