DS 5220 - Fall 2018 Olga Vitek

Homework 3

Analytical questions are best prepared using Latex/word, however photos of handwritten notes are also acceptable. For the questions involving programming, use a single notebook (Jupyter or R notebook) to answer all the programming questions. Run the code, explain the findings through markdown or visualizations and export it to PDF. Merge the notebook PDF with the rest of the files (latex or photos) of the homework. Submit the single PDF file through Blackboard.

Due on Blackboard before midnight on Tuesday October 16, 2018.
Each part of the problems 5 points

1. *[Analytical question.]* Show that for logistic regression with individual observations, the negative log-likelihood is convex (and therefore can be optimized by gradient descent).

2. *[Analytical question.]* Consider a problem of classifying a response Y with 3 classes, and two predictors X_1, X_2, using logistic regression.

   (a) Assume that X_1 and X_2 are continuous. Write the multi-class logistic regression, in softmax parametrization. State the total number of parameters.

   (b) Assume that X_1 is categorical with 5 categories, and X_2 is continuous. Write the multi-class logistic regression in softmax parametrization. State the total number of parameters.

3. *[Analytical question.]* Consider a problem of classifying a response Y with 3 classes, and two predictors X_1, X_2, using generative modeling. Assume that both X_1 and X_2 are continuous. For each of the following, state the model, and specify the total number of parameters:

   (a) Naïve Bayes with Gaussian distributions

   (b) Linear discriminant analysis

   (c) Quadratic discriminant analysis

4. *[Analytical question.]* Assume you have the following training set with three binary features X_1, X_2 and X_3, and a binary response Y.

X_1	X_2	X_3	Y
0	1	1	0
1	0	0	1
1	1	0	1
0	1	1	1
1	1	0	1
1	0	1	0
0	0	1	0

(a) Estimate $P(Y = 1|X_1 = 1, X_2 = 0, X_3 = 1)$ and $P(Y = 1|X_1 = 1, X_2 = 1, X_3 = 1)$ using Bayes rule, with the naïve Bayes assumption

(b) Estimate $P(Y = 1|X_1 = 1, X_2 = 0, X_3 = 1)$ and $P(Y = 1|X_1 = 1, X_2 = 1, X_3 = 1)$ using Bayes rule, without the naïve Bayes assumption

5. *[Implementation question.]*

   (a) Simulate the training set: N=50 values of X_1, distributed Uniformly on interval (0,3) and N=50 values of X_2 independent from X_1, distributed Uniformly on interval (0,3). Simulate the values of $Y \sim \mathit{Bernoulli}(\pi = 1/(1 + e^{-(-3+X_1+X_2)}))$. Repeat the above to simulate the validation set.

   (b) Implement batch gradient descent and stochastic gradient descent to estimate the parameters of logistic regression, as function of the learning rate α. Set the learning rate $\alpha = 0.08$ *[Note: consider a range of α to optimize convergence if needed.]*

   (c) Implement linear discriminant analysis.

   (d) Make a plot where the x axis is X_1, the y axis is X_2. Overlay the simulated observations from the validation set, labeled with their class. Overlay the true decision boundary (known from the process used to simulate the data) and the decision boundaries estimated by the three methods above. Report the proportion of correctly classified observations in the validation set, and interpret the results.

   (e) Repeat (a) 200 times. Each time, fit logistic regression with batch and gradient descent, and LDA, using your implementations. Record the % of correctly classified observations in the validation set with each method. Plot three histograms of % of correctly classified observations, and interpret the results.

6. *[Case study]* The dataset for this homework is posted on Piazza. It documents 482 initial public offerings (IPOs) of private companies. The goal is to predict which companies attract venture capital funding. *[Note: Use your own or existing library implementations of the methods.]*

   (a) Partition the dataset into a training, development and a validation subsets of equal size, by randomly selecting rows in the dataset. Explore the training set: report one-variable summary statistics, two-variable summary statistics, and discuss your findings.

   (b) Fit logistic regression on the training set. Consider transformations of variables, and the inclusion of higher-order terms if needed. Select the model with the best area under the ROC curve on the development set.

   (c) Fit linear discriminant analysis on the training set. Consider transformations of variables, and the inclusion of higher-order terms if needed. Select the model with the best area under the ROC curve on the development set.

   (d) Evaluate the performance of the classifiers using ROC curves on the training and on the validation set.

(e) Summarize your findings. How do the results differ between the training and the validation set? Which approach(es) perform(s) better on the validation set? What is are the reasons for this difference in performance? Which models are more interpretable?