UNIT – 2

Multivariate Classification and Logistic Regression

The goal of classification is to assign each cell in a study area to a class or category. Examples of a class or category include land-use type, locations preferred by bears, and avalanche potential.

There are two types of classification: supervised and unsupervised. In a supervised classification, you have a sampling of the features. For example, you know that there is a coniferous forest in the northwest region of your study area, so you identify it by enclosing it on the map with a polygon (or with multiple polygons). Another polygon is created to encompass a wheat field, another for urban buildings, and another for water. You continue this process until you have enough features to represent a class, and all classes in your data are identified. Each grouping of features is considered a class, and the polygon that encompasses the class is a training sample. Once you have identified your training samples, multivariate statistics are calculated on them to establish the relationships within and between the classes. The statistics are stored in a signature file.

In an unsupervised classification, you do not know what features are actually at any specified location, but you want to aggregate each of the locations into one of a specified number of groups or clusters. What determines to which class or cluster each location will be assigned is dependent on the multivariate statistics that are calculated on the input bands. Each cluster is statistically separate from the other clusters based on the values for each band of each cell within the clusters. The statistics establishing the cluster definition are stored in a signature file.

There are four steps in performing a classification:

1. Create and analyze the input data.
2. Produce signatures for class and cluster analysis.
3. Evaluate and, if necessary, edit classes and clusters.

4. Perform the classification.

   There are two input types to the classification: the input raster bands to analyze, and the classes or clusters into which to fit the locations. The input raster bands used in the multivariate analysis need to influence or be an underlying cause in the categorization of the classification. That is, slope, snow depth, and solar radiation can be factors that influence avalanche potential, while soil type may have no effect.

   A class corresponds to a meaningful grouping of locations. Examples of classes include forests, water bodies, fields, and residential areas. Classes derived from clusters include deer preference or erosion potential.

   Each location is characterized by a set or vector of values, one value for each variable, or band entered in the analysis. Each location can be visualized as a point in a multidimensional attribute space whose axes correspond to the variables represented by each input band. A class or cluster is a grouping of points in this multidimensional attribute space. Two locations belong to the same class or cluster if their attributes (vector of band values) are similar. A multiband raster and individual single band rasters can be used as the input into a multivariate statistical analysis.

   Locations corresponding to known classes may form clusters in attribute space if the classes can be separated, or distinguished, by the attribute values. Locations corresponding to natural clusters in attribute space can be interpreted as naturally occurring classes of strata.

   The goal of classification is to assign each cell in a study area to a class or category. Examples of a class or category include land-use type, locations preferred by bears, and avalanche potential.

   There are two types of classification: supervised and unsupervised. In a supervised classification, you have a sampling of the features. For example, you know that there is a coniferous forest in the northwest region of your study area, so you identify it by enclosing it on the map with a polygon (or with multiple polygons). Another polygon is created to encompass a wheat field, another for urban buildings, and another for water. You continue this process until you have enough features to represent a class, and all classes in your data are identified. Each grouping of features is considered a class, and the polygon that

encompasses the class is a training sample. Once you have identified your training samples, multivariate statistics are calculated on them to establish the relationships within and between the classes. The statistics are stored in a signature file.

In an unsupervised classification, you do not know what features are actually at any specified location, but you want to aggregate each of the locations into one of a specified number of groups or clusters. What determines to which class or cluster each location will be assigned is dependent on the multivariate statistics that are calculated on the input bands. Each cluster is statistically separate from the other clusters based on the values for each band of each cell within the clusters. The statistics establishing the cluster definition are stored in a signature file.

There are four steps in performing a classification:

1. Create and analyze the input data.
2. Produce signatures for class and cluster analysis.
3. Evaluate and, if necessary, edit classes and clusters.
4. Perform the classification.

There are two input types to the classification: the input raster bands to analyze, and the classes or clusters into which to fit the locations. The input raster bands used in the multivariate analysis need to influence or be an underlying cause in the categorization of the classification. That is, slope, snow depth, and solar radiation can be factors that influence avalanche potential, while soil type may have no effect.

A class corresponds to a meaningful grouping of locations. Examples of classes include forests, water bodies, fields, and residential areas. Classes derived from clusters include deer preference or erosion potential.

Each location is characterized by a set or vector of values, one value for each variable, or band entered in the analysis. Each location can be visualized as a point in a multidimensional attribute space whose axes correspond to the variables represented by each input band. A class or cluster is a grouping of points in this multidimensional attribute space. Two locations belong to the same class or cluster if their attributes (vector of band values) are similar. A

multiband raster and individual single band rasters can be used as the input into a multivariate statistical analysis.

Locations corresponding to known classes may form clusters in attribute space if the classes can be separated, or distinguished, by the attribute values. Locations corresponding to natural clusters in attribute space can be interpreted as naturally occurring classes of strata.

Logistic Regression

Logistic Regression is also known as Logit, Maximum-Entropy classifier is a supervised learning method for classification. It establishes a relation between dependent class variables and independent variables using regression.

The dependent variable is categorical i.e. it can take only integral values representing different classes. The probabilities describing the possible outcomes of a query point are modelled using a logistic function. This model belongs to a family of discriminative classifiers. They rely on attributes which discriminate the classes well. This model is used when we have 2 classes of dependent variables. When there are more than 2 classes, then we have another regression method which helps us to predict the target variable better.

There are two broad categories of Logistic Regression algorithms

1. Binary Logistic Regression when the dependent variable is strictly binary
2. Multinomial Logistic Regression when the dependent variable has multiple categories.

There are two types of Multinomial Logistic Regression

1. Ordered Multinomial Logistic Regression (dependent variable has ordered values)
2. Nominal Multinomial Logistic Regression (dependent variable has unordered categories)

Process Methodology:

Logistic regression takes into consideration the different classes of dependent variables and assigns probabilities to the event happening for each row of information. These probabilities are found by assigning different weights to each independent variable by understanding the relationship between the variables. If the correlation between the variables is high, then positive weights are assigned and in the case of an inverse relationship, negative weight is assigned.

As the model is mainly used to classify the classes of target variables as either 0 or 1, thus the Sigmoid function is obtained by implementing the log-normal function on these probabilities that are calculated on these independent variables.

The Sigmoid function:

P(y= 1) = Sigmoid(Z) = 1/(1 + e -z)

P(y= 0) = 1 –P(y =1) = 1 –(1/(1 + e -z)) = e –z/ (1 + e -z)

y = 1 if P(y=1|X) > .5, else y = 0

where the default probability cut off is taken as 0.5.

$$logLoss = \frac{-1}{N}\sum_{i=1}^{N}(y_i(logp_i) + (1 - y_i)log(1 - p_i))$$

This method is also called the Odds Log ratio.

Assumptions:

1. The dependent variable is categorical. Dichotomous for binary logistic regression and multi-label for multi-class classification
2. Attributes and log odds i.e. log(p / 1-p) should be linearly related to the independent variables
3. Attributes are independent of each other (low or no multicollinearity)
4. In binary logistic regression class of interest is coded with 1 and other class 0
5. In multi-class classification using Multinomial Logistic Regression or OVR scheme, class of interest is coded 1 and rest 0(this is done by the algorithm)

Note: the assumptions of Linear Regression such as homoscedasticity, normal distribution of error terms, a linear relationship between the dependent and independent variables are not required here.

Some examples where this model can be used for predictions.

1. **Predicting the weather:** you can only have a few definite weather types. Stormy, sunny, cloudy, rainy and a few more.
2. **Medical diagnosis:** given the symptoms predicted the disease patient is suffering from.

3. **Credit Default:** If a loan has to be given a particular candidate depend on his identity check, account summary, any properties he holds, any previous loan, etc

4. **HR Analytics:** IT firms recruit a large number of people, but one of the problems they encounter is after accepting the job offer many candidates do not join. So, this results in cost overruns because they have to repeat the entire process again. Now when you get an application, can you actually predict whether that applicant is likely to join the organization (Binary Outcome – Join / Not Join).

5. **Elections:** Suppose that we are interested in the factors that influence whether a political candidate wins an election. The outcome (response) variable is binary (0/1); win or lose. The predictor variables of interest are the amount of money spent on the campaign and the amount of time spent campaigning negatively.

Reference:

1. https://pro.arcgis.com/en/pro-app/latest/tool-reference/spatial-analyst/understanding-multivariate-classification.htm#:~:text=Each%20grouping%20of%20features%20is,stored%20in%20a%20signature%20file.