Homework 2:
Bayes Optimal Classifiers

Instructions:

1. The majority of questions in this assignment require coding. So you may submit a jupyter notebook with both your code and your answers to the questions. You may also submit a combined pdf that shows your code and your answers.
2. You may discuss this assignment with other students in the class, but you must submit your own answers to the questions below.
3. Include an honor pledge with your submission.
4. Submit on-line.
5. This homework is worth 100 points and the point totals for each question are shown in parentheses.

1. (30) The authors in [1] describe a marketing campaign by a bank in Portugal. Modify the Exercise3.2LDAQDA_IrisSoution python code to use the data (bank-full.csv) from this marketing campaign with only the predictor variables age, balance, and duration and the response variable, y, to create the Bayes optimal classifiers for each of the conditions listed below. Provide your python code for each of these cases and show the decision you would make using this code for a 58 year old customer with balance of €261 and a duration in the company of 261 days. Discuss how your decision changes or not for each of the situations below.

   (a) Assume Gaussian class conditional likelihoods with unequal variance-covariance matrices with each of the following additional assumptions applied singularly to each decision rule in this class:

      i. Equal class priors and equal costs for misclassification;
      ii. The prior for not selecting the new bank service is 0.9 and misclassification costs are equal; and
      iii. The prior for not selecting the new bank service is 0.9 and the cost of classifying a customer as not a new service candidate when they are is 15 times the cost of classifying a customer as a new service customer

   (b) Assume Gaussian class conditional likelihoods with equal variance-covariance matrices with each of the following additional assumptions applied singularly to each decision rule in this class:

      i. Equal class priors and equal costs for misclassification;

ii. The prior for not selecting the new bank service is 0.9 and misclassification costs are equal; and

iii. The prior for not selecting the new bank service is 0.9 and the cost of classifying a customer as not a new service candidate when they are is 15 times the cost of classifying a customer as a new service customer

2. (25) Use numpy and pandas to develop a Naïve Bayes classifier for edible mushrooms with the data in MushroomData.csv and MushroomVariables.txt. Use $\frac{2}{3}$ of the observations to train the classifier and test on $\frac{1}{3}$. Submit your code and your testing results. Hint: Treat the observations as documents and the features with their values as words in a similar manner to the approach used to classify restaurant reviews.

3. (20) Implement in python a Bayesian generalized least squares approach to fit the model:

$$y = \theta_0 + \theta_1 x + \theta_2 \sqrt{x} + \theta_3 log(x) + \theta_4 x^3 + \eta$$

where the true generating model has $\boldsymbol{\theta} = (-1.7, -0.1, 3.2, 0.7, -0.8))$. Assume you know the true basis functions but not the coefficients. Also assume the model variance-covariance is $\sigma_\eta \boldsymbol{I}$, where σ_η is constant. Similarly, assume the prior is Gaussian with parameters, $\boldsymbol{\mu}_{prior} = (1, 1, 1, 1, 1, 1)$ and variance-covariance is $\sigma_\theta \boldsymbol{I}$. Compare the results from predicting 20 new values with two different values each for σ_η and σ_θ and three different values for N producing 12 total simulation runs. Explain the differences you observe in graphical presentations of the predictions with error bars plotted with the actual model generating the data. Submit your code and your results.

4. (10) We want to develop a model for lung cancer based on the following joint probability: $p(S, R, C)$, where S = smoking; R = Radon exposure; and C = lung cancer.

   (a) Show a Bayesian network model for this joint probability that captures the intuitive causality.

   (b) How do we expand the model in 4a to account for miners who have both a higher likelihood of smoking and a higher likelihood of radon exposure?

5. (15) Submit Bayesian network representations for each of the probabilistic models you developed in Problems 1 (a), 2 and 3. Do not use automatic graphing methods. Either draw these networks by hand or with drawing software.

References

[1] S. Moro, R. Laureano and P. Cortez. "'Using Data Mining for Bank Direct Marketing: An Application of the CRISP-DM Methodology." In P. Novais et al. (Eds.), *Proceedings of the European Simulation and Modelling Conference* - ESM'2011, pp. 117-121, Guimaraes, Portugal, October, 2011.