Homework 3: Sampling Methods

Instructions: You may discuss this assignment with other students in the class, but you must submit your own answers to the questions below. Include an honor pledge with your submission. Submit on-line and in pdf, providing a pdf of your Python notebook. This homework is worth 100 points and the point totals for each question are shown in parentheses.

1. (33) Use the provided PyMC3 Bayesian Logistic Regression Classification Python notebook (https://www.kaggle.com/billbasener/pymc3-bayesian-logistic-regression-classification) to make conclusions about the predictor variables for using logistic regression for classification on the Iris data. In particular, determine which predictor variables might provide little or no predictive information because their coefficient might be zero. Support your statements using the trace plot, the forest plot, and the joint plots of the traces. Include copies of the plots to support your claim.

2. (33) With the CHD dataset(CHDdata.csv) from the previous homework, use Bayeisan model averaging with logistic regression in the Kaggle notebook (https://www.kaggle.com/billbasener/bayesian-model-averaging-logistic-regression) to determine the probability of inclusions of each of the factors.

   - Provide a bar chart using matplotlib.pyplot.bar to show the probabilities for each factor, and sort the factors by probability.
   - Create a bar chart sowing the model averaged coefficients for the predictor variables.
   - Standardize all of the numeric, continuous predictors using the mean and standard deviation, and create the previous two charts using the normalized variables. Do the results change?
   - Use the BMA (Bayesian Model Averaging) and OLS (Ordinary Least Squares) regression prediction methods at the bottom of the notebook to compare the accuracy of BMA to OLS. (Note that the full dataset is used for training and testing here.) Repeat the comparison using a proper test/train split and determine the prediction accuracy of BMA vs OLS. Do you see any effects of regularization when the size of the training set is small?

3. (33) With the CHD dataset(CHDdata.csv) from the previous homework, use PYMC3 to develop a sampling based estimate for the posterior distributions of the coefficients in a main effects logistic regression model. Use all predictor variables in the dataset and standardize all of the numeric, continuous predictors using the mean and standard deviation. For the continuous predictors, use a Gaussian prior with a mean vector of 0 and the identity matrix as the variance-covariance matrix. For the categorical predictors use uninformative priors, either a flat beta or Dirichlet as

appropriate. For results, show plots (either histograms or density plots) of the posterior distributions for each of the regression model parameters. Compare your results to part 2 of this assignment.

References

[1] Theodoridas, Sergios *Machine Learning: A Bayesian and Optimization-Perspective*, Elsevier, 2015.