CMPS 242: Machine Learning: Fall 2016: HW 3

Due: 08th November 2016

General Instructions (please review carefully):

- The assignment is to be attempted in groups of two. If you choose to not work with a partner, **two** points will be automatically deducted from your score for this homework. It is your responsibility to find a partner.
- Each group needs to submit only one set of solutions. However, each group member should completely understand the group's solutions. It is **strongly** recommended that each student work on all of the problems individually before integrating their solutions into the group consensus. Dividing up the problems so each student does just a subset of them is the wrong approach.
- LaTeX is preferred, but neatly handwritten solutions will also be accepted. All solutions need to be handed over in the class before the beginning of the lectures.
- The names of the group members, and their UCSC ID (@ucsc.edu email address) should prominently be written on the upper left corner of the first page.
- Multiple sheets should be stapled together in the upper **left** corner.
- Solutions to the problems should be clearly labeled with the problem number.
- Although no points are given for neatness, illegible and/or poorly organized solutions can be penalized at the TA discretion.
- Clearly acknowledge sources (web, people, books, etc.), and mention if you discussed the problems with other students or groups. In all cases, the course policy on collaboration applies, and you should refrain from getting direct answers from anybody or any source. If in doubt, please ask the instructors or TAs.

Question 1 (2 points): Consider the following table which details the use of contraception by age of currently married women in El Salvador in 1985.

	Contraceptive Method		
Age	Ster.	Other	None
15 - 19	3	61	232
20 - 24	80	137	400
25 - 29	216	131	301
30 - 34	268	76	203
35 - 39	197	50	188
40 - 44	150	24	164
45 - 49	91	10	183

Let t_i denote an indicator variable which takes on values $\{1, 2, 3\}$ where 1 is Sterilization, 2 is Other, and 3 is None. Consider the activation function:

$$a_{ij} = \alpha_j + \beta_j x_i + \gamma_j x_i^2$$

where x_i is the mid-point of the i-th age group, and $j \in \{1, 2, 3\}$. Recall that the multinomial logit model is given by

$$p\left(t_i = j | x_i\right) = \frac{\exp\left(a_{ij}\right)}{\sum_{j'} \exp\left(a_{ij'}\right)}$$

Implement multinomial logistic regression from scratch (for the underlying optimization you may use modules such as the ones available in `scipy`) and apply it to the given data. Comment on your observations.

Question 2 (4 points): Download the Enron-spam dataset from `http://www.aueb.gr/users/ion/data/enron-spam/`. Use the pre-processed datasets Enron1, ..., Enron5 as training data and Enron6 as test data.

- Implement the Naive Bayes algorithm that we discussed in the class (a good in-depth reference is `http://nlp.stanford.edu/IR-book/pdf/13bayes.pdf`). Remember to perform all calculations on the log scale to prevent underflow.
  - Report the accuracy on the test set.
  - How do you account for different prior probabilities for spam and ham?
  - Does the performance of the classifier change when you do add one Laplace smoothing?
  - What are the most discriminative words as per the naive Bayes classifier?

Question 3 (4 points): Work on the same dataset as the above problem.

- Implement the binary logistic regression algorithm that we discussed in the class. You may use any optimizer of your choice including optimization modules from `scipy`.
  - Report accuracy on test set.
  - How do you account for different prior probabilities for spam and ham?
  - What are the most discriminative words as per the logistic regression classifier?
- Implement the Bayesian logistic regression algorithm that we discussed in the class.
  - Report accuracy on test set.
  - How does the accuracy change when the strength of the prior changes?

Comment on the differences and similarities between the three classifiers that you implemented to solve the spam detection problem (naive bayes, binary logistic regression and bayesian logistic regression)?

Learning Outcomes After this homework you should

- understand how to implement three different classifiers for a binary classification problem
- understand how to implement a multiclass logistic regression classifier
- interpret the results of various classification algorithms

Changelog

- 25 October 2016: First version created.