IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

LOGISTIC REGRESSION PART-6

LECTURE 51

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

LOGISTIC REGRESSION

- Equivalent of linear regression for categorical outcome variable
  - Predictors can be categorical or continuous
- Applied in following tasks
  - Classification task
    - Predicting the class of a new observation
  - Profiling
    - Understanding similarities and differences among groups

LOGISTIC REGRESSION

- Steps for logistic regression
  - Estimate probabilities of class memberships
  - Classify observations using probabilities values
    - Most probable class method: assign the observation to the class with highest probability value
      - Equivalently, for a two-class case, cutoff value of 0.5 can be used
    - Class of interest: user specified cutoff value
      - For a two-class case, typically a value greater than average probability value for class of interest, but less than 0.5 can be used

LOGISTIC REGRESSION

- Logistic Regression Model
  - Used typically in cases when structured model is preferred over data-driven models for classification tasks
  - Categorical outcome variable cannot be directly modeled as a linear function of predictors
    - Inability to apply various mathematical operators
    - Variable type mismatches
    - Range reasonability issues
      - LHS range={0, …, m-1}
      - RHS range=(-∞, ∞)

LOGISTIC REGRESSION

- Logistic Regression Model
  - Instead of using outcome variable (Y) in the model, a function of Y, called *logit* is used
- Logit
  - Think about modeling probability value as a linear function of predictors, specifically in a two-class case

If P is the probability of class 1 membership

$$P = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

Where p is the no. of predictors

LOGISTIC REGRESSION

- Logit
  - LHS range improves from {0, 1} to [0, 1], however still cannot match RHS
  - Can we bring RHS range to [0,1]?
    - Nonlinear approach
  - Typically, a nonlinear function of the following form is used to perform the required transformation

$$P = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p)}}$$

This function is called *logistic response function*

LOGISTIC REGRESSION

- Logit
  - Rearrange the previous equation as below:

$$\frac{P}{1-P} = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}$$

LHS is expression for *odds*, another measure of class membership

$$odds = \frac{P}{1-P}$$

  - Odds of belonging to a class is defined as ratio of probability of class 1 membership to probability of class 0 membership
    - This metric is popular in sports, horse racing, gambling, and many other areas

LOGISTIC REGRESSION

- Logit
  - Previous equation can be rewritten as

$$odds = e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_p x_p}$$

    - Range is now (0, ∞)
  - Take log on both sides of previous equation

$$\log(odds) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_p x_p$$

    - Standard logistic model
    - Now, LHS and RHS both have same range (-∞, ∞)
  - Log(odds) is called logit
    - Logit is used as the outcome variable in the model instead of categorical Y

LOGISTIC REGRESSION

- Odds and logit can be written as a function of probability of class 1 membership
  - Open RStudio

- In logistic regression model, we predict the logit values and therefore corresponding probability of a categorical outcome
  - Predicted probabilities values become the basis for classification
  - A prediction model for classification task

LOGISTIC REGRESSION

- Estimation Technique
  - Least squares method used in multiple linear regression cannot be used
    - Non-linear formulation of logistic regression
  - Maximum likelihood method is used
    - Estimates are optimized in order to maximize the likelihood of obtaining the observations used in training the model
    - Less robust than estimation techniques used in linear regression
    - Reliability of estimates
      - Outcome variable categories should have adequate proportion
      - Adequate sample size w.r.t no. of estimates

LOGISTIC REGRESSION

- Estimation Technique
  - Maximum likelihood method is used
    - Collinearity issues similar to linear regression
- Interpretation of Results
  - Logit model
    - Additive factor (β)
      - If $\beta < 0$, increase in x => decrease in logit values
      - If $\beta > 0$, increase in x => increase in logit values
    - For any value of x, interpretative statements of results are same

LOGISTIC REGRESSION

- Interpretation of Results
  - Odds model
    - Multiplicative factor (e^{β})
      - If $\beta < 0$, increase in x => decrease in odds
      - If $\beta > 0$, increase in x => increase in odds
    - For any value of x, interpretative statements of results are same
  - Probability model
    - For a unit increase in a particular predictor, corresponding change in the probability value is not a constant, while holding all other predictors constant
      - Depends on the specific values of the predictor
    - Interpretative statements of results depend on specific values of x

LOGISTIC REGRESSION

- Odds and odds ratios
  - Odds is a ratio of two probability values (prob. of class 1/prob. Of Class 0)
  - Odds ratio is ratio of two odds (odds of class m1/odds of class m2)
    - Odds ratio > 1 => odds of class m1 are higher than class m2

- Open RStudio

LOGISTIC REGRESSION

- Linear Regression for a categorical outcome variable?
  - Can be done by treating the outcome variable as continuous and coding it numerically
  - However, anomalies will lead to spurious modeling
    - Predictions can take any value, not just dummy values {0,1}
    - Outcome variable or residuals don't follow normal distribution
      - binomial distribution
    - Variance of outcome variable is not constant across all records (violation of homoscedasticity)
      - $np(1-p)$

LOGISTIC REGRESSION

- Logistic Regression for Profiling Task
  - Apart from model performance on validation partition
  - Model's fit to data is assessed on training partition
    - However, still avoid overfitting
    - Usefulness of predictors is examined
  - Goodness of fit metrics
    - Overall fit of the model
      - Deviance (equivalent to SSE in linear regression)
      - 1 - Deviance/Null Deviance (equivalent to multiple R^2 in linear regression)
    - Single predictors

LOGISTIC REGRESSION

- Outcome variable with m classes (m>2)
  - Multinomial logistic regression
    - Separate binary logistic regression model for m-1 classes (one class is treated as reference class)
  - Ordinal logistic regression
    - Large no. of ordinal classes: treat ordinal variable as continuous variable and apply multiple linear regression

LOGISTIC REGRESSION

- Outcome variable with m classes (m>2)
  - Ordinal logistic regression
    - Small no. of ordinal classes: Proportional odds or cumulative logit method
      - Separate binary logistic regression model for m-1 cumulative probabilities

For a three class case: C1, C2, and C3 and a single predictor x1

$$logit(C1) = \alpha_0 + \beta_1 x_1$$
$$logit(C1 or\ C2) = \beta_0 + \beta_1 x_1$$

- RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

LOGISTIC REGRESSION PART-7

LECTURE 52

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

LOGISTIC REGRESSION

- Linear Regression for a categorical outcome variable?
  - Can be done by treating the outcome variable as continuous and coding it numerically
  - However, anomalies will lead to spurious modeling
    - Predictions can take any value, not just dummy values {0,1}
    - Outcome variable or residuals don't follow normal distribution
      - binomial distribution
    - Variance of outcome variable is not constant across all records (violation of homoscedasticity)
      - np(1-p)

LOGISTIC REGRESSION

- Logistic Regression for Profiling Task
  - Apart from model performance on validation partition
  - Model's fit to data is assessed on training partition
    - However, still avoid overfitting
    - Usefulness of predictors is examined
  - Goodness of fit metrics
    - Overall fit of the model
      - Deviance (equivalent to SSE in linear regression)
      - 1 - Deviance/Null Deviance (equivalent to multiple R^2 in linear regression)
    - Single predictors

LOGISTIC REGRESSION

- Outcome variable with m classes (m>2)
  - Multinomial logistic regression
    - Separate binary logistic regression model for m-1 classes (one class is treated as reference class)
  - Ordinal logistic regression
    - Large no. of ordinal classes: treat ordinal variable as continuous variable and apply multiple linear regression

LOGISTIC REGRESSION

- Outcome variable with m classes (m>2)
  - Ordinal logistic regression
    - Small no. of ordinal classes: Proportional odds or cumulative logit method
      - Separate binary logistic regression model for m-1 cumulative probabilities

    For a three class case: C1, C2, and C3 and a single predictor x1

$$logit(C1) = \alpha_0 + \beta_1 x_1$$
$$logit(C1 or\ C2) = \beta_0 + \beta_1 x_1$$

- RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

ARTIFICIAL NEURAL NETWORKS

LECTURE 53

DR. GAURAV DIXIT
DEPARTMENT OF MANAGEMENT STUDIES

ARTIFICIAL NEURAL NETWORKS

- Based on
  - Human learning and memory properties
  - Capacity to generalize from particulars
  - Biological activity of brain, where interconnected neurons learn from experience
- Can model complex relationships between outcome variable and set of predictors
  - Applications in Finance (credit card fraud) and engineering disciplines (autonomous vehicle movement)

ARTIFICIAL NEURAL NETWORKS

- Can model complex relationships between outcome variable and set of predictors
  - Flexible data driven model
    - Not required to specify the form of relationship
    - Useful technique, when functional form of relationship is complicated or unknown
    - Linear and logistic regressions can be conceptualized as special cases
- Neural Network Architectures
  - Multilayer feedforward networks

ARTIFICIAL NEURAL NETWORKS

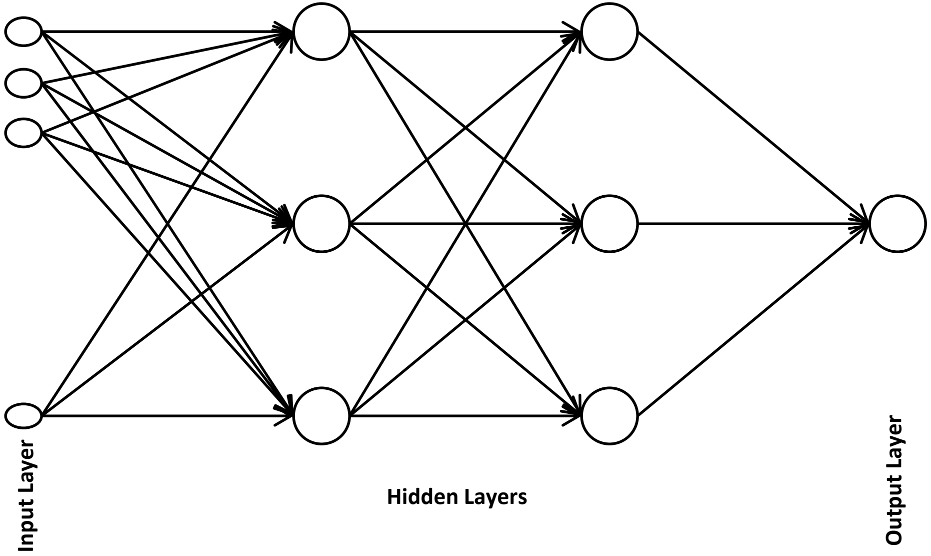
- Multilayer feedforward networks
  - Fully connected networks
    - Comprising of multiple layers of nodes
    - With one-way flow and no cycles
  - Input layer
    - First layer of the network
  - Hidden layers
    - Layers between input and output layer

ARTIFICIAL NEURAL NETWORKS

- Multilayer feedforward networks
  - Output layer
    - Last layer of the network
  - Nodes receive feed from previous layer and forward it to next layer after applying a particular function
  - Function used to map input values (received feed) to output values (forwarded feed) at a node is typically different for each type of layers

ARTIFICIAL NEURAL NETWORKS

- Multilayer feedforward networks



ARTIFICIAL NEURAL NETWORKS

- Multilayer feedforward networks
  - Each arrow from node i to node j has a value w_{ij} indicating weight of the connection
  - Each node in the hidden and output layers also has a bias value, θ_j (equivalent to intercept term)
- Computing output values at nodes of each layer type
  - Input layer nodes
    - No. of nodes are typically equal to no. of predictors, p
    - Each node will receive input values from its corresponding predictor
    - Output is same as input, that is, predictor’s value

ARTIFICIAL NEURAL NETWORKS

- Computing output values at nodes of each layer type
  - Hidden layer nodes
    - Sum of bias value and weighted sum of input values received from previous layer is computed

$$\theta_j + \sum_{i=1}^{p} w_{ij} x_i$$

    - Function g (referred as transfer function) is applied on this sum to produce the output values
    - Transfer function could be a monotone function, for example:
      - Linear function: $g(x) = bx$
      - Exponential function: $g(x) = e^{bx}$
      - Logistic or sigmoidal function: $g(x) = 1/1+e^{-bx}$

ARTIFICIAL NEURAL NETWORKS

- Computing output values at nodes of each layer type
  - Hidden layer nodes
    - θ_j and w_{ij} are typically initialized to small random values in the range 0.0±0.05
    - Network updates these values after learning from data during each iteration or round of training
  - Output layer nodes
    - Steps are same as for hidden layer nodes, except the fact that input values are received from last hidden layer
    - Output values produced by nodes are used as
      - Predictions in a prediction task
      - Scores to be used to classify a record in a classification task

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

ARTIFICIAL NEURAL NETWORK PART-2

LECTURE 54

DR. GAURAV DIXIT
DEPARTMENT OF MANAGEMENT STUDIES

ARTIFICIAL NEURAL NETWORKS

- Computing output values at nodes of each layer type
  - Hidden layer nodes
    - θ_j and w_{ij} are typically initialized to small random values in the range 0.0 ± 0.05
    - Network updates these values after learning from data during each iteration or round of training
  - Output layer nodes
    - Steps are same as for hidden layer nodes, except the fact that input values are received from last hidden layer
    - Output values produced by nodes are used as
      - Predictions in a prediction task
      - Scores to be used to classify a record in a classification task

ARTIFICIAL NEURAL NETWORKS

- Open RStudio
- Neural Network training process
  - Steps to compute neural network output values are repeated for all the records in the training partition
  - Prediction errors are used for learning after each iteration
- Linear and Logistic regression as special cases
  - A neural network with single output node and no hidden layers would approximate the linear and logistic regression models

ARTIFICIAL NEURAL NETWORKS

- Linear and Logistic regression as special cases
  - If a linear transfer function (g(x) = bx) is used, output would be

$$y = \theta + \sum_{i=1}^{p} w_i x_i$$

    - A formulation equivalent to multiple linear regression equation
    - However, estimation method (least squares) is different from neural network (back propagation)

ARTIFICIAL NEURAL NETWORKS

- Linear and Logistic regression as special cases
  - If a logistic transfer function ($g(x) = 1/1+e^{-bx}$) is used, output would be

$$P(y=1) = \frac{1}{1+e^{\theta+\sum_{i=1}^{p} w_i x_i}}$$

    - A formulation equivalent to logistic regression equation
    - However, estimation method (maximum-likelihood method) is different from neural network (back propagation)

ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Scale of [0,1] is typically recommended for neural network models for performance purposes
  - For numeric variables,

$$V_{norm} = \frac{V - \min(V)}{\max(V) - \min(V)}$$

ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Binary variables (categorical variables with two classes)
    - Create dummy variables: set of values {0, 1}
  - Nominal variables with m (>2) classes
    - Create m-1 dummy variables: set of values {0, 1}
  - Ordinal variables with m (>2) classes
    - Map the values to the set {0, 1/(m-1), 2/(m-1), …, (m-2)/(m-1), 1}

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

ARTIFICIAL NEURAL NETWORK PART-3

LECTURE 55

DR. GAURAV DIXIT
DEPARTMENT OF MANAGEMENT STUDIES

ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Scale of [0,1] is typically recommended for neural network models for performance purposes
  - For numeric variables,

$$V_{norm} = \frac{V - \min(V)}{\max(V) - \min(V)}$$

ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Binary variables (categorical variables with two classes)
    - Create dummy variables: set of values {0, 1}
  - Nominal variables with m (>2) classes
    - Create m-1 dummy variables: set of values {0, 1}
  - Ordinal variables with m (>2) classes
    - Map the values to the set {0, 1/(m-1), 2/(m-1), …, (m-2)/(m-1), 1}

ARTIFICIAL NEURAL NETWORKS

- Other transformations
  - Transformations which could spread the values more symmetrically can be done for performance purposes
    - Log transform of a right-skewed variable

- Estimation method
  - Least squares and maximum likelihood methods use a global metric of errors (e.g., SSE) to estimate the parameters

ARTIFICIAL NEURAL NETWORKS

- Estimation method
  - Neural networks use error values of each observation to update the parameters in an iterative fashion (referred as learning)
    - Error for the output node (prediction error) is distributed across all the hidden layer nodes
    - All hidden layer nodes share responsibility for part of the error (referred as node-specific error)
    - Node-specific errors are used to update the connection weights and bias values

ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - An algorithm to update weights and bias values of a neural network
  - Error values are computed from output layer back to hidden layers
    - All hidden layer and output layer nodes and all connection weights become part of learning process
  - Node-specific error for output node,

$$err = correction\ factor \times (actual\ value - predicted\ value)$$
$$\theta_{new} = \theta_{old} + learning\ rate \times err$$
$$w_{new} = wold + learning\ rate \times err$$

  - Learning rate controls the rate of change from previous iteration
    - Value is typically a constant in the range [0,1]

ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - Node-specific error for hidden nodes
    - Based on *err* value of output node instead of prediction error
    - Steps are same as those used for output node
- Methods for updating weight and bias values
  - Case updating
    - Updating is done after each case or record is run through the network (referred as a trial)
    - When all the records are run through the network, it is referred as ***one epoch, or sweep through the data***
    - Many epochs could be used to train the network

ARTIFICIAL NEURAL NETWORKS

- Methods for updating weight and bias values
  - Batch updating
    - Updating is done after all the records are run through the network
    - In place of prediction error of the record, sum of prediction errors for all records is used
    - Many epochs could be used to train the network
- Case updating vs. batch updating
  - Case updating yields more accurate results
    - With a longer runtime

ARTIFICIAL NEURAL NETWORKS

- Stopping Criteria for updating
  - Small incremental change in bias and weight values from previous iteration
  - Rate of change of error function values reaches a required threshold
  - Limit on no. of runs is reached

- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

ARTIFICIAL NEURAL NETWORK PART-4

LECTURE 56

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Scale of [0,1] is typically recommended for neural network models for performance purposes
  - For numeric variables,

$$V_{norm} = \frac{V - \min(V)}{\max(V) - \min(V)}$$

ARTIFICIAL NEURAL NETWORKS

- Normalization
  - Binary variables (categorical variables with two classes)
    - Create dummy variables: set of values {0, 1}
  - Nominal variables with m (>2) classes
    - Create m-1 dummy variables: set of values {0, 1}
  - Ordinal variables with m (>2) classes
    - Map the values to the set {0, 1/(m-1), 2/(m-1), …, (m-2)/(m-1), 1}

ARTIFICIAL NEURAL NETWORKS

- Other transformations
  - Transformations which could spread the values more symmetrically can be done for performance purposes
    - Log transform of a right-skewed variable

- Estimation method
  - Least squares and maximum likelihood methods use a global metric of errors (e.g., SSE) to estimate the parameters

ARTIFICIAL NEURAL NETWORKS

- Estimation method
  - Neural networks use error values of each observation to update the parameters in an iterative fashion (referred as learning)
    - Error for the output node (prediction error) is distributed across all the hidden layer nodes
    - All hidden layer nodes share responsibility for part of the error (referred as node-specific error)
    - Node-specific errors are used to update the connection weights and bias values

ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - An algorithm to update weights and bias values of a neural network
  - Error values are computed from output layer back to hidden layers
    - All hidden layer and output layer nodes and all connection weights become part of learning process
  - Node-specific error for output node,

$$err = correction\ factor \times (actual\ value - predicted\ value)$$
$$\theta_{new} = \theta_{old} + learning\ rate \times err$$
$$w_{new} = wold + learning\ rate \times err$$

  - Learning rate controls the rate of change from previous iteration
    - Value is typically a constant in the range [0,1]

ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - Node-specific error for hidden nodes
    - Based on *err* value of output node instead of prediction error
    - Steps are same as those used for output node
- Methods for updating weight and bias values
  - Case updating
    - Updating is done after each case or record is run through the network (referred as a trial)
    - When all the records are run through the network, it is referred as ***one epoch, or sweep through the data***
    - Many epochs could be used to train the network

ARTIFICIAL NEURAL NETWORKS

- Methods for updating weight and bias values
  - Batch updating
    - Updating is done after all the records are run through the network
    - In place of prediction error of the record, sum of prediction errors for all records is used
    - Many epochs could be used to train the network
- Case updating vs. batch updating
  - Case updating yields more accurate results
    - With a longer runtime

ARTIFICIAL NEURAL NETWORKS

- Stopping Criteria for updating
  - Small incremental change in bias and weight values from previous iteration
  - Rate of change of error function values reaches a required threshold
  - Limit on no. of runs is reached

- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

ARTIFICIAL NEURAL NETWORK PART-5

LECTURE 57

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

ARTIFICIAL NEURAL NETWORKS

- Estimation method
  - Neural networks use error values of each observation to update the parameters in an iterative fashion (referred as learning)
    - Error for the output node (prediction error) is distributed across all the hidden layer nodes
    - All hidden layer nodes share responsibility for part of the error (referred as node-specific error)
    - Node-specific errors are used to update the connection weights and bias values

ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - An algorithm to update weights and bias values of a neural network
  - Error values are computed from output layer back to hidden layers
    - All hidden layer and output layer nodes and all connection weights become part of learning process
  - Node-specific error for output node,

$$err = correction\ factor \times (actual\ value - predicted\ value)$$
$$\theta_{new} = \theta_{old} + learning\ rate \times err$$
$$w_{new} = wold + learning\ rate \times err$$

  - Learning rate controls the rate of change from previous iteration
    - Value is typically a constant in the range [0,1]

ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - Node-specific error for hidden nodes
    - Based on *err* value of output node instead of prediction error
    - Steps are same as those used for output node
- Methods for updating weight and bias values
  - Case updating
    - Updating is done after each case or record is run through the network (referred as a trial)
    - When all the records are run through the network, it is referred as ***one epoch, or sweep through the data***
    - Many epochs could be used to train the network

ARTIFICIAL NEURAL NETWORKS

- Methods for updating weight and bias values
  - Batch updating
    - Updating is done after all the records are run through the network
    - In place of prediction error of the record, sum of prediction errors for all records is used
    - Many epochs could be used to train the network
- Case updating vs. batch updating
  - Case updating yields more accurate results
    - With a longer runtime

ARTIFICIAL NEURAL NETWORKS

- Stopping Criteria for updating
  - Small incremental change in bias and weight values from previous iteration
  - Rate of change of error function values reaches a required threshold
  - Limit on no. of runs is reached

- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

ARTIFICIAL NEURAL NETWORK PART-6

LECTURE 58

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

ARTIFICIAL NEURAL NETWORKS

- Estimation method
  - Neural networks use error values of each observation to update the parameters in an iterative fashion (referred as learning)
    - Error for the output node (prediction error) is distributed across all the hidden layer nodes
    - All hidden layer nodes share responsibility for part of the error (referred as node-specific error)
    - Node-specific errors are used to update the connection weights and bias values

ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - An algorithm to update weights and bias values of a neural network
  - Error values are computed from output layer back to hidden layers
    - All hidden layer and output layer nodes and all connection weights become part of learning process
  - Node-specific error for output node,

$$err = correction\ factor \times (actual\ value - predicted\ value)$$
$$\theta_{new} = \theta_{old} + learning\ rate \times err$$
$$w_{new} = wold + learning\ rate \times err$$

  - Learning rate controls the rate of change from previous iteration
    - Value is typically a constant in the range [0,1]

ARTIFICIAL NEURAL NETWORKS

- Back Propagation
  - Node-specific error for hidden nodes
    - Based on *err* value of output node instead of prediction error
    - Steps are same as those used for output node
- Methods for updating weight and bias values
  - Case updating
    - Updating is done after each case or record is run through the network (referred as a trial)
    - When all the records are run through the network, it is referred as ***one epoch, or sweep through the data***
    - Many epochs could be used to train the network

ARTIFICIAL NEURAL NETWORKS

- Methods for updating weight and bias values
  - Batch updating
    - Updating is done after all the records are run through the network
    - In place of prediction error of the record, sum of prediction errors for all records is used
    - Many epochs could be used to train the network
- Case updating vs. batch updating
  - Case updating yields more accurate results
    - With a longer runtime

ARTIFICIAL NEURAL NETWORKS

- Stopping Criteria for updating
  - Small incremental change in bias and weight values from previous iteration
  - Rate of change of error function values reaches a required threshold
  - Limit on no. of runs is reached

- Open RStudio

ARTIFICIAL NEURAL NETWORKS

- A complete modeling is discussed in the lecture video based on this topic using data of used cars record

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Discriminant Analysis

LECTURE 59

DR. GAURAV DIXIT

DEPARTMENT OF MANAGEMENT STUDIES

Discriminant Analysis

- Statistical technique
  - Used for classification and profiling tasks
  - Model-based approach
  - Idea is
    - To find a separating line or hyperplane equidistant from centroids of different classes

      Or
    - To find a separating line or hyperplane that is best at discriminating the records into different classes
  - Classification procedure is based on distance based metrics
    - Based on the distance of a record from each class

Discriminant Analysis

- Classification
  - Best separation between items is found by measuring their distance from each class
  - An item is classified to the closest class
- Euclidean distance metric
  - Distance of a record $(x_1, \ldots, x_p)$ from centroid $(\bar{x}_1, \ldots, \bar{x}_p)$ of a class is computed

$$D_{eu}(x, \bar{x}) = \sqrt{(x_1 - \bar{x}_1)^2 + \cdots + (x_p - \bar{x}_p)^2}$$

Where centroid $\bar{x}$ is a vector of means of p predictors

Discriminant Analysis

- Issues with Euclidean distance metric
  - Distance values depend on the unit of a measurement
  - Based on mean and doesn't account for variance
    - Variability plays an important role in determining the closeness of a record to a particular class
  - Distance should be computed using std. dev. (z-scores) instead of unit of measurement
  - Correlation between variables is ignored

Discriminant Analysis

- "Statistical distance" (or Mahalanobis distance) can be used to overcome issues with Euclidean distance metric

$$D_{ml}(x, \bar{x}) = [x - \bar{x}]' \, S^{-1} \, [x - \bar{x}]$$

Where $[x - \bar{x}]'$ is transpose matrix of $[x - \bar{x}]$

- Column vectors are turned into row vectors

and S^{-1} is inverse matrix of S (covariance matrix between p predictors)

- Can be considered as p-dimensional extension of division operation

Discriminant Analysis

- Linear Classification Functions
  - Used as basis for separation of records into classes
    - Compute classification score measuring closeness of a record to each class
    - Highest classification score is equivalent of smallest statistical distance
  - Main idea is
    - To find linear functions of predictors that maximize ratio of between-class variability to within-class variability
- Open RStudio

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks...

IIT ROORKEE

NPTEL ONLINE CERTIFICATION COURSE

Discriminant Analysis Part-2

LECTURE 60

DR. GAURAV DIXIT
DEPARTMENT OF MANAGEMENT STUDIES

Discriminant Analysis

- Statistical technique
  - Used for classification and profiling tasks
  - Model-based approach
  - Idea is
    - To find a separating line or hyperplane equidistant from centroids of different classes

      Or

    - To find a separating line or hyperplane that is best at discriminating the records into different classes
  - Classification procedure is based on distance based metrics
    - Based on the distance of a record from each class

Discriminant Analysis

- Classification
  - Best separation between items is found by measuring their distance from each class
  - An item is classified to the closest class
- Euclidean distance metric
  - Distance of a record $(x_1, \ldots, x_p)$ from centroid $(\bar{x}_1, \ldots, \bar{x}_p)$ of a class is computed

$$D_{eu}(x, \bar{x}) = \sqrt{(x_1 - \bar{x}_1)^2 + \cdots + (x_p - \bar{x}_p)^2}$$

Where centroid $\bar{x}$ is a vector of means of p predictors

Discriminant Analysis

- Issues with Euclidean distance metric
  - Distance values depend on the unit of a measurement
  - Based on mean and doesn’t account for variance
    - Variability plays an important role in determining the closeness of a record to a particular class
  - Distance should be computed using std. dev. (z-scores) instead of unit of measurement
  - Correlation between variables is ignored

Discriminant Analysis

- "Statistical distance" (or Mahalanobis distance) can be used to overcome issues with Euclidean distance metric

$$D_{ml}(x, \bar{x}) = [x - \bar{x}]' \, S^{-1} \, [x - \bar{x}]$$

Where $[x - \bar{x}]'$ is transpose matrix of $[x - \bar{x}]$

- Column vectors are turned into row vectors

and S^{-1} is inverse matrix of S (covariance matrix between p predictors)

- Can be considered as p-dimensional extension of division operation

Discriminant Analysis

- Linear Classification Functions
  - Used as basis for separation of records into classes
    - Compute classification score measuring closeness of a record to each class
    - Highest classification score is equivalent of smallest statistical distance
  - Main idea is
    - To find linear functions of predictors that maximize ratio of between-class variability to within-class variability
- Open RStudio

Discriminant Analysis

- Assumptions and other issues
  - Predictors follow multivariate normal distribution for all classes
    - Given adequate sample points for all classes, relatively robust to violations of normality assumption
  - Correlation structure between predictors for each class should be same
  - Sensitive to outliers

Discriminant Analysis

- Further Comments on discriminant analysis
  - Application and performance aspects are similar to multiple linear regression
  - In discriminant analysis, coefficients of linear discriminant are optimized w.r.t class separation
    - In linear regression, coefficients are optimized w.r.t outcome variable
  - Estimation technique is least squares
    - Same as linear regression

Key References

- Data Science and Big Data Analytics: Discovering, Analyzing, Visualizing and Presenting Data by EMC Education Services (2015)
- Data Mining for Business Intelligence: Concepts, Techniques, and Applications in Microsoft Office Excel with XLMiner by Shmueli, G., Patel, N. R., & Bruce, P. C. (2010)

Thanks…