Transcriber Name: Saji Paul
Statistical Learning for Reliability Analysis
Prof. Monalisa Sarma
Subir Chowdhury School of Quality and Reliability
Indian Institute of Technology, Kharagpur

Lecture - 53
Tutorial on Classification Techniques

(refer time: 00:33)

Hello guys, so today is the tutorial class, tutorial class means interesting. Anyway, so now coming to our tradition of doing a tutorial class first we solve some objective equations then we solve some problems so today also we will be following the same trend.
(refer time: 00:47)

So, first the objective type questions see here. Which of the following statement is true or data which of the following statements are true? In supervised learning all input training data are level with some class names as it is true or false and supervised learning all input training data are level with some class name that is definitely true. Then what is what about the next statement in supervised learning all input training data are not necessarily level with class names.

That is definitely false. Supervised training means in the training data definitely the observation along with the observation falls in which class it is specified. Clustering is a type of supervised learning technique. Is it true? No, clustering is a type of unsupervised learning technique an unsupervised learning technique where technical class levels are not there. So, we try to group the objects based on the similar characteristics so that is that we call it clustering.

So, in unsupervised learning all input training data are level with some class name that is definitely false. So, our true is the only statement a.
(refer time: 01:57)

So, next, you have a data set of different flowers containing their petal lengths and colour. Different flowers there are sets of different flowers it contains their petal lengths and their colour. A model has to predict the type of flower for a given petal length and colour which type of model it would be? So, we have a data set that contains the petal length and a colour and this it corresponds to which flower that is specified.

Now we have to develop a model which will predict the type of flower given a petal length as well as the colour. So, which type of model is it, is it a Ch-square test definitely not Ch-square test we use for what we use we use Ch-square trace for correlation analysis and this is not a correlation. Then a classification model, is it a classification model, definitely it is a classification model we will have to predict to which class it falls.

Clustering model definitely not because the class levels are already specified into different flowers containing it is specified for different petal lengths and colour it belongs to which class it is already in a training that it is specified. So, it is not a clustering model regression analysis, definitely not, I think you can remember what is regression analysis what we do in regression analysis given x what will be the value of y that is what we find in regression analysis. So, it is a classification model.
(refer time: 03:22)

Next consider a table given below the table shows that even y has two outcomes namely A and B which is dependent on another event x with various outcome like x 1, x 2 and x 3. So, even y has two outcomes simply A and B and this is dependent on another even X that means Y we can say it is a class variable sort of it is just there are two events A and B we can say consider it is a different A is a class B is a class.

And it depends on the variable x and x has three variables outcome x 1 x 2 x 3. So, depending on the values of x 1 x 2 x 3 will be able to predict whether it falls in A or it falls in B. Now it is asking what is the question what is the probability of P X = x 1 given Y = B, given Y = B what is the probability that X will take a value x 1? So, this is an example of but there is an example of yesterday we have done posterior probability.

Remember we have done what is prior probability posterior probability if I would have asked what is the probability of Y. Then it is it would have been a prior probability probably Y = A it is a prior probability of Y = B that is a prior probability. Now what is the probability of X = x 1 given Y = B that is a positive probability. So, how do we find this out so what is the probability of X = x 1 given Y = B, how many occurrence of B are here?

There are 5 occurrence of B are here 1 2 3 4 5 and for given that Y = B what is the probability that X = x 1. X = x 1 how much we have just one just one so it is 1 / 5 that is 1 / 5 = 0.2 is the answer. We have done this type of problem and last two last lectures when we have tried to calculate the prostate probability and prior probability in fact for Bayesian analysis also, we have used a similar kind of example.
(refer time: 05:36)

Now for two events A and B the base theorem states that so, this is the formula which is a formula for base theorem. What is the probability of a given B how do we find out probability of a given B probability of A given B? So, what is this? This is probability (B | A) x probability of A divided by the total problem total probability B is not it. So, which is here given this is the correct answer.
(refer time: 06:14)

Naive Bayesian classifier cannot be applied to training data with, what it is given cannot be applied to training data with categorical attribute. Naive Bayesian classifier; remember naive Bayesian classifier cannot be applied to training data with categorical attribute. Which one is

correct? It cannot be applied to name base we cannot apply to which sort of things we can apply to category attribute definitely as we can apply to category.

Numerical attribute with continuous range of values even that also we can apply. There we have seen there are two approach out so if it is a continuous range of values and it is a numerical attribute is a continuous range of values how we can solve it. One we will either discretize it or we will keep use as a probability distribution function try to find out the mu n sigma and then we can find out the posterior probability.

So, this is also you can do it and null occurrence of some attribute values for some classes if there is a null occurrence of some attribute value still, we can use it M estimate. Remember the M estimate approach if there are null occurrence that we could solve it by using the m estimate approach M estimate approach remember the formula. We will see some examples as well. So, this is also a mixed set of attributes even that is, so which the answer is none of the above. So, all these things it can be applied to Bayesian classifier, so answer is none of the above.
(refer time: 07:49)

Time complexity of the name base classified for n features k class data. What is the time complexity for name-based classifier? I will not go how to find out the time complexity because that is a part of algorithm course and I think all of you know that also. So, if you are interested in finding out the time complexity of name-based classifier which is n features and k class data. So, n features this side we have features decide we have class we have seen the diagram. So, how many total what will be the complexity it is nothing but order of n cross k.
(refer time: 08:28)

So, which of the following is true about kNN algorithm? kNN algorithm does more computation on testing time rather than training time, is it true? Definitely not true kNN is called a lazy algorithm why because it does not work in training time at all it just stores the data during training time and during testing time only it does what work it has to be done. Again, an algorithm does more computation on training time rather than testing time this is correct.

kNN performs much better if all of the data are of different scales, is it true or false? See kNN performs much better if all of the data are of different scale see if all of the data are of different scale what happens. In kNN we try to find out the difference between the; we try to find out a distance among the attributes. Suppose in the two points it has two attributes say one we have a attribute say a particular attribute say age and second attribute say we have a salary.

Two attributes some sort of classification we are doing and two attributes age and salary and suppose a is range from say 30 to 40. Age could take the value from 30 to 40 and salary it can take the value from say 70000 maybe 70000 to maybe 10 lakhs. So, salary from 70000 to 1 lakh you can imagine the brains so when we try to find out. So, this each attribute will have will have this value each observation each observation will have the value say age and salary.

So, and then when we try to find out the find the difference between the given any testing data,

we try to find out the difference of this testing data with all of the training data. This salary because of its range of data it will overpower the age what to say the contribution the age attribute will make. So, that is that way what happened only salary will be the defining factor which will this deciding factor which will decide it which is the which value is nearest.

Though age is also one of the important equally important factor but because of its very lower range compared to the salary it becomes overpowered by the other attributes. So, that is why kNN and it does it works very badly when the data are of different scale. So, kNN performs much better it is definitely wrong kNN does not performs better kNN performs very bad when the data are of the different scale data should be usually another uniform skill.

And not necessarily in the same scale but it will be it should not be there should not be too much a variance it too much of variance then one overpowers the other and the overprint powering attribute becomes the defining factor deciding factor to decide object or observation will belong to which class. When kNN derives a mathematical relationship between the feature set and a class level during training, definitely not during training it does not do anything.

So, only a is correct it does more computation on testing time rather than training time, in training time it does not do anything, so only a is correct.
(refer time: 11:52)

Which of the following is true about kNN algorithm? It is a parametric and eager learner algorithm. See here I have in a note I have given what is eager learning and what is lazy learning, lazy learning I have already mentioned but still you can see. Lazy learning is a learning method in which generalization of the training data is delayed until a query is made to the system. Lazy learning means it does not do any work during the training time.

It only starts working when the train testing data has come so that is we call it lazy learning. That is what kNN does then and there is another kind of learning which you call eager learning and in the case of eager learning the system tries to generalize the training data before receiving queries. Before receiving any queries only, it tries to form the model. So, that once it receives the query just by utilizing the model it can find out the class.

So, that is we call it eager learning so kNN is definitely not a eager learning kNN is very much a lazy learning. So, which one is correct here it is a parametric and eager learner algorithm the kNN is not a parametric which I have already mentioned here the data as we are not bothered about data belongs to each distribution on and it is definitely not a eager learning algorithm also. It appearance it is a parametric and lazy learner algorithm it is a lazy learner but it is not parametric. So, answer is d, it is a non-parametric and lazy learner algorithm.
(refer time: 13:24)

Consider the following figure. The black triangles are the data points belonging to class 1 and the grey rectangles are the data points these are the grey rectangles data points belonging to class two, the test point is marked as green this is the test point green pentagon. What is the

class level of the test point as per kNN and classified which k = 1 and k = 3 if k = 1 then what will be the class level of this place point?

If k is only one which one is nearer it is only nearer to this green rectangle. So, then its if k = 1 its class level is class 2. So, the answer will be between these two definitely this and this will not be the answer. Now we will see what happens when k = 3, when k = 3 which are the nearest point, these are the nearest point then it belongs to if when k = 3 more number of class 1 points are nearer to the test point. So, that means when k = 3 that this test point belongs to class 1. So, my answer will be this.
(refer time: 14:45)

Which of the following statement is true about kNN classification algorithm? See if you see the objective questions in fact in all of my tutorial if you go through the objective question then if you forget anything you can just go back to the lecture and everything will be clear actually. I am trying to touch each and every topic what I am teaching in the lecture classes I am touching each and everything in my tutorial classes. It is sort of a quick recapitulation actually.

This which is not discussed only. Which of the following statement is true about kNN classification algorithm? Which is true? kNN plus it is easy to implant implement and robust to noisy and outlier this already we have seen kNN I already yesterday I mean last lecture we have already mentioned that it is very easy to implement implementation is very easy. You just have to find out the distance of the train test data the distance from the test data to all the training data.

Then sort the data in ascending order take the top k data and in the top k how many of them the in the top key data the class which has a major representation the test data will belong to that class. It is that easy implementation these are these are the three four steps. But then it has one disadvantage it has that if there is some noisy data if there is an outlier that is creates gets affected when we try to find out a neighbour.

Some of the funds some insignificant feature some very really occurrence features which actually sometimes this is different becomes the defining factor. So, it gets affected by this sort of outlier. So, that is one of the disadvantage of the kNN algorithm. So, that way a is definitely the right answer. Next let us see the other choices it is easy to implement and affect it largely due to noisy data and outlier.

There is some mistake here it is easy to implement an affect it largely due to noisy data and outlier. It affects but it does not it is robust to noisy data it is not robust so this is not a correct answer I have given here wrong this is not the correct answer. It is not robust to noisy data and outlier it affects it is but it at the same time it is not largely affected also it gets affected. So, I will be the answer will be it is easy to implement.

And it gets affected due to noisy data and outlier it is not largely. It is tough to implement and robust to noisy data and outlier definitely this is wrong. It is tough to implement it is not tough to

implement and affected largely due to noisy this is wrong. So, this three are wrong so it is easy to implement and affected largely this largely is wrong actually but this is the right answer. And a slide I will be changing it so it is easy to implement and its affected due to noisy data and outlier.

But it is not largely affected. If we take a very high k value for low value at low k value of course it is largely affected but if you take a very high k value it is not affected largely, I will be sending changing this in the slide.
(refer time: 18:07)

Then we will be solving some problem and the given data set provides the information of climate condition and rainfall of total 14 days. So, a test data of today is shown below, so what data it is given it is given the climatic condition based on a climatic condition whether there is a rainfall or not rainfall that is given data for 14 such days are given. And then we are given a test data of the climatic condition of today.

Then based on this climatic condition we have to predict whether there will be rainfall or there will not be rainfall. Basically, whether this data falls in the category rainfall or it falls in a category no rainfall that is what we need to find out. Using Bayesian classifier comment about if it will rain today or not, we will be using Bayesian classifier. Bayesians classifier already we have seen it is when the things are not deterministic and certainly, we cannot say with certainty because of this it will happen.

So, because in certain climatic condition we cannot say for certainty that climb it will rain or it will not rain. So, this is sort of in under this sort of situation when this sort of examples comes even here it is mentioned using Bayesian classifier. Even it is not mentioned you can directly go for a Bayesian classifier. When you see there is a when you can solve it using probabilistic value. The type of example where probabilistic value will give us a better solution than product immediately you can you should go for Bayesian classifier.

So, here it is already mentioned Bayesian and this example also shows for Bayesian classifier. Now so this is given, so Bayesian classifier we have already done some example in our class so I think you will be able to do it basically what we have to find out we will have to find out the posterior probability we will have to find out the prior probability. And then definitely Bayesian means will be used base theorem we know the formula for the base theorem.

And then when we know this two term posterior probability and prior probability we will be able to find out the probability of rainfall or probability of no rainfall. We will find out the probability of rainfall, we will find out for this particular data we will find out probability of what is the probability of rainfall, what is the probability of no rainfall. And the probability which is bigger we will consider that as the class for the given data.
(refer time: 20:26)

So, here these are the data given. So, first we will find out the prior probability how will you find out the probability, I think you can remember. Prior probability is because probably we only find

for the classes is not it there are how many classes there are just two classes. So, yesterday remember we found out we were talking about the air traffic data in our last lecture and we have given an example of air traffic data.

Air traffic data based on that there were many classes late very late on time cancel there were four classes. So, we found out the prior probability of those four classes. Always we remember we founded the prior probability of the classes. So, here how many classes two classes, no rainfall, yes rainfall. So, we will find out the prior probability of that so, what is the prior probability of yes first how many data's are here tier total there are 14 data's.

Out on 14 data's how many occurrences of yes if you count you will see there are nine occurrences of yes. So, if there are nine occurrences of yes so prior probability will be 9 / 14, prior probability will be 9 / 14. Similarly, prior probability of rain is equals to no and say out of 49 is yes then definitely 5 will be no so prior probability of rain is equals to noise 5 / 14 so we got the prior probabilities both the class.

Now once we got the prior probability of both the class then next what you have to find out is. Next, we will find out the posterior probability. What is the positive quality how will I find out? So, there are how many attributes climate, humidity, outlook, windy. So, we took the different attributes together. So, first let us consider for climate. When the climate attribute is hot what is the probability that essentially what we are trying to find out what is the probability that attribute climate attribute is hot.

Given that my classes rain what is the probability that my prior probability is my at climate attribute is hot. So, for rain how many hot a day you can just find out. So, here this is hot this line and it is rain and then this is another one only two we have. And total how many range occurrences are there? Total nine already we have seen. So, what will be the posterior probability 2 / 9, similarly we will find out for mild.

Mild how many occurrences before you rain, for this is rain and for hot no rain so no rain hot no rain this is one, this is one there is no more. So, it is how many occurrences of no rain was there it is 5 so it is 2 / 9. Similarly, for mind mild if you see how many occurrences given that it is class is rain how many what is the probability that attribute climate has mild climate. So, mild and yes this is mild and yes this is one and this is one, this is one, this is one, this is one, this is one, total how many 1 2 3 4 we got 4 so 4 / 9.

Similarly, you find out the posterior probability of all these attributes for both the classes rain is equals to yes, rain is equals to no, so we found out the posterior probably for both the classes then we found that the prior probability.
(refer time: 24:03)

Now we will have to find out what is the probability of this class this condition climate is equals to hot, humidity is equals to normal, outlook is equals to sunny, wind is equals to false. So, climate hot humidity in normal so climate hot means basically this one humidity normal then

what was that outlook, sunny and windy falls out looks, sunny, windy is false basically deserted probably 2 / 9 x 6 / 9 x 3 / 9 x 6 / 9 x 9 / 14.

That is for probability that it will rain probably that it will not rain 2 / 5 x 1 / 5 x 3 / 2 / 5 x 2 / 5 x 5 / 14 that is what.
(refer time: 24:55)

So, the class probably have yes given today's attribute we got this no today's attribute we got this, now which is bigger definitely this is bigger. So, this today's attribute we can say that it falls in the class rain means with these attributes it will rain probably it will rain maximum probability that it will rain. That is how we do Bayesian analysis Bayesian classification.
(refer time: 25:22)

So, as per Bayesian classification rain for today is higher compared to the class probability of no rain for today.
(refer time: 25:32)

So, on there is another question for page why I have kept this here if we will see how we have used M estimation technique. We have learned a technique remember M estimation when we have used M estimation when some of the posterior probabilities or some of the class occurrence itself is zero. Some of in our testing data in our observation if some of the class occurrence is zero if the class occurrence is zero definitely the corresponding posture probably will be zero.

And sometimes class occurrences are there but the positive probability is zero. Then how do we calculate that because the whole multiplication of that will give us zero one class totally becomes unrepresented. So, that then we use M estimation approach so this is the problem given where you can use M estimation approach. So, we will quickly go through this because it is the same type of example.

A food delivery application determines the overall quality of a particular delivery depending upon three user ratings delivery time, food quality and delivery person's behaviour. Based on this we give the from food overall quality. These three parameters are rated by as a user after each delivery on a scale of 1 to 4. Depending upon this user ratings level quality of the delivery is class ABC.

Depending upon this rating the overall quality we fall in it falls in three classes that is A B and C, depending on these three variables. The data from 15 such deliveries are given in the table, now suppose for a new delivery the received user ratings are. For new delivery disorder user ratings, now what is the overall quality whether it is A B or C can you predict overall quality of this delivery using naive Bayesian. So, similarly name based on how we have done I will not explain it since the same thing.
(refer time: 27:18)

So, you see using the naive Bayesian what we got here so, when we try to find out probably that class is equals to A for this attributes we got this 0.017 probability that class is equals to B for given this we got value 0 because this one occurrence zero occurrence you see when the for Z = 2 class = B it is nowhere it is given here it is B here z 4 here B it is z 4 here also B z is 3 B z 1 for B there is no occurrence of z 2.

If B has occurred there is no z = 2 that such occurrence is not there at all so zero occurrence what are zero occurrence I got here postal probability probably that behaviour of z = 2 given that B is that I got 0 making that so my class probability has become 0.
(refer time: 28:22)

Similarly, if I try for C also class C also, I will get 0 because of this posterior some of the posterior probability. In other words, there is not enough training data in favour of this cases Bayesian that is one limitation Bayesian we need lots of training data, training data should be used and there is one that is a limitation. The there is not enough training data in favour of this case. This is an issue of name-based classification that can happen due to small size of training data.
(refer time: 28:53)

Now such issue how we can solve it we can solve it using the M estimation. Yesterday we have already discussed what is M estimation so this is the formula for the M estimation, this is the formula for the M estimation. Usually without if we do without M estimation what is our formula P of A is equals to particular variable given a particular class, we just write see basically this is the n c i / n.

We usually do this what is n c i? n c i is the number of training examples from class here that take the value A j = x. And n is the total number of instance from class C i when we consider calculate positive probability we just use this function this expression and C i / n but in under such cases there we may get some zero values as well. So, we have modified it and we have used two different terms n c i + mp / m yesterday we have discussed.

What is m? m is a equivalence m is called the equivalent sample size and p is a user specified parameter, p is a parameter user specified parameter that totally depends on the applications that totally depends upon the number of total sample size class representation there are different various factor based on what we can give the p. If it is a very sensitive, we take a very small p if it is not that sensitive you can take a very bigger p, p is definitely but value will be between 0 to 1.

Now what is the equivalent sample size? Already I have mentioned equivalent sample sizes so, here how many classes we have it is given total we have 15 sample size and we have 3 classes. So, if it would have been a stable data 15 sample size 3 classes if you if it would have been a stable data maybe we would have got 5 samples for each class is not it. If it would have been a sample data that may not be true but when we try to find out the equivalent sample size that way we will talk.

So, here since there are 15 sample size and 3 classes for a for stable data it would have been 5 samples for each class. So, I will consider here my equivalent sample size as 5 so my m = 5.
(refer time: 31:09)

So, in this example p is called the price estimate of the probability in the absence of other information and uniform prior is used. So, the uniform this is one of the heuristic and uniform I can also take as p = 1 / k where is where k is the number of values that attributes can take. So, this is one of the heuristic how we can set k value, k is the number of values that the attributes can take this attributes can take how many values.

So, suppose X can takes 4 values, Y can take suppose 6 values, Z can take suppose say 3 values. Then we will consider a maximum maybe so we can take the k value as 6. So, that way this is one of the heuristic different methods to take k value so k value here we have taken 4 so my p =1 / 4 and M as I already told is a constant equivalent sample size we consider it a 5. This is also arbitrary it is not necessary that you can consider it this way. You can consider equivalent sample size you can that is also very much user dependent.
(refer time: 32:22)

So, now using this value m = 5 p = 1 / 4 we calculated again and we found for class is equals to A we found. So, for class is equals to A we got this is the probability I am not showing the calculation same calculation or we got this probability.
(refer time: 32:40)

For class is equals to B we got this probability.
(refer time: 32:44)

For class is equals to C we got this probability and which one is bigger class A so the given test that is classified as Class A as per naive Bayesian classifier. So, that is how we use M estimation approach.
(refer time: 32:58)

So, now next question the speed and agility ratings for 20 college students and whether they are selected in the college football team is given in the table. Based on the; speed and agility there whether they are selected in the college football team. So, if the speed is high as it is high maybe they are selected or maybe not so this data's are given. Using the kNN technique you can predict that a student with speed 6.75 and agility 3.0 will be selected in the team or not.

So, we will have to find out using the kNN technique. So, kNN technique means our this is our case data, speed = 6.75 as well agility = 3.00 this is our training data. So, that means it has two attributes and this is our test data 6.75 and 3.00 this is our test data does two attributes and this and the table what we have given it is this is our training data it has two attributes. So, in kNN what we try to do? We try to find out the distance of each training data from the test data.

So, we will find the distance of each training data from the test data then we will sort the data in ascending order.
(refer time: 34:09)

So, that is all this is how we found out the distance. Simple Euclidean distance $(x\ 1 - x\ 2)^2 + (y\ 1 - y\ 2)^2$ root that is a two attributes. So, we found out the Euclidean distance this is the value of the equilibrium distance then we have sorted it. Now we will have to take the first k top k. Now what is k here now as the number of neighbours that is value of k to be considered is not given it a problem, we should check the classification output for different values of k.

As it is values for k is not given it is just ask for kNN algorithm. So, we can check for different values of k as it is in my last lecture, I have already mentioned values how to find out the values of k different heuristic I have given. And the best part is that to find out the values of k considering a portion of the training data as if that they are not training data their test data with a class level is not known.

Assuming that assuming a portion of the training data the class level is not known and then we take different k values and we try to find out which k value gives the better choice better classification for this data. Because this data we know the classes we know the classes and we take different k values and we find out what class this data's gives. So, based on the accuracy the for k value which we get highest accuracy.

We can consider that as a k value for the given problem given application that is one heuristic and there are other heuristic as well which I have already mentioned in my last lecture.
(refer time: 35:40)

So, here just for an example maybe different values we have seen different values if I consider different values of k if I can consider value of k as 1, so this is in the sorted order. If I consider value of k as 1 then which one is the; this is which only one is only one means the it is only one neighbour so first neighbour is what first number is that it is in the yes. That means the first neighbour a first number is yes only one number that means my output will also be yes.

If I consider 3, 3 means this 3 out of 3, 2 no 1 yes then my output level will be no 5 then 1 2 3 4 5 out of this 5 how many 3 no 2 yes, so it will be no again likewise.
(refer time: 36:32)

So, that is all it is shown in the problem.
(refer time: 36:35)

Is a same this is I will just see what this Manhattan distance is so considered in following data sets the x and y coordinates of the one of the training points and the corresponding class levels are provided. The test point is given as 4.5 predict the class level of the test point using 3 NN classifier here it is given k = 3 means it is kNN and k instead of k it is given 3, hence k = 3. Using 3 NN classify considering Manhattan distance as the proximity measure.

So, what we have seen in the for kNN we have considered Euclidean distance as a proximity measure this there can be different heuristic to consider the proximately measure, one heuristic to consider proximity measure is Manhattan distance.
(refer time: 37:20)

Manhattan distance is that is how we calculate this is what Manhattan distance. Absolute value of x - x t that is the training data + y - y t x and y are all the training data all the training data and x t and y t is the test data. When so we; find out the absolute values of the difference of the attributes for all the training data. Other previously Euclidean distance what we do we take the square(x 1 - x 2)^2 + (y 1 - y 2)^2 + y 3 - y x 1 - x t test data (x 1 - x 3) ^{2} + (y 1 - y t)^2 plus.

Whatever how many attributes are there accordingly and we do square root, here in case of Manhattan this is the formula we do not do square root which and we do not do square also just absolute value this is how. Based on this stress procedure is same nothing to explain here.
(refer time: 38:14)

So, it is given will be able to see it.
(refer time: 38:16)

So, that is all for today, thank you guys.