Problem Set 2 **AQM**

Maximum Likelihood and Logistic Regression

Problem 1 – Maximum Likelihood for Linear Regression

Consider a simple linear regression model, $Y = \beta_0 + \beta_1 X + \epsilon$ and our observation set consists of n predictor-response pairs, $(x_1, y_1), (x_2, y_2), ..., (x_n, y_n)$. Assume that the errors ϵ_i are normally distributed with mean 0 and variance σ^2: $\mathcal{N}(0, \sigma^2)$.

Consider a guess (b_0, b_1, s^2) for the true parameters $(\beta_0, \beta_1, \sigma^2)$, then compute the corresponding probability $P(y_i|x_i; b_0, b_1, s^2)$ of observing a response y_i for an input x_i? Also obtain the likelihood of observing the full dataset $\mathcal{L}(b_0, b_1, s^2)$ and subsequently, by maximizing the log-likelihood function with respect to the parameters, obtain the maximum-likelihood estimators for β_0, β_1 and σ^2.

Problem 2 – Optimizing the Likelihood Function

Suitably adapting the optimization routine (gradient descent, Newton-Raphson or any other optimization approach of your choice) that you have written in Problem Set 1, use it to maximize the log-likelihood function to numerically obtain the coefficients for a simple linear regression of Sales on TV in the **advertising.csv** dataset and verify that the maximum likelihood approach yields the same estimators as the sum of least-squares minimization method.

Problem 3 – Model Validation Methods

1. Using the **BreastCancer.csv** dataset, fit a logistic regression model using all predictor variables to obtain the confusion matrix (as performed in class). Compute the percentage of observations wrongly classified.

2. Firstly, let us adopt the *validation set approach* to test our model. Split the dataset randomly into a training and a test set. Train the logistic model on the training set and compute the percentage of observations wrongly classified in the test set.

3. Now, split the dataset randomly into $K = 10$ groups of roughly equal size. Perform *k-fold cross validation* using each of the K groups as the test set in turn, and the rest $K - 1$ groups as the training set. Using the percentage of wrongly classified observations as the metric, obtain the k-fold cross validation estimate, that is, the average misclassification percentage across the 10 runs.

4. What value should we choose for K? With $K = N$, the cross-validation estimator is approximately unbiased for the true (expected) prediction error, but can have high variance because N "training sets" are so similar to one another. The computational burden is also

considerable, requiring N applications of the learning method. On the other hand, with $K = 5$ say, cross-validation has lower variance, but bias could be a problem depending on how the performance of the learning method varies with the size of the training set (Hasty et al., The Elements of Statistical Learning, pg. 241-243). Using the **BreastCancer.csv** dataset, come up with a method on evaluating this bias-variance trade-off empirically, then select the K-fold that makes the most sense for this study.

Problem 4 – Cross-Validation Intuition

Consider a regression problem with 10,000 predictor variables, where you are forced to decide on one of two possible approaches to model the data:

1. (a) Perform variable selection and find the best subset of predictors that demonstrate strong correlation to response variable.
   (b) Using the subset of predictors, you train a multivariate regression.
   (c) You perform Leave-one-out cross-validation (LOOCV) to estimate the unknown tuning parameters for your model and estimate the prediction error.
2. (a) Perform variable selection such that the variables with the lowest standard deviation across samples get dropped.
   (b) Using the subset of predictors, you train a multivariate regression.
   (c) You perform 10-fold cross-validation to estimate the unknown tuning parameters for your model and estimate the prediction error.

Explain your rational as to why you would choose one method over the other.