A Short Monograph on Bias-variance and cross-validation

TO SERVE AS A REFRESHER FOR MODEL TUNING

Index

Contents

1. Introduction 4

1.1 Introduction to Predictive Modelling 4

1.2 Supervised and Unsupervised Learning 5

2. The Problem of Prediction: Bias Variance Trade-off 7

2.1 Bias of a model 7

2.2 Variance of a model 8

2.3 Bias Variance Trade-off 9

2.4 Training and Test data sets 10

2.5 Cross-Validation 10

List of Figures

Fig. 1: Statistical Learning Flowchart........6

Fig. 2: Overfitting vs Underfitting vs Ideal Model........8

1. Introduction

1.1 Introduction to Predictive Modelling

Technology has given us the power to capture data from almost everything – road accidents, grocery store bar codes, customer loyalty programs as well as political opinions expressed on Twitter and Facebook. However, amassing data does not help us in any way unless we can understand the information contained in the data. Predictive modeling and data mining help to detect the hidden patterns in the data and to forecast for yet unobserved situations.

In this monograph and few subsequent monographs, various topics in predictive modeling and data mining will be taken up But before we go into the details of predictive modeling, several major concepts need to be addressed.

What is the objective of predictive modeling?

Let us examine two different cases. Identification of spam or detection of fraudulent credit card transactions is two situations where the application of predictive modeling is important. In the former case, the objective is to detect spam email through a filter and classify it as such. In the latter case, the banks would want to identify frauds immediately and can flag the transactions. In both cases, prediction accuracy needs to be very high, though how the spam was filtered or fraud was detected may not be so important. In this case, the model may be complex and the interpretability of the model low. Often these models are known as ‘black box’ models. Automation will work well here.

Consider another case where a physician needs to predict whether a post-menopausal woman above 65 years of age is at a risk of knee replacement surgery given various other health indicators. Here a ‘black box’ model may not be acceptable despite having high accuracy. The reason being, it is not enough to know who is at risk but it is mandatory to mitigate her risk of knee replacement. Unless a physician can understand which health indicators are more important, she will not be able to provide an effective treatment regime for her patient. Interpretability is a must in this situation. A black box model may be rejected in favor of an interpretable model, albeit accuracy in the former may be higher.

That is not to say predictive accuracy needs to be sacrificed totally to improve interpretability. Somewhere a balance needs to be struck. The suitability of the predictive model is an important issue that may need to be addressed case by case.

Predictive Analytics or Predictive Modeling is a process of extracting information from large complex data sets using a variety of computational tools to make predictions and estimates about future outcomes.

1.2 Supervised and Unsupervised Learning

All problems of data mining, pattern recognition, or predictive modeling come under the umbrella known as *Statistical Learning*.

Statistical learning problems can be partitioned into *Supervised* or *Unsupervised Learning*.

The problems discussed in the previous section belong to the first category. The objective in each case is to predict a response, typically denoted by Y. Corresponding to each unit of observation there is a set of independent variables or predictors (X), based on which Y is estimated (see the monograph on Regression). Whenever in the data set the response is available, the problem falls under supervised learning.

Supervised learning can be further divided into two classes, depending on the nature of the response Y. If Y is a continuous variable, the problem falls under the category Regression. On the other hand, if the response is qualitative, binary, or multi-class, the problem falls under the category Classification. Spam identification (Yes/No) and detection of fraud (Yes/No) are both classification problems.

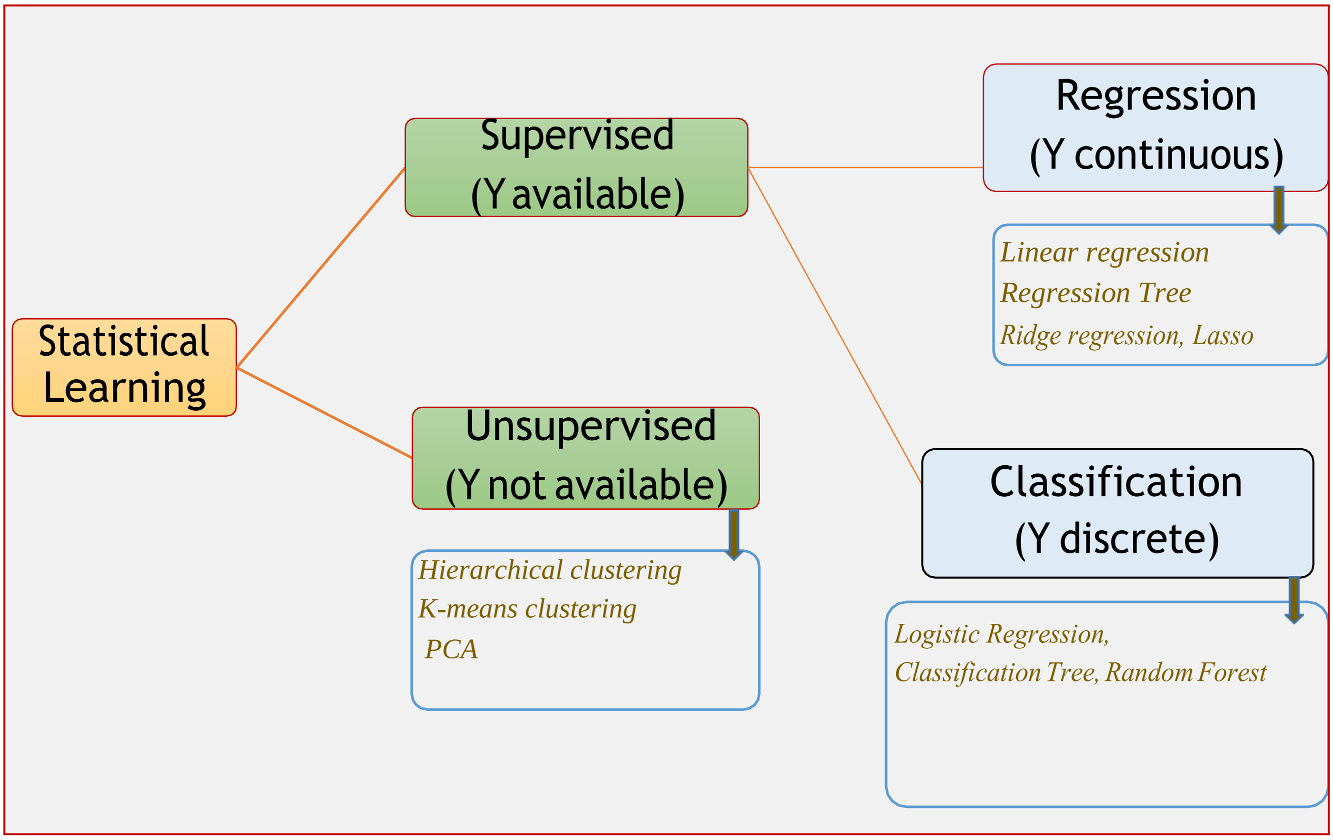
The problem of assigning a risk value to a patient may be considered a classification problem if the risk is defined as low, medium, or high. If, however, a continuous risk probability is to be estimated for each patient, the problem is considered a regression problem.

Unsupervised learning problems are those where there is no response. One example of an unsupervised learning problem is to categorize loyalty customers in a Gold, Silver, and Bronze classification, depending on their propensity of spending in a store. Detection of possible clusters in a multivariate data set is an unsupervised learning problem.

An illustrative figure is shown on the next page with a few techniques of different types of learning.

A few examples of classification and regression techniques. By no means the examples provided are comprehensive.

(Special note: *Random Forest may also be used where the response is continuous and logistic regression may also include lasso regularization. However, for initial understanding these examples are illustrative*.)



2. The Problem of Prediction: Bias Variance Trade-off

However accurate the prediction models are, there will always be a *prediction error*, the difference between the predicted and actual observations. Recall that for regression this error is also known as residual. Prediction error has two components: *bias* and *variance*. To minimize prediction error, both bias and variance need to be at a low level. Unfortunately, however, both bias and variance cannot be minimized simultaneously.

Identification of a suitable prediction model involves a trade-off between its bias and variance.

Gaining a proper understanding of bias and variance would help in avoiding the mistakes of overfitting or under-fitting. The problem of bias and variance are intimately associated with the problem of under-fitting and overfitting. A good model is one for which the bias and variance are as small as possible, and for which, the predictive ability is good.

The notions of bias and variance of a model are explained below.

Let us consider a model $Y = (X)+\epsilon$, where f is an arbitrary function of the independent variables X and ϵ is the random error component. (Most models in predictive problems are of this type). The function f may be completely unknown or one may only have partial knowledge about its form. For example, suppose f to be linear, i.e. $(x) = a+bx$. Unless the numerical values of a or b are known, only the type of the function f is known but not the complete function. Let an estimate $\hat{f}$ of f is obtained. For any given value of X, the predicted value $\hat{Y}=\hat{f}\,(X)$. Usually, $\hat{f}$ is computed using the data y, and hence $\hat{f}\,(x)$ is a random variable.

All the subsequent definitions and explanations in the next three sections will be based on this model.

2.1 Bias of a model

Bias is the average difference between the predicted values of a model and the observed values.

$$Bias=E(\hat{f}\,(x)-f(x))$$

where E denotes expectation or mean. A model is called unbiased if $Bias=0$. Bias can be both negative and positive, so a desirable condition for bias is that the absolute value of Bias is close to 0.

A model with high bias pays very little attention to the dataset and oversimplifies the model. This phenomenon is called **under-fitting**. Under-fitting may happen for various reasons. The most common reasons include having only a limited amount of data to model a complex structure or fitting a linear model to non-linear data. An oversimplified model exhibits small variability.

2.2 Variance of a model

Variance is defined to be the variability of predicted values which quantifies the instability of the model.
Formally,

$$Variance=(\hat{f}\ (x)-E(\hat{f}\ (x)))^2$$

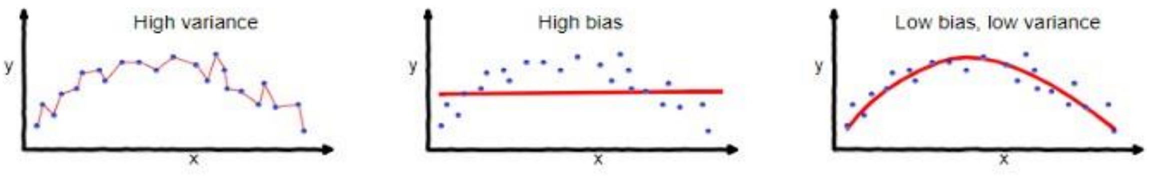
This is the error variance and is often denoted by the mean square error (MSE). Unlike bias, the variance can never be negative.

A model with high variance pays a lot of attention to the available data. Such models perform very well on data used to fit the model but have limited power to predict for the data that has not been observed. Hence it has high error rates on data used for prediction. This phenomenon is called **overfitting**. Overfitting happens when a model is too close to the observed data and captures the noise along with the underlying pattern in data.

Let us demonstrate the concepts of bias and variance through a visualization.

Observe n paired data points $(x_1, y_1),(x_2,y_2),\ldots,(x_n,y_n)$. The model $y = (x) + \epsilon$ is fitted to the data and the estimate $\hat{f}$ is obtained. If a new observation x is one of x_1, $x_2,\ldots$, then $\hat{f}$ can be chosen so that $E(\hat{f}\ (x)) = f(x)$ exactly and hence the bias is zero when evaluated at the observed x value. But if x is different from x_1, $x_2,\ldots$, or x_n, then our estimate $\hat{f}$ may behave poorly. The reason is that $\hat{f}$ becomes very short-sighted since it tries to fit the observed data (x_i,y_i) as perfectly as possible, and does not take into consideration any new data point which could be used in the future.

The following picture illustrates these concepts.



overfitting

underfitting

Good balance

The leftmost panel shows an almost perfect fit to the observed dataset and thus bias is very small. But for this model, the variability is very high. As soon as one new observation is added to the data, the function *f(x)* may change considerably.

The middle panel shows a model that completely ignores the curvature in the dataset and plots a straight line through it. Clearly, it does not fit the data at all. This model varies only slightly for a different sample and the impact of the addition of several points may be negligible. Though in theory variance of a model cannot be zero, for all practical purposes it may be considered such. But the bias for this model is very high.

The rightmost panel shows a much better fit. It does not have zero bias, nor does it enjoy negligible variance, but it strikes a balance somewhere between the two and also fits the data well.

The objective of predictive modelling is to find such an optimum f(x).

2.3 Bias Variance Trade-off

If a model is too simple and has only a few parameters, then it may have high bias and low variance. On the other hand, if a model has a large number of parameters then it is likely to have high variance and low bias. So it is essential to find the right balance without overfitting or under-fitting the data. This trade-off between bias and variance is closely associated with a trade-off in the complexity of the model.

The total squared error of a model can be expressed as

$$Total\ squared\ error = Bias^2 + Variance + Irreducible\ error$$

It is not possible for any statistical model to manipulate irreducible error, which is inherent in the data.

The goal of predictive modeling is to reduce $Bias^2 + Variance$. As described above, due to the bias-variance trade-off, no model can reduce both bias and variance simultaneously, and therefore, a good model will try to minimize the sum of $Bias^2 + Variance$.

In general, an unbiased model with the smallest possible variance is preferred. However, in special situations, a biased model is deliberately chosen for which $Bias^2 + Variance$ is smaller than any unbiased model.

Such a model has the best predictive power, i.e. it is able to provide the best possible estimated value for a new observation.

2.4 Training and Test datasets

The discussion above establishes an important fact. Unless the predictive ability of a model is tested on an independent data set, which is different than the one used to build the model, a vital aspect of the model is ignored. This necessitates splitting of the existing data set into two or more parts.

Training data: A training dataset is used to fit one or more models and estimate their parameters.

Test data: A test dataset is used to assess the performance of the developed model. The test set should be as close as possible to the training dataset (more formally, having the same distribution) but no overlap with the training dataset.

Typically training and test data are partitions of the observed data into a random 80:20 split. Other possible splits may be 75:25 or 70:30 or in some other ratio, all taken randomly. If the data is very large even 50:50 split into training and test is also possible. Training data is larger so that the model parameters are estimated with considerable accuracy. The purpose of having test data is to check how close the predicted values are to the observed values.

Choice of test data may also be modified according to special applications. In certain predictive methods that are intended to be applied for a while(e.g. credit scoring), the test data is taken to be the most recent period. For example: Q1, Q2, and Q3 data are used in the training set, but Q4 data is used as the test set for a model that is intended to project credit risk for the next quarter. In time series, one of the most specialized predictive models, only the most recent periods are used as test data.

In the test data set no parameter estimation is performed.

2.5 Cross Validation

One criticism of training and test data split is that the proposed model may depend on the split since the split is done once only. One suggested alternative is to perform k-fold cross-validation, which is an extension of the train-test split.

k-fold cross-validation is a method of getting multiple sets of training and test data out of the original data set. The steps are as follows.

i) Split the train dataset randomly into k equal (or almost equal) parts.

ii) Choose one of these k parts as the test dataset and the other $k-1$ parts together as the training dataset.

iii) Fit the model on the training dataset thus obtained and assess its performance on the test dataset. Usually, one predicts the values in the test dataset using this model and takes the square root of the error sum of squares (RMSE), but other measures of prediction errors are also possible

iv) Repeat this process k times, once for each of the k splits as the test data, and the complementary set as the training data

v) Finally, take an average of all the k RMSEs to get a final estimate of prediction error.

The most important advantage of k-fold cross-validation is that each data point is included at least once in the training and in the test data.

The value of k is usually taken to be 10. But due to computational complexity k = 5 is also a common choice. Note that the higher the value of k is, the smaller is the size of test data. That should be another consideration is choosing an optimum value of k.