SAS Advanced Predictive Modeling Exam

Neural Networks - 20%

Describe key concepts underlying neural networks

- Use SAS procedures to perform nonlinear modeling
  - Use the NLIN procedure for non-linear regression
- Explain advantages and disadvantages of using neural networks compared to other approaches
  - Explain two ways to respond to the black-box objection
  - Compare and contrast variable selection, degrees of freedom to traditional approaches
  - Explain advantages of the Widrow-Hoff Delta rule

Use two architectures offered by the Neural Network node to model either linear or non-linear input-output relationships

- Define the linear perceptron neural network
  - Define combination functions (linear, additive, equal slopes)
  - Define activation functions (logistic, tanh, arctan, softmax, exponential, identity)
  - Explain the difference between activation and link functions
- Be able to demonstrate how a linear perceptron is a generalized linear model that is able to model many target distributions
  - Explain the difference between a general and generalized model
  - Demonstrate the power of the NEURAL procedure in SAS
- Construct multilayer perceptrons
  - Define the three layers in a basic multilayer perceptron (input, hidden, output)
  - Explain how you can obtain a skip-layer network
- Construct radial basis function networks
  - Compare ordinary and normalized
- Identify advantages of using a radial basis function network over using a multilayer perceptron (invert order)

Use optimization methods offered by the SAS Enterprise Miner Neural Network node to efficiently search the parameter space in a neural network

- Describe the problem of local minima
- Explain the rationale behind the initialization settings
- Explain how early stopping and weight decay can be used to help avoid bad local minima

- Describe parameter estimation methods and determine best method to use
- List the assortment of error functions that are available in the Neural Networks node and determine the appropriate one to use based upon statistical considerations
    - Find the parameter set that minimizes the specified error function
    - Ordinary least squares
    - Maximum likelihood /Minimizing Deviance
    - Robust estimation methods
        - Huber's M-estimation (HUBER)
    - Determine the appropriate activation and error function combination to apply based on the target data
- List the optimization (training) techniques available in the Neural Networks node and determine the appropriate method to use based upon statistical considerations
    - iterative updating
    - back propagation
        - Conjugate gradient
        - Quasi-Newton
        - Levenberg-Marquardt

Construct custom network architectures by using the NEURAL procedure (PROC Neural)

- Working with SAS Enterprise Miner, use selected NEURAL procedure statements and PROC DMDB to construct neural networks
    - ARCHI
    - CONNECT
    - HIDDEN
    - INPUT
    - PRELIM
    - TARGET
    - TRAIN
- Define Sequential Network Construction (SNC) and use it to build an MLP(Multilayer Perceptron)
- Use weight interpretation to select relevant input variables
- Define a generalized additive neural network (GANN) and be able to explain the use of the GANN paradigm

Based upon statistical considerations, use either time delayed neural networks, surrogate models to augment neural networks

- Given a particular scenario/problem, use the time delayed neural network (TDNN) model to conduct time series analysis
- Apply a surrogate model to help understand a neural network's predictions
    - Interpret a neural network with a continuous target
    - Interpret a neural network with a discrete target

Use the HP Neural Node to perform high-speed training of a neural network

Logistic Regression - 30%

Score new data sets using the LOGISTIC and PLM procedures

- Use the SCORE statement in the PLM procedure to score new cases
- Use the CODE statement in PROC LOGISITIC to score new data
- Describe when you would use the SCORE statement vs the CODE statement in PROC LOGISTIC
- Use the INMODEL/OUTMODEL options in PROC LOGISTIC
- Explain how to score new data when you have developed a model from a biased sample

Identify the potential challenges when preparing input data for a model

- Identify problems that missing values can cause in creating predictive models and scoring new data sets
- Identify limitations of Complete Case Analysis
- Explain problems caused by categorical variables with numerous levels
- Discuss the problem of redundant variables
- Discuss the problem of irrelevant and redundant variables
- Discuss the non-linearities and the problems they create in predictive models
- Discuss outliers and the problems they create in predictive models
- Describe quasi-complete separation
- Discuss the effect of interactions
- Determine when it is necessary to oversample data

Use the DATA step to manipulate data with loops, arrays, conditional statements and functions

- Use ARRAYs to create missing indicators
- Use ARRAYS, LOOP, IF, and explicit OUTPUT statements

Improve the predictive power of categorical inputs

- Reduce the number of levels of a categorical variable
- Explain thresholding
- Explain Greenacre's method
- Cluster the levels of a categorical variable via Greenacre's method using the CLUSTER procedure
    - METHOD=WARD option
    - FREQ, VAR, ID statement
    - Use of ODS output to create an output data set
- Convert categorical variables to continuous using smooth weight of evidence

Screen variables for irrelevance and non-linear association using the CORR procedure

- Explain how Hoeffding's D and Spearman statistics can be used to find irrelevant variables and non-linear associations
- Produce Spearman and Hoeffding's D statistic using the CORR procedure (VAR, WITH statement)
- Interpret a scatter plot of Hoeffding's D and Spearman statistic to identify irrelevant variables and non-linear associations

Screen variables for non-linearity using empirical logit plots

- Use the RANK procedure to bin continuous input variables (GROUPS=, OUT= option; VAR, RANK statements)
- Interpret RANK procedure output
- Use the MEANS procedure to calculate the sum and means for the target cases and total events (NWAY option; CLASS, VAR, OUTPUT statements)
- Create empirical logit plots with the GPLOT procedure
- Interpret empirical logit plots

Apply the principles of honest assessment to model performance measurement

- Explain techniques to honestly assess classifier performance
- Explain overfitting
- Explain differences between validation and test data
- Identify the impact of performing data preparation before data is split

Assess classifier performance using the confusion matrix

- Explain the confusion matrix
- Define: Accuracy, Error Rate, Sensitivity, Specificity, PV+, PV-
- Explain the effect of oversampling on the confusion matrix
- Adjust the confusion matrix for oversampling

Model selection and validation using training and validation data

- Divide data into training and validation data sets using the SURVEYSELECT procedure
- Discuss the subset selection methods available in PROC LOGISTIC
- Discuss methods to determine interactions (forward selection, with bar and @ notation)
- Create interaction plot with the results from PROC LOGISTIC
- Select the model with fit statistics (BIC, AIC, KS, Brier score)

Create and interpret graphs (ROC, lift, and gains charts) for model comparison and selection

- Explain and interpret charts (ROC, Lift, Gains)
- Create a ROC curve (OUTROC option of the SCORE statement in the LOGISTIC procedure)
- Use the ROC and ROCCONTRAST statements to create an overlay plot of ROC curves for two or more models
- Explain the concept of depth as it relates to the gains chart

Establish effective decision cut-off values for scoring

- Illustrate a decision rule that maximizes the expected profit
- Explain the profit matrix and how to use it to estimate the profit per scored customer
- Calculate decision cutoffs using Bayes rule, given a profit matrix
- Determine optimum cutoff values from profit plots
- Given a profit matrix, and model results, determine the model with the highest average profit

Predictive Analytics on Big Data - 40%

Build and interpret a cluster analysis in SAS Visual Statistics

- Assign roles for cluster analysis
- Set cluster matrix properties (number, seed, etc)
- Select the proper inputs for the k-means algorithm for a given cluster analysis scenario
- Choose the number of clusters for a given cluster analysis scenario
- Set Parallel coordinate properties for cluster analysis
- Interpret a cluster matrix
- Interpret a parallel coordinates plot
- Display summary statistics for clusters
- Interpret summary statistics for clusters
- Assign cluster IDs to the data within Visual Statistics
- Score observations into clusters based on the results from Visual Statistics

Explain SAS high-performance computing

- Identify limitations of traditional computing environments
- Describe the characteristics of SAS High-Performance Analytics procedures
- Compare SMP and MPP computing modes
- Distinguish between HPA and the LASR related operation

Perform principal component analysis

- Explain how principal component analysis is performed
- List the benefits and problems of principal component analysis

- Distinguish between clustering, variable clustering, and principal component analysis
- Determine the number of principal components to retain
- Compare IMSTAT, Visual Statistics, and High Performance Computing nodes in Enterprise Miner

Analyze categorical targets using logistic regression in SAS Visual Statistics

- Assign roles for logistic regression
- Assign properties for logistic regression
- Filter data used for logistic regression
- Interpret logistic regression results (fit summary, residual plots, ROC/Lift charts, etc)
- Use Group-By variables to perform binary logistic regression

Analyze categorical targets using decision trees in SAS Visual Statistics

- Assign roles for decision trees
- Assign properties for decision trees
- Interpret decision trees results (trees, leaf statistics, assessment, etc)
- Identify variable importance with decision trees for use in other analysis techniques
- Splitting criteria used by Visual Statistics

Analyze categorical targets using decision trees in PROC IMSTAT

- Use the DECISIONTREE statement to create decision trees
- Define input variables with the INPUT and NOMINAL options
- Create and retrieve saved trees for input data scoring with the SAVE, TREETAB, and ASSESS options
- Evaluate the output of ODS tables (DTREE, DTreeVarImpInfo, DTREESCORE, etc) from decision trees
- Use the ASSESS statement to create data sets for evaluating the decision tree model
- Perform honest assessment on PROC IMSTAT decision trees
- Assess decision trees using ODS statistical graphics (SGPLOT)

Analyze categorical targets using logistic regression in PROC IMSTAT

- Assign variables to roles for logistic regression in PROC IMSTAT
- Create logistic regression in PROC IMSTAT using the LOGISTIC statement
- Use selected options of the LOGISTIC STATEMENT (ROLEVAR, INPUTS, SCORE, CODE, SHOWSELECTED, SLSTAY=)
- Assess logistic regression models using ODS statistical graphics (SGPLOT)
- Perform honest assessment on PROC IMSTAT logistic regression

Build random forest models with PROC IMSTAT

- Describe random forests
- Use the RANDOMWOOODS statement to build a forest of trees

- Score data with the RANDOMWOODS score code
- List benefits of forests
- Interpret random forests
- Identify variable importance with forest for use in other analysis techniques

Analyze interval targets with SAS Visual Statistics

- Build linear regression models in SAS Visual Statistics
- Assign roles for linear regression models
- Set properties for linear regression models
- Assess a linear regression model (evaluate Fit summary statistics, residual plot, influence plot, summary table, etc)
- Assess linear model assumption violations and recognize when linear model is inadequate
- Build generalized linear models in SAS Visual Statistics
- Assign roles for generalized linear models
- Set properties for generalized linear models
- Assess a generalized linear model (evaluate Fit summary statistics, residual plot, assessment, etc)

Analyze interval targets with PROC IMSTAT

- Use GENMODEL and GLM statements
- Distinguish between GENMODEL and GLM statements and the results of each procedure
- Assign variables to roles for GENMODEL and GLM statements in PROC IMSTAT
- Create models with GENMODEL and GLM statements in PROC IMSTAT
- Use selected options of the GENMODEL and GLM statements in PROC IMSTAT
- Assess models using ODS statistical graphics (SGPLOT)
- Perform honest assessment on PROC IMSTAT linear models

Analyze zero inflated models with HPGLM in Enterprise Miner

- Identify when it would be appropriate to use mixture distribution
- Describe the link functions and distributions available in the HP GLM node
- Build a zero inflated generalized linear model in EM
- Describe restrictions on roles and levels in input data sources for generalized linear models in EM
- Assess a zero inflated generalized linear model (evaluate Fit summary statistics, residual plot, assessment, etc)

Open Source Models in SAS - 10%

Incorporate an existing R program into SAS Enterprise Miner

- Enable R language statements to connect SAS to R
- Use the Open Source Integration node in SAS Enterprise Miner
  - Modes of operation (training, output)
  - Use Predictive Modeling Markup Language (PMML) in Open Source Integration Node
- Use Enterprise Miner variable handles to alter an R script
  - Use Enterprise Miner to run a random forest in R

Incorporate an existing Python program into SAS Enterprise Miner

- Determine steps to perform in SAS to incorporate a Python model
- Determine nodes in Enterprise Miner to incorporate a Python model
- Determine the necessary set up requirements for running Python models in SAS

Note: All 30 main objectives will be tested on every exam. The 185 expanded objectives are provided for additional explanation and define the entire domain that could be tested.