PRACTICAL MACHINE LEARNING

(A case study demonstrating the applications of supervised learning – logistic regression, fisher's linear discriminant analysis, k-nearest neighbors, artificial neural network, and unsupervised learning – k-means clustering for some botanical data.)

Submitted by: Vishal Kumar

Due Date: 19th April, 2016

Date of Submission: 15th April, 2016

Instructor's Remarks:

5. PRACTICAL MACHINE LEARNING (5 CASES)

5.1. LEARNING A 3-CLASS FLOWER SPECIES CLASSIFIER BASED ON SOME ATTRIBUTES OF A FLOWER USING LOGISTIC REGRESSION

5.2. LEARNING A 3-CLASS FLOWER SPECIES CLASSIFIER BASED ON SOME ATTRIBUTES OF A FLOWER USING FISHER'S LINEAR DISCRIMINANT ANALYSIS

5.3. LEARNING A 3-CLASS FLOWER SPECIES CLASSIFIER BASED ON SOME ATTRIBUTES OF A FLOWER USING K-NEAREST NEIGHBORS

5.4. LEARNING A 3-CLASS FLOWER SPECIES CLASSIFIER BASED ON SOME ATTRIBUTES OF A FLOWER USING ARTIFICIAL NEURAL NETWORK

5.5. CATEGORIZING FLOWERS INTO 3 CATEGORIES BASED ON SOME ATTRIBUTES OF A FLOWER USING K-MEANS CLUSTERING

Practical Machine Learning

Machine learning is a scientific discipline that explores the construction and study of algorithms that can learn from data. Such algorithms operate by building a model from example inputs and using that to make predictions or decisions, rather than following strictly static program instructions. Machine learning is closely related to and often overlaps with computational statistics: a discipline that also specializes in prediction-making.

The **regression** problem which we already know is called as a **classification** problem if the response is a discrete variable. In simpler words we want to classify an observation (univariate or multivariate) into one of several possible classes, or simply we want to estimate the probability given an observation that it belongs to one of the several possible classes.

Due to the fact that a discrete variable can't be normally distributed the application of linear regression becomes invalid as the assumption of normality of the observations is no longer satisfied. Indeed the response follows a multinomial distribution. Moreover, the expectation of response becomes uninterpretable in terms of a linear function of the features and also the variances do not remain constant across observations and hence causing heteroscedasticity. Hence such problems are outside the ambit of linear regression. There are several tools namely **logistic classifier**, **discriminant analysis**, **nearest neighbor** approach, **neural network** etc. available to be deployed in such situations. The classification problems are quite naturally divided into two types:

Binary, where response has two possible classes meaning by an observation either belongs to a class or it doesn't, and **Multiclass**, where the response has more than two possible classes.

In this case study, we have considered the famous (Fisher's or Anderson's) iris data set which gives the measurements in centimeters of the variables sepal length and width and petal length and width, respectively, for 50 flowers from each of 3 species of iris. The species are Iris "***setosa***", "***versicolor***", and "***virginica***". Thus, we will conduct Multiclass Classification Techniques.

Definitions:

A **Confusion Matrix**, also known as a contingency table or an crosstabs, is a specific table layout that allows visualization of the performance of an algorithm, typically a **supervised** learning one (in **unsupervised** learning it is usually called a matching matrix). Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. The name stems from the fact that it makes it easy to see if the system is confusing two classes (i.e. commonly mislabeling one as another).

$$\text{Percentage of Misclassification } = \frac{No.of\ misclassified\ predicted\ group}{Total\ no.of\ observations} \times 100$$

Case 1: Consider the dataset "flower species.xlsx" which has 150 observations on four features namely sepal length, sepal width, petal length and petal width for three different species of flowers (50 observations for each species). Learn a 3class classification hypothesis on given data, which predicts the species type based on the given four features using **Logistic Regression.**

Obtain the estimated classes for all the observations in the given data. Construct the observed vs. predicted classification table and calculate the class wise and overall percentage of misclassification.

Solution: Multinomial logistic regression is a classification method that generalizes logistic regression to multiclass problems, i.e. with more than two possible discrete outcomes. That is, it is a model that is used to predict the probabilities of the different possible outcomes of a categorically distributed dependent variable meaning that it falls into any one of a set of categories which cannot be ordered in any meaningful way, given a set of independent variables.

Multinomial logistic regression is a particular solution to the classification problem that assumes that a linear combination of the observed features and some problem-specific parameters can be used to determine the probability of each particular outcome of the dependent variable. The best values of the parameters for a given problem are usually determined from some training data.

After predicting classes for all the observations in the given data using **Multinomial Logistic Regression**, we construct the Observed vs. Predicted Classification Table as follows:

Classification

	Predicted			
Observed	1	2	3	Percent Correct
1	50	0	0	100.0%
2	0	49	1	98.0%
3	0	1	49	98.0%
Overall Percentage	33.3%	33.3%	33.3%	98.7%

From the above Cross-table we make following **conclusions**:

Case wise Percentage of Misclassification:

$$\text{Setosa} = \frac{0}{50} \times 100 = 0\%$$

$$\text{Versicolor} = \frac{1}{50} \times 100 = 2\%$$

$$\text{Virginica} = \frac{1}{50} \times 100 = 2\%$$

$$\textbf{Overall Percentage of Misclassification} = \frac{2}{150} \times 100 = 1.33\%$$

Case 2: Consider the dataset "flower species.xlsx" which has 150 observations on four features namely sepal length, sepal width, petal length and petal width for three different species of flowers (50 observations for each species). Learn a 3class classification hypothesis on given data, which predicts the species type based on the given four features using **Fisher's Discriminant Analysis (FDA)**.

Obtain the estimated classes for all the observations in the given data. Construct the observed vs. predicted classification table and calculate the class wise and overall percentage of misclassification.

Solution: Discriminant Analysis is a classification technique to classify an observation (univariate or multivariate) into one of several possible classes by means of "some" optimal way to separate different populations. Some real life examples of classification problem are loan classification – high risk, medium risk and low risk, warning systems for financial crisis, medical diagnostics – critical and non-critical patients.

Linear discriminant analysis is a generalization of Fisher's linear discriminant, a method used in statistics, pattern recognition and machine learning to find a linear combination of features that characterizes or separates two or more classes of objects or events.

The main purpose of a discriminant function analysis is to predict group membership based on a linear combination of the interval variables. The procedure begins with a set of observations where both group membership and the values of the interval variables are known. The end result of the procedure is a model that allows prediction of group membership when only the interval variables are known.

After predicting classes for all the observations in the given data using **Fisher's Discriminant Analysis**, we construct the Observed vs. Predicted Classification Table as follows:

Species_new * Predicted Group for Analysis 1 Crosstabulation

Count

		Predicted Group for Analysis 1			
		1	2	3	Total
Species_new	1	50	0	0	50
	2	0	48	2	50
	3	0	1	49	50
Total		50	49	51	150

From the above Cross-table we make following **conclusions**:

Case wise Percentage of Misclassification:

$$\text{Setosa} = \frac{0}{50} \times 100 = 0\%$$

$$\text{Versicolor} = \frac{2}{50} \times 100 = 4\%$$

$$\text{Virginica} = \frac{1}{50} \times 100 = 2\%$$

$$\textbf{Overall Percentage of Misclassification} = \frac{3}{150} \times 100 = 2\%$$

Case 3: Consider the dataset "flower species.xlsx" which has 150 observations on four features namely sepal length, sepal width, petal length and petal width for three different species of flowers (50 observations for each species). Learn a 3class classification hypothesis on given data, which predicts the species type based on the given four features using **K-Nearest Neighbors (K-NN).**

Obtain the estimated classes for all the observations in the given data. Construct the observed vs. predicted classification table and calculate the class wise and overall percentage of misclassification.

Solution: k- Nearest Neighbor (kNN) classification is one of the most fundamental and simple classification methods and should be one of the first choices for a classification study when there is little or no prior knowledge about the distribution of the data. K-nearest-neighbor classification was developed from the need to perform discriminant analysis when reliable parametric estimates of probability densities are unknown or difficult to determine. In pattern recognition, the k-Nearest Neighbors algorithm is a non-parametric method used for classification.

In k-NN classification, the output is a class membership. An object is classified by a majority vote of its neighbors, with the object being assigned to the class most common among its k nearest neighbors (k is a positive integer, typically small). If k = 1, then the object is simply assigned to the class of that single nearest neighbor.

After predicting classes for all the observations in the given data using **K-Nearest Neighbors**, we construct the Observed vs. Predicted Classification Table as follows:

Species_new * Predicted Value for Species_new Crosstabulation

Count

		Predicted Value for Species_new			
		1	2	3	Total
Species_new	1	50	0	0	50
	2	0	49	1	50
	3	0	4	46	50
Total		50	53	47	150

From the above Cross-table we make following **conclusions**:

Case wise Percentage of Misclassification:

$$\text{Setosa} = \frac{0}{50} \times 100 = 0\%$$

$$\text{Versicolor} = \frac{1}{50} \times 100 = 2\%$$

$$\text{Virginica} = \frac{4}{50} \times 100 = 8\%$$

$$\textbf{Overall Percentage of Misclassification} = \frac{5}{150} \times 100 = 3.33\%$$

Case 4: Consider the dataset "flower species.xlsx" which has 150 observations on four features namely sepal length, sepal width, petal length and petal width for three different species of flowers (50 observations for each species). Learn a 3class classification hypothesis on given data, which predicts the species type based on the given four features using **Artificial Neural Network (ANN).**

Obtain the estimated classes for all the observations in the given data. Construct the observed vs. predicted classification table and calculate the class wise and overall percentage of misclassification.

Solution: Artificial neural networks (ANNs) are a family of statistical learning algorithms inspired by biological neural networks and are used to estimate or approximate functions that can depend on a large number of inputs and are generally unknown. Artificial neural networks are generally presented as systems of interconnected "neurons" which can compute values from inputs, and are capable of machine learning as well as pattern recognition.

ANN is a computing system made up of a number of simple, highly interconnected processing elements, which process information by their dynamic state response to external inputs.

Neural networks are typically organized in layers. Layers are made up of a number of interconnected 'nodes' which contain an 'activation function'. Patterns are presented to the network via the 'input layer', which communicates to one or more 'hidden layers' where the actual processing is done via a system of weighted 'connections'. The hidden layers then link to an 'output layer' where the answer is output.

After predicting classes for all the observations in the given data using **Artificial Neural Network**, we construct the Observed vs. Predicted Classification Table as follows:

Classification

		Predicted			
Sample	Observed	1	2	3	Percent Correct
Training	1	50	0	0	100.0%
	2	0	49	1	98.0%
	3	0	1	49	98.0%
	Overall Percent	33.3%	33.3%	33.3%	98.7%

Dependent Variable: Species_new

From the above Cross-table we make following **conclusions**:

Case wise Percentage of Misclassification:

$$\text{Setosa} = \frac{0}{50} \times 100 = 0\%$$

$$\text{Versicolor} = \frac{1}{50} \times 100 = 2\%$$

$$\text{Virginica} = \frac{1}{50} \times 100 = 2\%$$

$$\textbf{Overall Percentage of Misclassification} = \frac{2}{150} \times 100 = 1.33\%$$

Case 5: Consider the dataset "flower species.xlsx" which has 150 observations on four features namely sepal length, sepal width, petal length and petal width for three different species of flowers (50 observations for each species). Assuming that true species is not known for the given data, classify the data in 3 clusters using **K-Means Clustering.**

Obtain the estimated classes for all the observations in the given data. Construct the observed vs. predicted classification table and calculate the class wise and overall percentage` of misclassification

Solution: k-Means Clustering is one of the simplest unsupervised learning algorithms that solve the well-known clustering problem. The procedure follows a simple and easy way to classify a given data set through a certain number of clusters (assume k clusters) fixed a priori. The main idea is to define k centroids, one for each cluster. These centroids should be placed in a cunning way because of different location causes different result. So, the better choice is to place them as much as possible far away from each other.

The next step is to take each point belonging to a given data set and associate it to the nearest centroid. When no point is pending, the first step is completed and an early grouping is done. At this point we need to re-calculate k new centroids as centers of the clusters resulting from the previous step. After we have these k new centroids, a new binding has to be done between the same data set points and the nearest new centroid.

A loop has been generated. As a result of this loop we may notice that the k centroids change their location step by step until no more changes are done. In other words centroids do not move any more.

After predicting classes for all the observations in the given data using **K-Means Clustering**, we construct the Observed vs. Predicted Classification Table as follows:

Species_new * Cluster Number of Case Crosstabulation

Count

		Cluster Number of Case			
		1	2	3	Total
Species_new	1	0	50	0	50
	2	2	0	48	50
	3	36	0	14	50
Total		38	50	62	150

From the above Cross-table we make following **conclusions**:

Case wise Percentage of Misclassification:

$$\text{Setosa} = \frac{0}{50} \times 100 = 0\%$$

$$\text{Versicolor} = \frac{2}{50} \times 100 = 4\%$$

$$\text{Virginica} = \frac{14}{50} \times 100 = 28\%$$

$$\textbf{Overall Percentage of Misclassification} = \frac{16}{150} \times 100 = 10.66\%$$

From the above Cross-tables we make following deductions:

We calculate how many misclassifications were made by different methods of classification.

	Multinomial Logistic	Discriminant	KNN	Neural Network	Cluster
Setosa	0%	0%	0%	0%	0%
Versicolor	2%	4%	2%	2%	4%
Virginica	2%	2%	8%	2%	28%
Overall Misclassification	**1.33%**	**2%**	**3.33%**	**1.33%**	**10.66%**

Conclusion: We observe for the given "iris" data maximum misclassification were made by cluster analysis and maximum correct classifications were made by Multinomial logistic regression and Neural Network. As Neural Network is an advance algorithm to implement we will stick to **Multinomial Logistic Regression** to predict flower species for the given data set.

The Observed data (Species, given below) and Predicted Class category using the above discussed methods are given as below, **misclassifications are highlighted** in the table:

S.no.	Species	Multinomial Logistic	Discriminant	KNN	Neural Network	Cluster
1	setosa	setosa	setosa	setosa	setosa	setosa
2	setosa	setosa	setosa	setosa	setosa	setosa
3	setosa	setosa	setosa	setosa	setosa	setosa
4	setosa	setosa	setosa	setosa	setosa	setosa
5	setosa	setosa	setosa	setosa	setosa	setosa
6	setosa	setosa	setosa	setosa	setosa	setosa
7	setosa	setosa	setosa	setosa	setosa	setosa
8	setosa	setosa	setosa	setosa	setosa	setosa
9	setosa	setosa	setosa	setosa	setosa	setosa
10	setosa	setosa	setosa	setosa	setosa	setosa
11	setosa	setosa	setosa	setosa	setosa	setosa
12	setosa	setosa	setosa	setosa	setosa	setosa
13	setosa	setosa	setosa	setosa	setosa	setosa
14	setosa	setosa	setosa	setosa	setosa	setosa
15	setosa	setosa	setosa	setosa	setosa	setosa
16	setosa	setosa	setosa	setosa	setosa	setosa
17	setosa	setosa	setosa	setosa	setosa	setosa
18	setosa	setosa	setosa	setosa	setosa	setosa
19	setosa	setosa	setosa	setosa	setosa	setosa
20	setosa	setosa	setosa	setosa	setosa	setosa
21	setosa	setosa	setosa	setosa	setosa	setosa
22	setosa	setosa	setosa	setosa	setosa	setosa
23	setosa	setosa	setosa	setosa	setosa	setosa
24	setosa	setosa	setosa	setosa	setosa	setosa
25	setosa	setosa	setosa	setosa	setosa	setosa
26	setosa	setosa	setosa	setosa	setosa	setosa
27	setosa	setosa	setosa	setosa	setosa	setosa
28	setosa	setosa	setosa	setosa	setosa	setosa
29	setosa	setosa	setosa	setosa	setosa	setosa
30	setosa	setosa	setosa	setosa	setosa	setosa

31	setosa	setosa	setosa	setosa	setosa	setosa
32	setosa	setosa	setosa	setosa	setosa	setosa
33	setosa	setosa	setosa	setosa	setosa	setosa
34	setosa	setosa	setosa	setosa	setosa	setosa
35	setosa	setosa	setosa	setosa	setosa	setosa
36	setosa	setosa	setosa	setosa	setosa	setosa
37	setosa	setosa	setosa	setosa	setosa	setosa
38	setosa	setosa	setosa	setosa	setosa	setosa
39	setosa	setosa	setosa	setosa	setosa	setosa
40	setosa	setosa	setosa	setosa	setosa	setosa
41	setosa	setosa	setosa	setosa	setosa	setosa
42	setosa	setosa	setosa	setosa	setosa	setosa
43	setosa	setosa	setosa	setosa	setosa	setosa
44	setosa	setosa	setosa	setosa	setosa	setosa
45	setosa	setosa	setosa	setosa	setosa	setosa
46	setosa	setosa	setosa	setosa	setosa	setosa
47	setosa	setosa	setosa	setosa	setosa	setosa
48	setosa	setosa	setosa	setosa	setosa	setosa
49	setosa	setosa	setosa	setosa	setosa	setosa
50	setosa	setosa	setosa	setosa	setosa	setosa
51	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
52	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
53	versicolor	versicolor	versicolor	versicolor	versicolor	virginica
54	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
55	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
56	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
57	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
58	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
59	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
60	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
61	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
62	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
63	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
64	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
65	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
66	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
67	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
68	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
69	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
70	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
71	versicolor	versicolor	virginica	versicolor	versicolor	versicolor
72	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
73	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
74	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
75	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
76	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
77	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
78	versicolor	versicolor	versicolor	versicolor	versicolor	virginica
79	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor

80	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
81	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
82	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
83	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
84	versicolor	virginica	virginica	virginica	virginica	versicolor
85	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
86	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
87	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
88	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
89	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
90	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
91	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
92	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
93	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
94	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
95	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
96	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
97	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
98	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
99	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
100	versicolor	versicolor	versicolor	versicolor	versicolor	versicolor
101	virginica	virginica	virginica	virginica	virginica	virginica
102	virginica	virginica	virginica	virginica	virginica	versicolor
103	virginica	virginica	virginica	virginica	virginica	virginica
104	virginica	virginica	virginica	virginica	virginica	virginica
105	virginica	virginica	virginica	virginica	virginica	virginica
106	virginica	virginica	virginica	virginica	virginica	virginica
107	virginica	virginica	virginica	versicolor	virginica	versicolor
108	virginica	virginica	virginica	virginica	virginica	virginica
109	virginica	virginica	virginica	virginica	virginica	virginica
110	virginica	virginica	virginica	virginica	virginica	virginica
111	virginica	virginica	virginica	virginica	virginica	virginica
112	virginica	virginica	virginica	virginica	virginica	virginica
113	virginica	virginica	virginica	virginica	virginica	virginica
114	virginica	virginica	virginica	virginica	virginica	versicolor
115	virginica	virginica	virginica	virginica	virginica	versicolor
116	virginica	virginica	virginica	virginica	virginica	virginica
117	virginica	virginica	virginica	virginica	virginica	virginica
118	virginica	virginica	virginica	virginica	virginica	virginica
119	virginica	virginica	virginica	virginica	virginica	virginica
120	virginica	virginica	virginica	versicolor	virginica	versicolor
121	virginica	virginica	virginica	virginica	virginica	virginica
122	virginica	virginica	virginica	virginica	virginica	versicolor
123	virginica	virginica	virginica	virginica	virginica	virginica
124	virginica	virginica	virginica	virginica	virginica	versicolor
125	virginica	virginica	virginica	virginica	virginica	virginica
126	virginica	virginica	virginica	virginica	virginica	virginica
127	virginica	virginica	virginica	virginica	virginica	versicolor
128	virginica	virginica	virginica	virginica	virginica	versicolor

129	virginica	virginica	virginica	virginica	virginica	virginica
130	virginica	virginica	virginica	virginica	virginica	virginica
131	virginica	virginica	virginica	virginica	virginica	virginica
132	virginica	virginica	virginica	virginica	virginica	virginica
133	virginica	virginica	virginica	virginica	virginica	virginica
134	virginica	versicolor	versicolor	versicolor	versicolor	versicolor
135	virginica	virginica	virginica	versicolor	virginica	virginica
136	virginica	virginica	virginica	virginica	virginica	virginica
137	virginica	virginica	virginica	virginica	virginica	virginica
138	virginica	virginica	virginica	virginica	virginica	virginica
139	virginica	virginica	virginica	virginica	virginica	versicolor
140	virginica	virginica	virginica	virginica	virginica	virginica
141	virginica	virginica	virginica	virginica	virginica	virginica
142	virginica	virginica	virginica	virginica	virginica	virginica
143	virginica	virginica	virginica	virginica	virginica	versicolor
144	virginica	virginica	virginica	virginica	virginica	virginica
145	virginica	virginica	virginica	virginica	virginica	virginica
146	virginica	virginica	virginica	virginica	virginica	virginica
147	virginica	virginica	virginica	virginica	virginica	versicolor
148	virginica	virginica	virginica	virginica	virginica	virginica
149	virginica	virginica	virginica	virginica	virginica	virginica
150	virginica	virginica	virginica	virginica	virginica	versicolor