Introduction to Statistical Learning Model Jan. 30th, 2021
Boxi Huang's Reading Note bh563@cornell.edu

Page Covered: PP 21-37

Statistical Learning:

statistical learning refers to a set of approaches for estimating f (The True Function)

Why Estimate f (AKA, the true function)?

There are two main reasons for estimating the function f: (1) Prediction (2) Inference

Prediction:

In many situations, a set of input X is commonly available, but not the output Y. So we can generate Y^hat = f^hat (X) under the condition that the error term averages to zero.

Essentially, we want to provide/create a mechanism to predict future Y using X. (Supervised Learning heavily happens here)

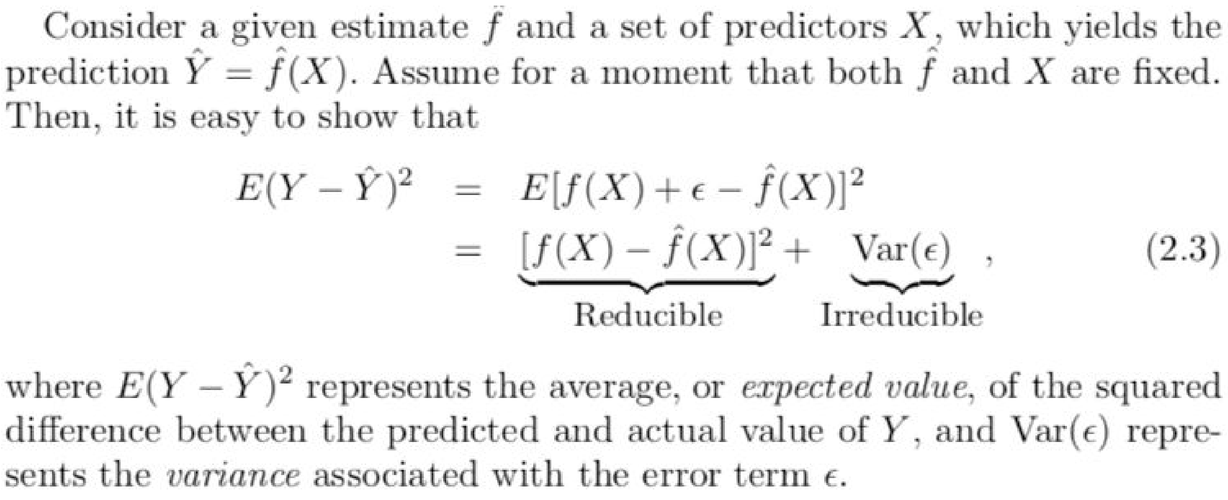
The accuracy of Y^hat as a prediction for Y depends on two quantities: Reducible & Irreducible Errors.

Reducible Errors:

1. We can lower this error by choosing the most appropriate statistical learning technique to estimate f.
2. We can lower this error by further training models with more iterations and greater datasets.

Irreducible Errors:

1. Due to the fact that each datapoint contains variability, which can lead to inaccurate model training (Slightly off coefficient for linear regression, etc), we are not able to create a model that is 100% identical to the true model. Because variability within data points can not be reduced, this is an irreducible error.
2. Since there is no way to access the true model, it is also difficult for us to reduce the difference/gap between our y^hat and y.



Consider a given estimate $\hat{f}$ and a set of predictors X, which yields the prediction $\hat{Y} = \hat{f}(X)$. Assume for a moment that both $\hat{f}$ and X are fixed. Then, it is easy to show that

$$\begin{aligned} E(Y - \hat{Y})^2 &= E[f(X) + \epsilon - \hat{f}(X)]^2 \\ &= \underbrace{[f(X) - \hat{f}(X)]^2}_{\text{Reducible}} + \underbrace{\mathrm{Var}(\epsilon)}_{\text{Irreducible}} \ , \end{aligned} \tag{2.3}$$

where $E(Y - \hat{Y})^2$ represents the average, or *expected value*, of the squared difference between the predicted and actual value of Y, and $\mathrm{Var}(\epsilon)$ represents the *variance* associated with the error term ϵ.

The book is mostly focused on minimizing the reducible errors in prediction. Keep in mind that irreducible error will always provide an upper bound on the accuracy of our prediction for Y.

Inference:
Instead of making a prediction for Y, we want to understand the relationship between X and Y; Or namely, we want to understand how Y changes as a function of X1...Xp. So we are actually interested in f^hat 's exact form.

A few questions are usually asked in inference:

- Which predictors are associated with the response?
- What is the relationship between the response and each predictor?
- Can the relationship between Y and each predictor be adequately summarized using a linear equation? Or is the relationship more complicated?

How do we estimate f?

Our goal is to apply a statistical learning method to the training data in order to estimate the unknown function f. Namely, we want to find a function f^hat such that Y ~= f^hat (X) for any observation (X, Y).

This "How Question", can be characterized into (1) Parametric (2) Non-parametric methods.

Parametric Method

(1) We make an assumption about the model (functional form, shape) and settle to it.

(2) After a model has been selected, we need a procedure that uses the training data to fit or train the model. In the case of linear regression, we need to estimate the parameters vectors B0, B1,...Bp.

The model-based approach is referred to as parametric, which reduces the problem of estimating f down to one of estimating a set of parameters. This reduces the complexity and ambiguity of the problems to a set of parameters we can track and modify further.

Disadvantage: The assumptions we made for model shape/functional form might not be an appropriate method/ or might not be fully developed to estimate the true function, which might create large reducible errors.

Solution: Chose a more flexible model requires estimating a greater number of parameters. (Can lead to overfitting if too flexible)

Non-parametric Methods
The non-parametric method approaches the problem by seeking an estimate of f that gets as close to the data points as possible without being too rough. In this way, the non-parametric method eliminates the assumption part (Better than the parametric method) and has the ability to fit a wide range of possible shapes of f.

Disadvantages:

(1) Computation expensive
(2) Slow
(3) Do not reduce to a small number of parameters, so a large number of observations (data) is required.

Supervised Learning vs Unsupervised Learning

Supervised Learning:
For each observation of the predictor measurement(s), x_i, i = 1...n. There is an associated response measurement y_i. We wish to fit a model that relates the response to the predictors, with the aim of accurately predicting the response for future observations (prediction) or better understanding the relationship between the response and the predictors (Inference).

Ex: Linear Regression, Logistic Regression, GAM, boosting, SVM...more

Unsupervised Learning: Describe the situation in which for every observation i= 1,...n, we observe a vector of measurements x_i, but no associated response y_i. We lack a response variable that can supervise our analysis, therefore, we seek to understand the relationships between variables or between observations.

Ex: Clustering Analysis

Variables can be classified into two types: (1) Quantitative (2) Qualitative

(1) Quantitative variables takes on numerical values, such as ages, height, income, monetary value or price. We tend to refers to problems with a quantitative response as **regression problems.**

(2) Qualitative variables take on values in one of K different classes, or categories, such as gender, brans, whether default on loan, or cancer diagnosis. We tend to refers to problems with a qualitative response as **classification problems.**

The distinction between regression and classification is not crispy and so we can always encoding (OHE) our qualitative data.

Assessing Model Accuracy

Measuring the Quality of Fit

In order to quantify the extent to which the predicted response value for a given observation is close to the true response value for that observation. In regression, the most commonly-used measure is the mean squared error (MSE)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2,$$

MSE will be small if the predicted response are very close to the true response, and large if it is far. We are most often interested in lowering the test MSE and we'd like to select the model for which the average of the square of the difference between predicted response and true response is minimized.

By adjusting the flexibility of the models (Adding more features to predict, increase dimensions of the features, etc), we can create multiple fits of the data. The degree of freedome is a quantity that summarizes the flexibility of a curve.

As we increases the flexibility of the model, the test MSE initially decreases but then level off and start increasing again. On the other hand, training MSE will monotonically decrease.

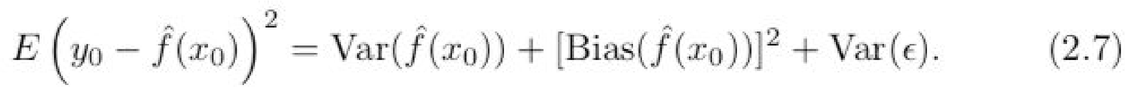
From the first point that test MSE starts to increase but training MSE decreases, the model is **OVERFITTING.**

Overfiting: The model is working too hard to find the patterns in teh training data and start to pick up some patterns that are just caused by random chance rather than by true properties of the unknow function f, ex: Noise.
Solution: (1) Increasing dataset size (2) Readjust Flexibility by conducting variable selections (3) Switch a model to perform

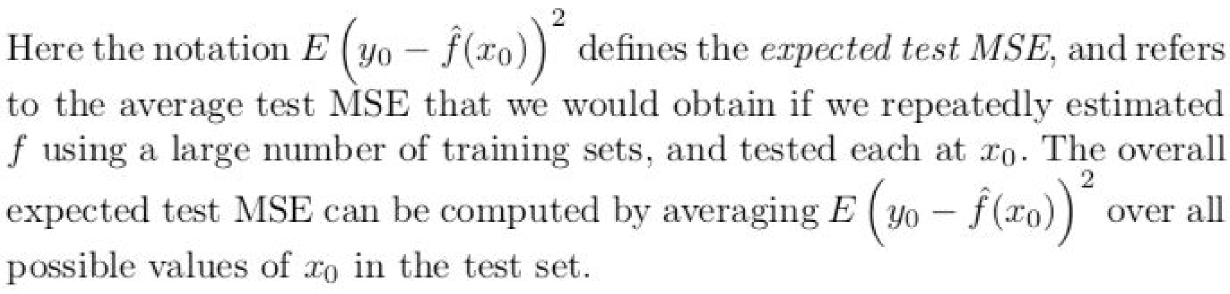
Cross Validation is a method for estimating test MSE using the training data.

Bias-Variance Trade-Off

Expected test MSE, for a given value X0, can always be decomposed into the :sume of three fundamental quantities: the variance, the squared bias of f^(x0) and the variance of the error terms e.



$$E\left(y_0 - \hat{f}(x_0)\right)^2 = \mathrm{Var}(\hat{f}(x_0)) + [\mathrm{Bias}(\hat{f}(x_0))]^2 + \mathrm{Var}(\epsilon). \qquad (2.7)$$



Here the notation $E\left(y_0 - \hat{f}(x_0)\right)^2$ defines the *expected test MSE*, and refers to the average test MSE that we would obtain if we repeatedly estimated f using a large number of training sets, and tested each at x_0. The overall expected test MSE can be computed by averaging $E\left(y_0 - \hat{f}(x_0)\right)^2$ over all possible values of x_0 in the test set.

Goal: Achieve a model that has low variance and low bias.
Insight: Since both variance and square of bias is non-negative, we know that expected test MSE can never lie below the irreducible error.

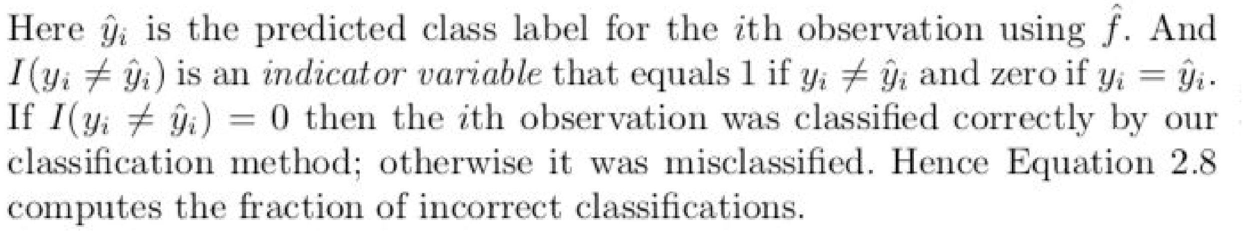
Variance: refers to the amount by which f^ would change if we estimated it using a different training data set. A highly flexible model tends to have high variance.

Bias: refers to the error that is introduced by approximating a real-life problem, which maybe very complicated, by a much simpler model.

Increase Flexibility → Increase Variance, Decrease Bias
Initially, bias decrease faster than variance increase, approach the global optimal point. Once we pass this point, bias barely decrease and variance increase significantly.

Bias-Variance Trade-off in Classification
The most common approach for quantifying the accuracy of our estimate f^hat is the training error rate, the proportion of mistakes that are made if we apply our estimate f^hat to the training observations:

$$\frac{1}{n}\sum_{i=1}^{n} I(y_i \neq \hat{y}_i). \tag{2.8}$$



Here $\hat{y}_i$ is the predicted class label for the ith observation using $\hat{f}$. And $I(y_i \neq \hat{y}_i)$ is an *indicator variable* that equals 1 if $y_i \neq \hat{y}_i$ and zero if $y_i = \hat{y}_i$. If $I(y_i \neq \hat{y}_i) = 0$ then the ith observation was classified correctly by our classification method; otherwise it was misclassified. Hence Equation 2.8 computes the fraction of incorrect classifications.

The Bayes Classifier
The point is to create a simple classifier that assigns each observation to the most likely class, given its predictor values. Namely, we should simply assign a test observation with predictor vector x0 to the class j for which

$$\Pr(Y = j | X = x_0)$$

Is largest. In a two-class problem when we have two classes, class 1 and class 2. Bayes Classifier corresponds to predicting class 1 if P(Y= 1 | X = x0) > 0.5 and class 2 otherwise.

Bayes Decision Boundary: The line that the conditional probability is exactly 50%
Bayes Error Rate: produces the lowest possible test error rate.

$$\Pr(Y = j | X = x_0)$$

In real data, it is hard to obtain conditional distribution of Y given X, so Bayes classifier is the unattainable gold standard.

Therefore, most classifiers try to estimate the conditional distribution of Y given X, and classify a given observation to the class with highest estimated probability.

K-nearest neighbor classifier is one of those classifiers.
KNN classifier first identifies the K points in the training data that are closest to x0, represented by N0. It then estimates the conditional probability for class j as the fraction of points in N0 whose response values equal j:

$$\Pr(Y = j | X = x_0) = \frac{1}{K} \sum_{i \in \mathcal{N}_0} I(y_i = j).$$

Finally, KNN applies Bayes rule and classifies test observation x0 to the class with the largest probability.

If K is small, the classifier model is too flexible and thus, has a high variance but low bias. As K grows, the method is more restricted, resulting in a low-variance but high-bias classifier.