CS 5350/6350: Machine Learining Fall 2015

Homework 5

Handed out: Nov 20, 2015
Due date: Dec 8, 2015

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 ask the instructor or the TAs questions about the homework.
- 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 glory 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 Naïve Bayes

Consider the Boolean function $f_{TH(3,7)}$. This is a threshold function defined on the 7 dimensional Boolean cube as follows: given an instance x, $f_{TH(3,7)}(x) = 1$ if and only if 3 or more of x's components are 1.

1. [4 points] Show that $f_{TH(3,7)}$ has a linear decision surface over the 7 dimensional Boolean cube.

2. [7 points] Assume that you are given data sampled according to the uniform distribution over the Boolean cube $\{0,1\}^7$ and labeled according to $f_{TH(3,7)}$. Use naïve Bayes to learn a hypothesis that predicts these labels. What is the hypothesis generated by the naïve Bayes algorithm? (You do not have to implement the algorithm here. You may assume that you have seen all the data required to get accurate estimates of the probabilities).

3. [4 points] Show that the hypothesis produced in the previous question does not represent this function.

4. [5 points] Are the naïve Bayes assumptions satisfied by $f_{TH(3,7)}$? Justify your answer.

2 EM Algorithm

There are two grocery stores in the neighborhood of the U: Smith's and Trader Joe's. Each store has n checkout lanes. The number of customers for each lane per unit time, say one day, is distributed according to Poisson distribution with parameter λ. That is, for the i'th checkout lane,

$$P(\#\text{ of customers for lane } i = x_i|\lambda) = \frac{\lambda^{x_i} e^{-\lambda}}{x_i!}$$

with parameters λ_S and λ_T for Smith's and Trader Joe's, respectively.

1. [10 points] Given a record of customer counts $(x_1, \cdots, x_n)$, where x_i denotes the number of customers went through lane i, what is the most likely value of λ?

2. [10 points] Assume now that you are given a collection of m records $\{(x_{j1}, \cdots x_{jn})\}$, where $j = 1, \cdots, m$. You do not know which record is from Smith's and which is from Trader Joe's. Assume that the probability of a record is from the Smith's is η. In other words, it means that the probability that a record is from Trader Joe's is $1 - \eta$. Explain the generative model that governs the generation of this data collection. In doing so, name the parameters that are required in order to fully specify the model.

3. [10 points] Assume that you are given the parameters of the model described above. How would you use it to cluster records to two groups, the Smith's and the Trader Joe's?

4. [10 points] Given the collection of m records without labels of which store they came from, derive the update rule of the EM algorithm. Show all of your work.

3 Experiment

We looked maximum a posteriori learning of the logistic regression classifier in class. In particular, we showed that learning the classifier is equivalent to the following optimization problem:

$$\min_{\mathbf{w}} \Big\{ \sum_{i=1}^{m} \log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i)) + \frac{1}{\sigma^2} \mathbf{w}^T \mathbf{w} \Big\}$$

In this question, you will derive the stochastic gradient descent algorithm for the logistic regression classifier, and also implement it with cross-validation. Detailed instructions on cross-validation procedure can be found in homework 1, and instructions on SGD can be found in homework 4.

1. [5 points] What is the derivative of the function $\log(1 + \exp(-y_i \mathbf{w}^T \mathbf{x}_i))$ with respect to the weight vector?

2. [5 points] The inner most step in the SGD algorithm is the gradient update where we use a single example instead of the entire dataset to compute the gradient. Write down the objective where the entire dataset is composed of a single example, say $(\mathbf{x}_i, y_i)$. Derive the gradient of this objective with respect to the weight vector.

3. [10 points] Write down the pseudo code for the stochastic gradient algorithm using the gradient from previous part.

4. [20 points] Implement your pseduo code as a training algorithm with cross-validation on the σ parameter. This parameter basically helps trade off between generalizability and model fit.

   Use the two `astro` data sets (original and scaled) from homework 4 to train and evaluate the learner. In your writeup, please report on the accuracy of your system, what value of sigma you chose based on cross validation, how many epochs you chose to run SGD, and a plot of the NEGATIVE log likelihood after each epoch of SGD.

As mentioned in previous homeworks, you may use any programming language for your implementation. Upload your code along with a script so the TAs can run your solution in the CADE environment.

4 HAPPY HOLIDAYS Extra Credit

25pts You've seen stochastic gradient decent applied to logistic regression. Now we ask why this is a viable strategy for optimizing this objective. Prove that SGD will find the optimal value for this function. This can be done by demonstrating that the objective is convex. There are many ways to prove this. One of the most straightforward ways to show this is demonstrate that the Hessian is positive-semidefninite.

1. [5 ppoins] Find the Gradient of:

$$\sum_{i=1}^{m} \log(1 + \exp(-y_i w^T x_i)) + \frac{1}{\sigma^2} w^T w$$

2. [5 points] Find the Hessian of (a).

3. [10 points] prove that the Hessian is positive semidefinite.

4. [5 points] Why does this prove that we are GAURANTEED to have within $\epsilon > 0$ of the right answer?