

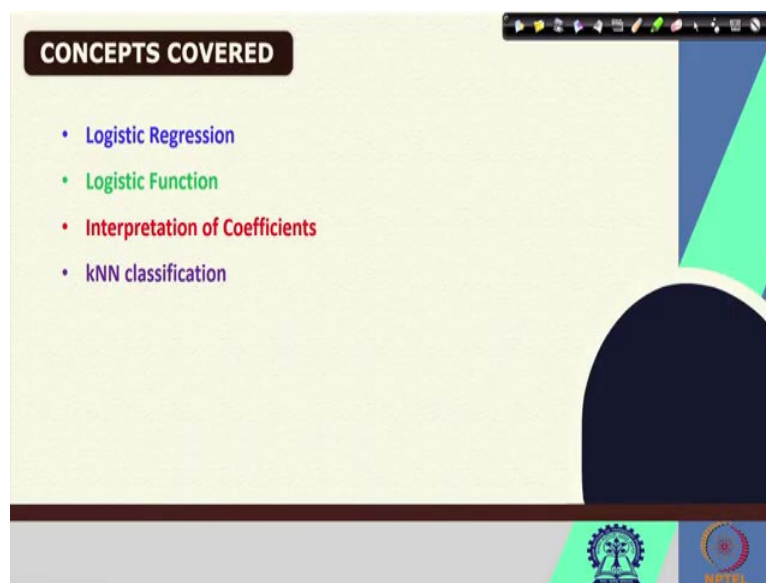
Machine Learning for Soil and Crop Management
Professor Somsubhra Chakraborty
Agricultural and Food Engineering Department
Indian Institute of Technology, Kharagpur
Lecture – 17

Application of Classification and Clustering Methods in Agriculture (Contd.)

Welcome friends to this 17th lecture of NPTEL online certification course of Machine Learning for Soil and Crop Management. We are currently this, in week four, and in this week we are discussing the Application of classification and clustering Methods in Agriculture. In our first lecture of this week, that is lecture 16th, we have discussed the basics of classification.

What is a classifier? What is the basic difference between a classification and clustering problem? Also, we have discussed the linear discriminant analysis which is 1 of the widely accepted or utilized linear classification problem. So, today, we are going to discuss some other linear discriminant, linear classification problems.

(Refer Slide Time: 01:29)



So, these are the major concepts which we are going to cover today. We are going to first cover the logistic regression, and then we are going to see what logistic function is. Then, we are going to see how to interpret the coefficients of a logistic regression. And then we are going to discuss another very important classification algorithm which is nonlinear classification is kNN or K nearest neighbor classification. So, let us start with the logistic regression.

(Refer Slide Time: 02:15)

KEYWORDS

- Logistic Regression
- Odds ratio
- kNN
- Recall
- Kappa coefficient

The slide features a light green background with a dark blue and green geometric design on the right side. A circular inset in the bottom right shows a man in a white shirt speaking. At the bottom, there are logos for a university and NPTEL.

Of course, some of the important keywords for these, for this lecture is logistic regression, then odds ratio we are going to learn. What is odds ratio? We are going to, we are going to learn what is kNN, we are going to learn some important classification metrics like recall, Kappa coefficient RLC curve, so we are going to discuss all these.

(Refer Slide Time: 02:45)

LOGISTIC REGRESSION

- LR is a process of modeling the probability of a discrete outcome given an input variable
- Generally used when the dependent variable is dichotomous (binary)
- The most common logistic regression models a binary outcome: take two values such as true/false, yes/no, etc
- Multinomial logistic regression: model scenarios where there are more than two possible discrete outcomes
- LR: a useful analysis method for classification problems

The slide features a light green background with a dark blue and green geometric design on the right side. A circular inset in the bottom right shows a man in a white shirt speaking. At the bottom, there are logos for a university and NPTEL.

Now, what is the logistic regression? So, logistic regression is a although it is a digression, it is can be used for classification problem. And the, it is a linear I would say method classification

problem. So, logistic regression is a process of modeling the probability of discrete outcome given an input variables. So, you know in case of normal regression problems our target variable is a continuous variable.

So, Y is always continuous. However, when our Y is not continuous it is a, it is a, it is a discrete variable then the logistic regression comes into play. So, generally the logistic regression is useful when the dependent variable or Y is dichotomous or binary. Sometimes we know, that the logistic, when our output is either yes, no, 0 or 1 then true, false this type of problem generally we deal with logistic regression.

When and you can see this is a binary outcome, either yes or no, either true or false either 0 or 1. So, in this kind of condition, we generally use the logistic regression model. However, when there are more than two possible discrete outcomes, then we go with the multinomial logistic regression. Generally, we are going to in this lecture we are going to discuss the binary logistic regression. So, logistic regression is a useful method for classification problems.

(Refer Slide Time: 05:18)

LOGISTIC REGRESSION

- There are many important research topics for which the dependent variable is "limited."
- For example: voting, morbidity or mortality, and participation data is not continuous or distributed normally.
- Binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable: coded 0 (did not vote) or 1 (did vote)

Now, what is the difference between a normal regression and logistic regression is that in case of the normal regression of a target or Y is continuous, however in case of logistic regression the dependent variable the outcome of the dependent variable is limited, as we have seen by dichotomous or binary or sometime multiple are there, but they are discrete in nature. For

example, vote or no vote, morbidity or mortality and participation data in a continuous or distributed normally.

So, so, under here you can see that in this case the data is not continuous or distributed normally. Here there are either yes or no vote or no vote morbidity or mortality. So, you can see these binary outcomes. So, binary logistic regression is a type of regression analysis where the dependent variable is a dummy variable coded 0. Suppose, yes or did not vote we do we code them with the dummy variables 0.

And also the other 1 we code them as 1 which actually another dummy variable. So, here let us consider that there are two outcomes what is vote and not and do not vote. So, in case of vote we can assign the dummy variable 1 and in case of did not vote, we can assign the dummy variables 0. So, this is the logistic regression.

(Refer Slide Time: 07:01)

LINEAR PROBABILITY MODEL

- In the OLS regression:
 $Y = \gamma + \phi X + e$; where $Y = (0, 1)$
- The error terms are heteroskedastic
- e is not normally distributed because Y takes on only two values
- The predicted probabilities can be greater than 1 or less than 0!

Now, in case of if we, if we define this regression in terms of ordinary least squares regression, then our model will assume this form we know that Y equal to γ plus ψ x plus e . Where this is the, the ψ is the slope, γ is the offset. So, however, in case of logistic regression the Y has only two values 0 and y_1 , where in case of normal regression, linear regression we generally assume that Y is a continuous variable.

And the values of Y are independent to each other. So, we can see here in this condition, the error terms are heteroskedastic. Remember in case of linear regression, we assume 1 of the major assumption is the error terms are homoscedastic or distributed evenly. However, when these are the cases when there are only binary outcome, then the error terms will be heteroskedastic.

So, because here in this case, the error will not be normally distributed because Y takes on only two values. So, since Y can take only 1 of these two values, the error term is not normally distributed. We remember we assume that the error was, is normally distributed in case of now, in case of normal linear regression or ordinary least squares regression.

So, in this condition when the error will not be normally distributed and the error terms are heteroskedastic because Y have only two values possible values, then we will see the predicted probabilities can be greater than 1 or less than 0, which is an unrealistic situation because we know that the probability of any occurrence of a, occurrence of any event cannot go beyond 1 or go less than 0.

It generally goes, varies between 0 to 1. So, if we are sticking with this binary variable, I mean binary outcome I want to do a linear regression or ordinary least squares regression, based on these data test set, then you will see that the predicted probabilities can be some time greater than 1 or less than 0 which is an unrealistic situation. So, to rectify this problem, we generally use the logistic regression.

(Refer Slide Time: 10:17)

LOGISTIC FUNCTION

- Also called the sigmoid function
- Useful to describe properties of population growth in ecology, rising quickly and maxing out at the carrying capacity of the environment
- It's an S-shaped curve that can take any real-valued number and map it into a value between 0 and 1, but never exactly at those limits

$$f(x) = \frac{L}{1 + e^{-k(x-x_0)}}$$

$f(x)$ = output of the function
 L = the curve's maximum value
 k = logistic growth rate or steepness of the curve
 x_0 = the x value of the sigmoid midpoint
 x = real number

The slide also includes a graph of the logistic function, a small video inset of a speaker, and logos for IIT Bombay and NPTEL.

But before we go for the logistic regression, let us first see what is the logistic function. So, here you can see this is a logistic function, you can see this is a logistic function. So, it is a sigmoid function you can clearly see. And it is useful to describe generally, generally this sigmoid function is used for describing the population growth in ecology in, when there is a rising quickly and, and maxing out at the carrying capacity of the environment.

So, here it is reaching a plateau when it is reaching the carrying capacity of the environment. So, it is called a sigmoid function. Generally, we use this logistic function in case of ecological application. So, similarly this logistic function so, the what is the mathematical? So, mathematical representation of the logistic function we can see here. So, it is an S shaped curve that can take any real valued number and map it into the value between 0 to 1, but never exactly at those limits.

So, this is how we can limit or we can confine the probability of occurrence of any event within 0 to 1. And it does not allow the probability to go beyond 1 or less than 1 and thus rectifying the problem of discrete dichotomous outcome or binary outcome of the target variable. So, this is the mathematical form, here you can see L is the curve maximum value. Here small k is the logistic growth rate or steepness of the curve. And the x, the x the x value at the sigma midpoint here and also x is when x is the real number. So, this is how you calculate the logistic function.

(Refer Slide Time: 12:22)

LOGISTIC REGRESSION

The "logit" model solves these problems:

$$\ln\left(\frac{p}{1-p}\right) = \alpha + \beta X + e$$

- p is the probability that the event Y occurs, $p(Y=1)$
- $p/(1-p)$ is the "odds ratio"
- $\ln[p/(1-p)]$ is the log odds ratio, or "logit"

Handwritten notes: $\frac{p}{1-p}$ in a circle, $\ln\left(\frac{p}{1-p}\right) = \text{logit}$

Video inset: A man in a white shirt speaking.

Logos: IIT Bombay and NPTEL.

LINEAR PROBABILITY MODEL

- In the OLS regression:
 $Y = \gamma + \phi X + e$; where $Y = (0, 1)$
- The error terms are heteroskedastic
- e is not normally distributed because Y takes on only two values
- The predicted probabilities can be greater than 1 or less than 0!

Video inset: A man in a white shirt speaking.

Logos: IIT Bombay and NPTEL.

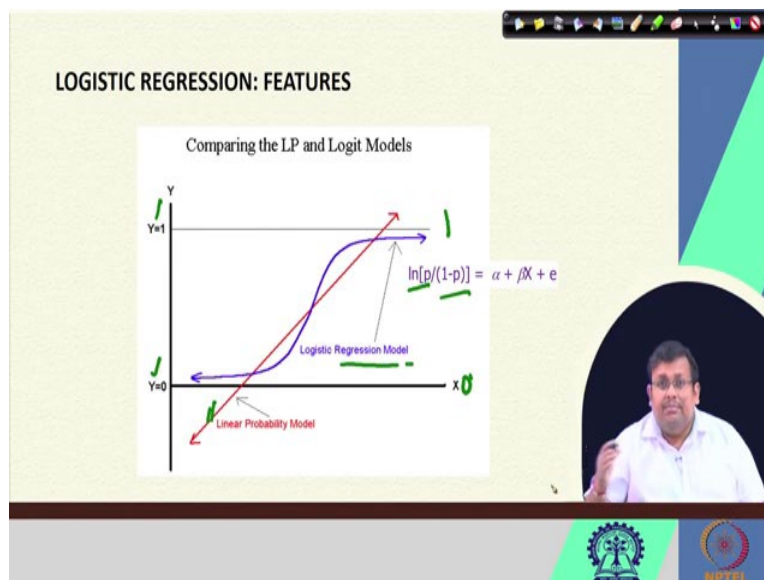
So, why this called a logistic problem? Logistic, in the logistic regression these OLS regression can be reshaped into this form, what is this form? This form is the logistic regression form. So, here you can see, we are taking the logarithm of p by $1 - p$ which is $\alpha + \beta x + e$. So, here we are we are, we are, we are producing a logit model to solve the problem of exceeding the probability beyond 1 and, reducing the probability less than 0.

So, we are taking this a logit model, this is called a logic model. So, here p what is p ? P is the probability that the event Y occur. So, p is the probability that an event Y will occur. And the

value, and the value and the ratio of p by 1 minus p is known as the odds ratio or means probability. So, here this is known as the odds ratio. So, when we are taking the log of these odds ratio p minus 1 by p it is also known as log odds ratio or in simple form we call it logit.

So, this is called a logit, when we are taking the logarithm of this ratio of probabilities. So, we can re arrange, the prop that the model in this form we call it a logit model. And what, why we do these? Because when we do this type of representation in terms of logit model that will restrict the probability within 0 to 1 .

(Refer Slide Time: 14:48)



So, graphically, you can see it here in case, suppose, there, there are two as I have mentioned, here you can see there are two outcome Y equal to 0 and Y equal to 1 . And it is when we plot Y versus x , we can see here this is the linear probability model. So, this is a linear probability model and which is exceeding the probability below 0 and above 1 .

And here you can see when we are using the same in the form of logistic regression model, where log of odds when you are expressing in terms of α plus βX plus e , β is the coefficient. Then you can see, we are restricting the probability through the sigmoid function within this probability between 0 to 1 . So, this is the benefit of using these logistic regression, as compared to the linear probability model.

(Refer Slide Time: 16:01)

LOGISTIC REGRESSION: FEATURES

- The logistic distribution constrains the estimated probabilities to remain between 0 and 1.
- The estimated probability is:
$$p = 1/[1 + \exp(-\alpha - \beta X)]$$
- if you let $\alpha + \beta X = 0$, then $p = .50$
- as $\alpha + \beta X$ gets really big, p approaches 1
- as $\alpha + \beta X$ gets really small, p approaches 0

Small video inset showing a man speaking.

So, the logistic distribution contributes, sorry the logistic distribution constrains the estimated probabilities to remain between 0 and 1. So, we know that estimated probability from this logistic function if you simplify, we can get p equal to $1 / (1 + \exp(-\alpha - \beta x))$. So, from here, we can see that if these $\alpha + \beta x$ equal to 0, then this probability generally goes to 0.50.

And when the $\alpha + \beta x$ gets really big, the probability of, the estimated probability approaches 1. And when the $\alpha + \beta x$ gets really small, the probability or p generally approaches 0. So, this is how we can restrict the probability of estimated probability between 0 to 1.

(Refer Slide Time: 17:25)

LOGISTIC REGRESSION: FEATURES

- Since:
 $\ln\left[\frac{p}{1-p}\right] = \alpha + \beta X + e$
The slope coefficient (β) is interpreted as the rate of change in the "log odds" as X changes :confusing
- Since:
 $p = \frac{1}{1 + \exp(-\alpha - \beta X)}$

Now, now, some other features let us discuss some other features. Now, since we know that log of based on the logistic regression, the sum of the log of sorry, the log of p by 1 minus p will be alpha plus beta x plus e eta or error. So, the slope coefficient, so, in this form if we consider this form and want to describe the coefficient in terms of this logistic regression, then we can say that the slope coefficient which is beta is interpreted as a rate of change in the log odds of x changes.

So, here beta is representing the rate of change of log of words with x as, as x is changing. But, this could be somewhat confusing to understand. Now, since we know this is the simplified form of this probability, we can further simplify it like this.

(Refer Slide Time: 18:50)

LOGISTIC REGRESSION: FEATURES

- An interpretation of the logit coefficient which is usually more intuitive is the "odds ratio"
- Since,
$$\left[\frac{p}{1-p} \right] = \exp(\alpha + \beta X)$$

 $\exp(\beta)$ is the effect of the independent variable on the "odds ratio"

So, another interpretation of this logistic or logit coefficient, which is usually more intuitive is that odds ratio. So, instead of using the same representation we can do another way instead of, instead of expressing the logistic regression in the previous form, we can just present the odds ratio and take the exponential in this side. So, p by $1 - p$ will be exponential of $\alpha + \beta x$.

So, here we can say that exponential β is the effect of the independent variable on the odds ratio. So, this is how we can more reasonably express the coefficient of a logistic regression. So, again, the exponential of β is the effect of the independent variable on the odds ratio. So, this will be much more meaningful than using the original formula, while explaining the β coefficient. So, instead of explaining the β coefficient here, we are explaining in terms of exponential of β .

(Refer Slide Time: 20:24)

LDA vs LOGISTIC REGRESSION

1. Logistic Regression is conventionally used for two-class and binary classification problems
2. Though LR can be extrapolated and used in multi-class classification, this is rarely performed
3. LDA is considered a better choice whenever multi-class classification is required
4. In the case of binary classifications, both LR and LDA are applied

So, we have seen both now, the logistic linear discriminant analysis we have also seen the logistic regression. So, logistic regression is conventionally, what is the difference between these two? So, logistic regression is conventionally used for two class and binary classification problem. So, that is why we generally go with the linear discriminant analysis. So, though logistic regression can be extrapolated and used in multiclass classification also, this is rarely performed.

Generally we stick with the binary classification problems. So, LDA is considered is a better choice whenever multi class, classification is required. And in the case of binary classification, we can use both of them. In case of binary classification, both a linear, logistic regression and linear discriminant analysis both of them are applied. However, in case of multi class classification, we generally prefer the linear discriminant analysis over the logistic regression.

(Refer Slide Time: 21:44)

kNN CLASSIFICATION

- Non-parametric classification algorithm
- Classifies based on a similarity measure
- Classified by “MAJORITY VOTES” for its neighbor classes
- Assigns to the most common class amongst its K-nearest neighbors (by measuring “distant” between data)

Ref: <https://www.slideshare.net/tilanigunawardena/k-nearest-neighbors>

Now, let us discuss, we have discussed some parametric classification algorithm called linear classification algorithm. Now, let us see some nonparametric classification algorithm. So, the nonparametric classification algorithm here, the kNN classification or K nearest neighbor classification which classify based on the similarity measure. So, it classify based on the similarity measure and classify by majority of the votes for its neighbor classes.

And assign the most common class among these K nearest neighbors. So, but, first of all it identify suppose, we are assigning any value of K, so, it first identify. Suppose K is 3 so, first it will identify the nearest three classes for any unknown sample and then it will assign the label of the class based on the maximum voting. So, this is how, this kNN classification generally works, by measuring the distance between the data set.

So, the major question in case of kNN is suppose there is a suppose, there is a new sample and we want to assign it based on the distance of this sample to one of these two classes. And it will identify the nearest class and then based on that distance it will classify the sample, unknown sample to one of these classes. So, let us move ahead and see the kNN classification.

(Refer Slide Time: 24:05)

kNN CLASSIFICATION

1. Step 1 – Load the training as well as test data
2. Step 2 – Choose the value of K i.e. the nearest data points. K can be any integer.
3. Step 3 – For each point in the test data do the following –
 - 3.1 – Calculate the distance between test data and each row of training data with the help of any of the method namely: Euclidean, Manhattan or Hamming distance. The most commonly used method to calculate distance is Euclidean.
 - 3.2 – Now, based on the distance value, sort them in ascending order.
 - 3.3 – Next, it will choose the top K rows from the sorted array.
 - 3.4 – Now, it will assign a class to the test point based on most frequent class of these rows.
4. Step 4 – End

Ref: http://www.scholarpedia.org/article/K-nearest_neighbor

So, here kNN classification the step by steps are given. Suppose there are multiple, there are two classes, red and blue, and they are mixed together. So, we want to assign a new sample which is denoted by this green dot to one of these two classes. So, how we can do that using kNN classification? So, there are several steps, first of all, we need to load the training as well as the test data.

So, suppose this green data is the test data and all other samples are the training data set. So, we need to first choose the value of K, that is nearest data points K can be any integer, 3, 4 any integer you can select. In the next step you can calculate the distance between this test data, between this test data and each row of the training data.

So, you can calculate the distance between this training between, this test data and between this test data and the linear distance I am sorry, the distance between the test data and each row of the training data with the help of the method. So, there are several methods for calculating the distance, Euclidean distance are there, Manhattans or hamming distance are there. However, in most of the cases, we generally use the Euclidean distance.

Now, based on the distance values, we then arrange or sort them these, adding them into ascending order. And then it will choose the top K rows and from the, from the, from the sorted array. So, if they are it is, it is when the K value is 3, it will select the top three rows and then it

will see what is the maximum outcome out of these three rows or maximum class levels predicted class levels.

So, if we see that the three rows, out of the three rows two of them are showing the yes and one of them are showing no or in other words, if two of them are showing the red class and only one of them are assigned to blue class, then these unknown or validation samples will be assigned to this kNN, to this red group.

So, this is how, so, again, let me let me summarize this, first we load the training data, then we calculate the distance between the test data with these, between these each row of the training data. And then we ascend them, we sort them in the ascending order, and then we select the top three or top four based on the K values and then we assign the value or assign the class level based on the majority class for those three or four observations. So, this is how the K nearest neighbor algorithm basically works.

(Refer Slide Time: 27:44)

ADVANTAGES OF kNN

1. It is very simple algorithm to understand and interpret
2. It is very useful for nonlinear data because there is no assumption about data in this algorithm
3. It is a versatile algorithm as we can use it for classification as well as regression
4. It has relatively high accuracy but there are much better supervised learning models than KNN
5. Since the KNN algorithm requires no training before making predictions, new data can be added seamlessly which will not impact the accuracy of the algorithm

So, what are the advantages of K nearest neighbor algorithm? So, it is a very simple algorithm to understand and interpret. It is very useful for nonlinear data, because there is no assumption about data in this algorithm. It is a versatile algorithm as we can see, as we can use it for classification as well as regression.

It has relatively high accuracy, but, there are much better supervised learning models than kNN. And since the kNN algorithm requires no training, before making prediction, new data can be added seamlessly, which will not impact the accuracy of the algorithm. So, these are some advantages of the algorithm.

(Refer Slide Time: 28:27)

DISADVANTAGES OF kNN

1. It is computationally a bit expensive algorithm because it stores all the training data
2. Accuracy depends on the quality of the data
3. High memory storage required as compared to other supervised learning algorithms
4. Prediction is slow in case of big N
5. It is very sensitive to the scale of data as well as irrelevant features

However, there are some disadvantages of kNN also. First of all, it is computationally a bit expensive algorithm because it stores all the training data. And here the accuracy depends on the quality of the data. And high memory storage is required as compared to other supervised learning algorithm. Then prediction is slow in case of big, big values of N, and also it is a very sensitive to the scale of the big number of data and when it is very sensitive to the scale of the data as well as the, as well as well as irrelevant features.

(Refer Slide Time: 29:15)

CLASSIFICATION PERFORMANCE METRICS

A Confusion matrix is commonly used to visualize the performance of a classification algorithm. Figure 2 illustrates the confusion matrix for a multi-class model with N classes [34]. Observations on correct and incorrect classifications are collected into the confusion matrix $C(c_j)$, where c_j represents the frequency of class j being identified as class i . In general, the confusion matrix provides four types of classification results with respect to one classification target k :

- True Positive (TP)—correct prediction of the positive class ($c_{k,k}$)
- True Negative (TN)—correct prediction of the negative class ($\sum_{i \neq k} c_{i,i}$)
- False Positive (FP)—incorrect prediction of the positive class ($\sum_{i \neq k} c_{i,k}$)
- False Negative (FN)—incorrect prediction of the negative class ($\sum_{i \neq k} c_{k,i}$)

Figure 2. Confusion matrix for a multi-class model with N classes [34].

So, these are the some of the important advantages and disadvantages of, for K nearest neighbor classification. Now, the classification problem we can define in terms of different performance metrics or confusion matrix, we have already discussed what is the confusion matrix, and what are the two positive true negative false positive and false negative. We already know from our previous discussion of confusion matrix, we generally use for identifying or calculating the accuracy different performance metrics for classification.

(Refer Slide Time: 29:53)

CLASSIFICATION PERFORMANCE METRICS: A BETTER REPRESENTATION

	Predicted:	
	NO	YES
n=165		
Actual: NO	50	10
Actual: YES	5	100

	Predicted:		
	NO	YES	
n=165			
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

<https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

So, as it is a simple example here is given. Suppose, we are classifying some disease and suppose these is true negative, where actually there are no disease and we are predicted the disease there is no disease. So, it is a true negative when there is actually no disease, but we predicted yes there are disease. So, it is a false positive and here you can see this is a false negative, because there is actually disease but we are predicted no.

And here when the actual disease is there, and we have also predicted disease. So, here you can see, this is called the true positive. So, true positive, true negative, false positive, false negative four types of outcome we can generate from any confusion matrix.

(Refer Slide Time: 30:48)

CLASSIFICATION PERFORMANCE METRICS

Several performance metrics can be derived from these four outcomes. The ones of interest to us are listed below, for per-class classifications:

- Accuracy: This metric simply measures how often the classifier makes a correct prediction.

$$\text{Overall Accuracy} = \frac{\sum_{i=1}^N C_{ii}}{\sum_{i=1}^N \sum_{j=1}^N C_{ij}} \quad (1)$$
- Recall (Sensitivity or True Positive Rate): This metric denotes the classifier's ability to predict a correct class

$$\text{Recall}_{\text{class}} = \frac{TP_{\text{class}}}{TP_{\text{class}} + FN_{\text{class}}} \quad (2)$$
- Precision: This metric represents the classifier's certainty of correctly predicting a given class

$$\text{Precision}_{\text{class}} = \frac{TP_{\text{class}}}{TP_{\text{class}} + FP_{\text{class}}} \quad (3)$$
- False Positive Rate (FPR): This metric represents the number of incorrect positive predictions out of the total true negatives

$$\text{FPR}_{\text{class}} = \frac{FP_{\text{class}}}{FP_{\text{class}} + TN_{\text{class}}} \quad (4)$$

Abraham et al. (2020)

And using those values we can calculate some important performance metrics like overall accuracy, recall, precision, FPR. FPR stands for the false positive rate. So, accuracy stands for this metric simply measures how often the classifier makes a correct prediction. And recall is basically the sensitivity or true positive rate, this matrix denotes the classifiers ability to predict a correct class, so formula is given here.

Precision, it is a measure of classifier's certainty of correctly predicting a given class. False positive rate or FPR is the metric representation; it is a representation of the number of incorrect positive predictors out of the total true negatives.

(Refer Slide Time: 31:39)

CLASSIFICATION PERFORMANCE METRICS

- True Negative Rate (TNR or Specificity): This metric represents the number of correct negative predictions out of the total true negatives
$$TNR_{class} = \frac{TN_{class}}{FP_{class} + TN_{class}} \quad (5)$$
- F1-Score: This metric is a harmonic mean of precision and recall. Although the F1-score is not as intuitive as accuracy, it is useful in measuring how precise and robust the classifier is. Abraham et al. (2020)
$$F1 - Score_{class} = \frac{2 * TP_{class}}{2 * TP_{class} + FN_{class} + FP_{class}} \quad (6)$$
- Matthews Correlation Coefficient (MCC): For binary classification, MCC summarizes into a single value the confusion matrix. This is easily generalizable to multi-class problems as well.
$$MCC_{class} = \frac{TP_{class} * TN_{class} - FP_{class} * FN_{class}}{\sqrt{(TP_{class} + FP_{class}) * (TP_{class} + FN_{class}) * (FP_{class} + TN_{class}) * (FN_{class} + TN_{class})}} \quad (7)$$
- Cohen's Kappa (κ): This metric compares an Observed Accuracy with an Expected Accuracy (random chance)
$$\kappa_{class} = \frac{p_o - p_e}{1 - p_e} \quad (8)$$

where p_o represents the accuracy and p_e represents a factor that is based on normalized marginal probabilities.

Also, there are true negative rate, f1 score, Matthews's correlation coefficient, Cohen's Kappa, there are different types of performance metrics. And we will discuss it, in our next class these things in details, but here I want to just focus on this Cohen's Kappa, because it is widely used. This metric compares and observed accuracy with the expected accuracy. Expected accuracy means accuracy by random chance.

We are also going to discuss this in details in our, in our upcoming lectures and also when we are going to discuss the digital soil mapping. So, these are all indicative of the performance of the classification algorithm.

(Refer Slide Time: 32:28)

CLASSIFICATION PERFORMANCE METRICS

ROC Curve: this is a commonly used graph that summarizes the performance of a classifier over all possible thresholds. It is generated by plotting the True Positive Rate (y-axis) against the False Positive Rate (x-axis) as you vary the threshold for assigning observations to a given class.

The slide features a video inset of a man in a white shirt speaking. At the bottom, there are logos for IIT Bombay and WIPAC.

So, there is another important performance metrics called ROC curve. And this is generally commonly used to graph or summarize the performance of a classifier over all the possible threshold. And it is generated by plotting the true positive rate against the false positive rate. We have already discussed these are important metrics, as you vary the threshold for assigning the observation to a given class.

So, ROC curve you also can see in some of the literature, where they have used the ROC curve for as an as in performance metrics of any classification algorithm. So, in our upcoming lectures, we are going to, we are going to discuss this performance metrics in details.

(Refer Slide Time: 33:12)

APPLICATIONS OF KNN IN SOIL

Table 3. Comparison of performance metrics for classification using k-Nearest Neighbors (kNN), Support Vector Machine (SVM) and decision trees.


	k-Nearest Neighbor (kNN) k = 4						
	Recall	Precision	FPR	TNR	F1 Score	MCC	Kappa
HSG A	0.84	0.86	0.03	0.97	0.84	0.82	0.89
HSG B	0.85	0.84	0.20	0.80	0.84	0.65	0.88
HSG C	0.72	0.73	0.09	0.91	0.72	0.63	0.56
HSG D	0.73	0.83	0.01	0.99	0.77	0.76	0.89
Macro Average	0.79	0.81	0.08	0.82	0.79	0.72	0.56
Micro Average	0.80	0.80	0.07	0.93	0.80	0.73	0.73

	Support Vector Machines (SVM) Gaussian Kernel						
	Recall	Precision	FPR	TNR	F1 Score	MCC	Kappa
HSG A	0.90	0.79	0.05	0.95	0.84	0.81	0.87
HSG B	0.86	0.71	0.42	0.56	0.77	0.46	0.59
HSG C	0.35	0.65	0.06	0.94	0.45	0.36	0.66
HSG D	0.54	0.98	0.00	1.00	0.69	0.72	0.91
Macro Average	0.66	0.78	0.13	0.87	0.69	0.59	0.58
Micro Average	0.74	0.74	0.09	0.91	0.74	0.65	0.65

	Decision Tree						
	Recall	Precision	FPR	TNR	F1 Score	MCC	Kappa
HSG A	0.80	0.84	0.03	0.97	0.86	0.83	0.88
HSG B	0.91	0.83	0.22	0.78	0.87	0.70	0.80
HSG C	0.67	0.82	0.05	0.95	0.74	0.67	0.59
HSG D	0.66	0.85	0.01	0.99	0.74	0.73	0.90
Macro Average	0.79	0.84	0.08	0.82	0.80	0.73	0.55
Micro Average	0.79	0.79	0.07	0.93	0.79	0.72	0.72

Soil samples were classified into one of the four hydrologic groups using soil texture calculations

Abraham et al. (2020)



Just I want to, wrap up this lecture by showing some soil and plant application of K nearest, K nearest neighbors method. So, you can see here, here in this example, soil samples were classified into one of the four hydrologic groups. So, here you can see four hydrologic model says HSG A, HSG B, HSG C, HSG D, based on this K nearest neighbor and all. So, they have used the support vector machine classification, decision tree based classification.

We already know decision tree and support vector machine classification. And here you can see the K nearest neighborhood classification they have used using the K value of 4. So, different performance may increases are given like recall position FPR, TNR, F1 score MCC Kappa and from there they have calculated which classification algorithm performed better. So, this is the application of KNN in soil.

(Refer Slide Time: 34:14)

APPLICATIONS OF KNN IN PLANT

Table 3. Results obtained with different classifiers using RGB histogram features for image-level classification.

Channel	R			G			B			RGB		
	TFR	TNR	ACC	TFR	TNR	ACC	TFR	TNR	ACC	TFR	TNR	ACC
Class	F	N	%	F	N	%	F	N	%	F	N	%
Five KNN	99.1%	100%	99.3%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Cubic SVM	94.3%	100%	97%	97.3%	100%	96.3%	99.1%	100%	99.3%	99.1%	100%	99.3%
Boosted Tree	100%	100%	100%	100%	100%	100%	100%	9%	10.8%	100%	9%	10.8%
Bagged Tree	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Complex Tree	99.1%	100%	99.3%	100%	100%	100%	99.1%	100%	99.3%	99.1%	100%	99.3%

Table 4. Combined results obtained using HSV histogram features for image-level classification.

Channel	H			S			V			HSV		
	TFR	TNR	ACC	TFR	TNR	ACC	TFR	TNR	ACC	TFR	TNR	ACC
Class	F	N	%	F	N	%	F	N	%	F	N	%
Five KNN	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%	100%
Cubic SVM	97.3%	100%	98.2%	94.4%	100%	96%	97.3%	100%	98.2%	94.3%	100%	97%
Boosted Tree	100%	0%	10.8%	98.2%	100%	99%	99.1%	100%	99.3%	100%	0%	10.8%
Bagged Tree	100%	100%	100%	98.2%	100%	99%	100%	100%	100%	100%	100%	100%
Complex Tree	98.2%	100%	99%	98.2%	100%	99%	100%	100%	100%	98.2%	100%	99%

Table 5. Results obtained with different classifiers using LBP features for image-level classification.

Features	RLBP			GLBP			SLBP			LBP		
	TFR	TNR	ACC	TFR	TNR	ACC	TFR	TNR	ACC	TFR	TNR	ACC
Class	F	N	%	F	N	%	F	N	%	F	N	%
Five KNN	100%	100%	100%	100%	100%	100%	98.2%	100%	99%	100%	100%	100%
Cubic SVM	96.4%	100%	98%	91.8%	100%	95.4%	95.3%	100%	97.3%	100%	100%	100%
Boosted Tree	100%	0%	10.8%	100%	0%	10.8%	100%	0%	10.8%	100%	0%	10.8%
Bagged Tree	99.1%	100%	99.3%	98.2%	100%	99%	99.1%	100%	99.3%	99.1%	100%	99.3%
Complex Tree	99.1%	100%	99.3%	98.2%	100%	99%	98.2%	100%	99%	99.1%	100%	99.3%

(a) Normal (b) Rust
(c) Carbon (d) Monilia

Almadhor et al. (2021)

And also here, application of KNN in plant features for image-level classification. Images of guava were used for classifying the samples. Basically, the image features of different guava diseases were used, and they were classified using different types of classification algorithms like K nearest neighbor and then support vector machine, then boosted tree, back tree, complex tree. So, people, scientists are using different types.

So, they have used, first instance they have used the RGB histogram, which is a specific color model. In the second table is showing the HSB histogram. In the third is to, in the third table it is showing LBP features. So, these features they have used from the images. And from collecting from different diseases and then they have tried to predict the target variable. So, these are some examples of KNN application in plant.

(Refer Slide Time: 35:35)



REFERENCES

Abraham, S.; Huynh, C.; Vu, H. Classification of Soils into Hydrologic Groups Using Machine Learning. *Data* **2020**, *5*, 2. <https://doi.org/10.3390/data5010002>

Almadhor, A.; Rauf, H.T.; Lali, M.I.U.; Damaševičius, R.; Alouffi, B.; Alharbi, A. AI-Driven Framework for Recognition of Guava Plant Diseases through Machine Learning from DSLR Camera Sensor Based High Resolution Imagery. *Sensors* **2021**, *21*, 3830

<https://www.slideshare.net/tilanjuncwardena/k-nearest-neighbors/>
http://www.scholarpedia.org/article/K-nearest_neighbor
<https://www.dataschool.io/simple-guide-to-confusion-matrix-terminology/>

So, guys, I hope that you have learned something new while we discuss the logistic regression and also K nearest neighbor classification. These logistic, while logistic regression is a linear classification method. However, this kNN method is a nonparametric method. So, these are the references which I used, you can, you can this you can, you can go back to these references and you can have more information regarding these technologies.

Thank you and, we will, we will go from here in our next class, and we will also discuss in details about different performance metrics in our next lectures. So, stay tuned and let us join in our next lecture to discuss other classification and clustering algorithms. Thank you.