INDIAN STATISTICAL INSTITUTE, SQC & OR Unit
Street No.8, Habsiguda, Hyderabad - 500 007

Certification Program on

Business Analytics

2018 Batch

Periodical Examination Answers

Submitted by:
Mallesham Yamulla

1. Analyze whether the following statements are correct or wrong. Justify your answer briefly (maximum Five lines). You need to clearly state "correct or Wrong"

a. As a rule, parametric models will tend to outperform non-parametric approaches when there are a small number of observations per predictor.

Yes, the given statement is correct.

Due to the curse of dimensionality in non-parametric methods, parametric methods will tend to outperform over non parametric.

Let's assume that there are 1000 training observations, here when it's p is 1 it gives the sufficient information to accurately estimate model performance.

on the other hand, spreading these 1000 observations over p=20 dimensions results in curse of dimensionality where the given observation has no near by neighbours.

b. Confidence intervals are always wider than prediction intervals as, they include both, the error in the estimate for f(X) (reducible error) and the uncertainty as to how much an individual point will differ from the population regression plane (irreducible error).

No, the given statement is wrong.

A prediction interval is a range that likely contains the value of the dependent variable for a single new observation given specific values of the independent variables. With this type of interval, we're predicting ranges for individual observations rather than the mean value.

A confidence interval of the prediction is a range that likely contains the mean value of the dependent variable given specific values of the independent variables.

Like regular confidence intervals, these intervals provide a range for the population average.

There is greater uncertainty when we predict an individual value rather than the mean value. therefore, a prediction interval is always wider than the confidence interval of the prediction.

c. R (Cor(X,Y)) can also be used instead of R2 in order to assess the fit of the linear model, as it measures the linear relation between X and Y; in case of both simple linear regression and multiple linear regression.

No, the given statement is wrong.

Two of the most common numerical measures of model fit are the RSE and R2, the fraction of variance explained. These quantities are computed and interpreted in the same fashion as for simple linear regression.

Recall that in simple regression, R2 is the square of the correlation of the response and the variable. In multiple linear regression, it turns out that it equals Cor(Y, Yˆ)2, the square of the correlation between the response and the fitted linear model; in fact one property of the fitted linear model is that it maximizes this correlation among all possible linear models.

d. It is advisable to use Linear Discriminant Analysis or Quadratic Discriminant Analysis as against Logistic Regression when the sample size is very large.

No, the given statement seems to be wrong,

LDA and QDA are often preferred over logistic regression when we have more than two non-ordinal response classes, If n is small and the distribution of the predictors X is approximately normal in each of the classes, the LDA & QDA models yields more stable results than the logistic regression model

e. When we perform cross validation, our goal is to only locate the minimum point in the estimated test MSE curve.

Yes, the given statement is correct.

When we perform cross-validation, our goal might be to determine how well a given statistical learning procedure can be expected to perform on independent data; in this case, the actual estimate of the test MSE is of interest.

But at other times we are interested only in the location of the minimum point in the estimated test MSE curve.

This is because we might be performing cross-validation on a number of statistical learning methods, or on a single method using different levels of flexibility, in order to identify the method that results in the lowest test error.

For this purpose, the location of the minimum point in the estimated test MSE curve is important, but the actual value of the estimated test MSE is not.

f. Test error estimate resulting from LOOCV tends to have higher variance than does the test error estimate resulting from k-fold CV.

Yes, the given statement is correct.

The error estimate from CV is all about trying to get a sense of how well the model of will generalize. With LOOCV, each iteration uses training samples that are very similar, so the models themselves will be incredibly similar. however we will have lower bias because each training sample has more observations.

The training samples in each iteration of k-fold validation are ideally pretty different from one another and from the full training sample, and so the variance is lower.

g. Principal Component Analysis, like linear regression, does not depend on whether the variables have been individually scaled.

No, the given statement is wrong.

PCA is influenced by the magnitude of each variable. therefore, the results obtained when we perform PCA will also depend on whether the variables have been individually scaled.

h. One of the primary reasons for the popularity and adaptability of Business Analytics at present is the urgent need to capture the basic and one-dimensional VOC of the customer.

Yes, the given statement is correct.

To determine the voice of the customer, an organization typically analyzes data reflective of what a customer does as well as what he says. This includes close examination of data gathered from monetary transactions, market research, social media monitoring and customer feedback channels.

i. In big data analytics, as large volumes of data (population) are normally analyzed, statistical methods based on samples such as inferential statistics have become irrelevant and do not play a significant role in decision making.

No, the given statement is wrong.

Inferential analytics takes different theories on the world into account to determine the certain aspects of the large population. When we use inferential analytics, we'll be required to take a smaller sample of information from the population and use that as a basis to infer parameters about the larger population.

Fundamentally, both ML and Statistics work together with data to solve problems

j. General Linear Models and Generalized Linear Models do have the similar objective of providing statistical modelling approaches under varied situations.

For generalized linear models the distribution of residuals is assumed to be Gaussian. If it is not the case, it turns out that the relationship between Y and the model parameters is no longer linear. But if the distribution of residuals is one from the exponential family such as binomial, Poisson, negative binomial, or gamma distributions, there exists some functions of

mean of Y, which has linear relationship with model parameters. This function is called link function.

For example, a binomial residual can use a logit or a probit link function. A Poisson residual uses a log link function.

2. In the context of Data Analytics, briefly differentiate and explain each with at least one real-life example

a. Statistical Learning and Machine Learning
b. Supervised Learning and Unsupervised Learning
c. Parametric Models and Non-Parametric Models
d. Regression Models & Classification Models

Statistical Learning and Machine Learning

Statistical learning consist of a set of tools which will helps us out to look through the data and make use of its insights for the decision making.

With machine learning, we give the machine one or more algorithms for learning how to interpret data. We then feed the machine training data. Using the algorithm and training data, the machine creates a model for interpreting that data.
When given additional data, the machine can adjust its model, thus "learning" to accurately interpret data that's fed to it in the future.

A machine uses training data to develop a model for extracting insight from data or making predictions based on future input. It uses test data to validate the model and fine-tune it.

1. Machine learning is already at work delivering benefits to you in the form of more personalized services,
2. better product recommendations,

3. interactive maps and driving directions,
4. personal digital assistants, anti-spam utilities, and much more.

Machines learn in a variety of ways, including supervised, unsupervised, and semi-supervised learning.

Supervised Learning and Unsupervised Learning

Supervised learning:

In Supervised learning we act like a tutor for the machine, we provide guidelines for the machine to follow along with some training data and we let the machine know when it made a mistake.

In supervised learning we show the machine the connection between a known outcome and the variables that impact that outcome.

Examples:

For example the salary that we get at work depending on the various factors such as experience, education and other skill set so and so forth. So here, salary is referred to as Dependent variable, similarly the factors are referred to as independent variables.

We can predict what/how would be the blocking rate of spams in our network they are classified manually and automatically, what has made it to go up and go down?.

We generally go for supervised learning when we have a general idea of the relationship between independent variables and dependent variables.

Unsupervised learning:

In unsupervised learning, we feed the machine a set of data, and it discovers patterns in that data on its own and figures out its own rules and strategies for interpreting the data.
With machine learning we feed algorithms, unlabeled data and let the machine interpret it. the unlabeled data means that we have one or more independent variables but no dependent variables.

Examples:
In our google photos, the photos would grouped to a place name where a particular photo is taken out. and it shows up a list of photos in a place wise.

Parametric Models and Non-Parametric Models

Parametric Models:

Assumptions can greatly simplify the learning process, but can also limit what can be learned. Algorithms that simplify the function to a known form are called parametric machine learning algorithms.

The algorithms involve two steps:

1. Select a form for the function.
2. Learn the coefficients for the function from the training data.

An easy to understand functional form for the mapping function is a line, as is used in linear regression

$$f(X) = \beta 0 + \beta 1X1 + \beta 2X2 + ... + \beta pXp$$

Where $\beta 0$, $\beta 1$ and $\beta 2$ are the coefficients of the line that control the intercept and slope, and X1 and X2 are two input variables.

Assuming the functional form of a line greatly simplifies the learning process. Now, all we need to do is estimate the coefficients of the line equation and we have a predictive model for the problem.

Often the assumed functional form is a linear combination of the input variables and as such parametric machine learning algorithms are often also called “linear machine learning algorithms“.

The problem is, the actual unknown underlying function may not be a linear function like a line. It could be almost a line and require some minor transformation of the input data to work right. Or it could be nothing like a line in which case the assumption is wrong and the approach will produce poor results.

Some more examples of parametric machine learning algorithms include:

1. Logistic Regression
2. Linear Discriminant Analysis
3. Perceptron
4. Naive Bayes
5. Simple Neural Networks

Benefits of Parametric Machine Learning Algorithms:

1. Simpler: These methods are easier to understand and interpret results.
2. Speed: Parametric models are very fast to learn from data.
3. Less Data: They do not require as much training data and can work well even if the fit to the data is not perfect.

Limitations of Parametric Machine Learning Algorithms:

1. Constrained: By choosing a functional form these methods are highly constrained to the specified form.
2. Limited Complexity: The methods are more suited to simpler problems.
3. Poor Fit: In practice the methods are unlikely to match the underlying mapping function

Non- Parametric Models:

Algorithms that do not make strong assumptions about the form of the mapping function are called nonparametric machine learning algorithms. By not making assumptions, they are free to learn any functional form from the training data.

Nonparametric methods seek to best fit the training data in constructing the mapping function, whilst maintaining some ability to generalize to unseen data. As such, they are able to fit a large number of functional forms.

An easy to understand nonparametric model is the k-nearest neighbors algorithm that makes predictions based on the k most similar training patterns for a new data instance. The method does not assume anything about the form of the mapping function other than patterns that are close are likely have a similar output variable.
Some more examples of popular nonparametric machine learning algorithms are:

1. k-Nearest Neighbors
2. Decision Trees like CART and C4.5
3. Support Vector Machines

Benefits of Nonparametric Machine Learning Algorithms:

1. Flexibility: Capable of fitting a large number of functional forms.
2. Power: No assumptions about the underlying function.
3. Performance: Can result in higher performance models for prediction.

Limitations of Nonparametric Machine Learning Algorithms:

1. More data: Require a lot more training data to estimate the mapping function.
2. Slower: A lot slower to train as they often have far more parameters to train.
3. Overfitting: More of a risk to overfit the training data and it is harder to explain why specific predictions are made.

Regression Models & Classification Models

In regression problems, the machine tries to come up with an approximation based on the data input. For example, we may have a machine learning program that predicts stock prices. Here, the answer to the question isn't a specific class or category, it's a value in a continuous range of values.

In classification problems the machine tries to figure out in which group a new input belongs. Classification problems are further divided into binary and multi-class classification. with binary classification the machine has only two class such as Spam or Not, Machines has more than two classes such as satisfaction ratings satisfied, very satisfied, dissatisfied

Introduction to BIAS and Variance,the different methods adopted in model validation

In supervised machine learning an algorithm learns a model from training data.

The goal of any supervised machine learning algorithm is to best estimate the mapping function (f) for the output variable (Y) given the input data (X). The mapping function is often called the target function because it is the function that a given supervised machine learning algorithm aims to approximate.

The prediction error for any machine learning algorithm can be broken down into three parts:

1. Bias Error
2. Variance Error
3. Irreducible Error

The irreducible error cannot be reduced regardless of what algorithm is used. It is the error introduced from the chosen framing of the problem and may be caused by factors like unknown variables that influence the mapping of the input variables to the output variable.

Bias Error

Bias are the simplifying assumptions made by a model to make the target function easier to learn.

Generally, parametric algorithms have a high bias making them fast to learn and easier to understand but generally less flexible. In turn, they have lower predictive performance on complex problems that fail to meet the simplifying assumptions of the algorithms bias.

1. Low Bias: Suggests less assumptions about the form of the target function.
2. High-Bias: Suggests more assumptions about the form of the target function.

Examples of low-bias machine learning algorithms include: Decision Trees, k-Nearest Neighbors and Support Vector Machines.

Examples of high-bias machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.

[1]

Variance Error

Variance is the amount that the estimate of the target function will change if different training data was used.

The target function is estimated from the training data by a machine learning algorithm, so we should expect the algorithm to have some variance. Ideally, it should not change too much

[1] Indian Statistical Inst-Hyderabad, Certification Program on Business Analytics-2018

from one training dataset to the next, meaning that the algorithm is good at picking out the hidden underlying mapping between the inputs and the output variables.

Machine learning algorithms that have a high variance are strongly influenced by the specifics of the training data. This means that the specifics of the training have influences the number and types of parameters used to characterize the mapping function.

1. Low Variance: Suggests small changes to the estimate of the target function with changes to the training dataset.
2. High Variance: Suggests large changes to the estimate of the target function with changes to the training dataset.

Generally, nonparametric machine learning algorithms that have a lot of flexibility have a high variance. For example, decision trees have a high variance, that is even higher if the trees are not pruned before use.

Examples of low-variance machine learning algorithms include: Linear Regression, Linear Discriminant Analysis and Logistic Regression.

Examples of high-variance machine learning algorithms include: Decision Trees, k-Nearest Neighbors and Support Vector Machines.

Bias-Variance Trade-Off

The goal of any supervised machine learning algorithm is to achieve low bias and low variance. In turn the algorithm should achieve good prediction performance.

You can see a general trend in the examples above:

1. Parametric or linear machine learning algorithms often have a high bias but a low variance.

2. Non-parametric or non-linear machine learning algorithms often have a low bias but a high variance.

The parameterization of machine learning algorithms is often a battle to balance out bias and variance.

Below are two examples of configuring the bias-variance trade-off for specific algorithms:

1. The k-nearest neighbors algorithm has low bias and high variance, but the trade-off can be changed by increasing the value of k which increases the number of neighbors that contribute t the prediction and in turn increases the bias of the model.
2. The support vector machine algorithm has low bias and high variance, but the trade-off can be changed by increasing the C parameter that influences the number of violations of the margin allowed in the training data which increases the bias but decreases the variance.

There is no escaping the relationship between bias and variance in machine learning.

1. Increasing the bias will decrease the variance.
2. Increasing the variance will decrease the bias.

There is a trade-off at play between these two concerns and the algorithms you choose and the way you choose to configure them are finding different balances in this trade-off for your problem.

In reality, we cannot calculate the real bias and variance error terms because we do not know the actual underlying target function. Nevertheless, as a framework, bias and variance provide the tools to understand the behavior of machine learning algorithms in the pursuit of predictive performance.

Model validation methods:

Validation set approach:

The validation set approach is a very simple strategy for this task. It involves randomly dividing the available set of observations into two parts, a training set and a validation set or hold-out set.
The model is fit on the training set, and the fitted model is used to predict the responses for the observations in the validation set.
The resulting validation set error rate—typically assessed using MSE in the case of a quantitative response—provides an estimate of the test error rate.

The validation set approach is conceptually simple and is easy to implement. But it has two potential drawbacks as given below.

1. The validation estimate of the test error rate can be highly variable, depending on precisely which observations are included in the training set and which observations are included in the validation set.

2. In the validation approach, only a subset of the observations—those that are included in the training set rather than in the validation set—are used to fit the model. Since statistical methods tend to perform worse when trained on fewer observations, this suggests that the validation set error rate may tend to overestimate the test error rate for the model fit on the entire data set.

Leave-One-Out Cross-Validation:

1. Leave-one-out cross validation is K-fold cross validation taken to its logical extreme, with K equal to N, the number of data points in the set. That means that N separate times, the function approximator is trained on all the data except for one point and a prediction is made for that point.

2. As before the average error is computed and used to evaluate the model. The evaluation given by leave-one-out cross validation error (LOO-XVE) is good, but at first pass it seems very expensive to compute. Fortunately, locally weighted learners can make LOO predictions just as easily as they make regular predictions.

3. That means computing the LOO-XVE takes no more time than computing the residual error and it is a much better way to evaluate models. We will see shortly that Vizier relies heavily on LOO-XVE to choose its metacodes.

K-fold cross validation :

1. K-fold cross validation is one way to improve over the holdout method. The data set is divided into *k* subsets, and the holdout method is repeated *k* times.
2. Each time, one of the *k* subsets is used as the test set and the other *k-1* subsets are put together to form a training set. Then the average error across all *k* trials is computed.
3. The advantage of this method is that it matters less how the data gets divided. Every data point gets to be in a test set exactly once, and gets to be in a training set *k-1* times.
4. The variance of the resulting estimate is reduced as *k* is increased. The disadvantage of this method is that the training algorithm has to be rerun from scratch *k* times, which means it takes *k* times as much computation to make an evaluation.

5. A variant of this method is to randomly divide the data into a test and training set k different times.
6. The advantage of doing this is that you can independently choose how large each test set is and how many trials you average over.