Overview of Statistical Learning

Seung Hyun Sung

11/4/2021

Contents

> In essence, statistical learning refers to a set of approaches for estimating f. In this chapter we outline some of the key theortical concepts that arise in estimating f, as well as tools for evaluating the estimates obtained. — pg 17 *An Introduction to Statistical Learning with Apllications in R.*

1A Introduction to Regression Models

Depending on the family of Regression model and its complexity of $f(x)$, we may be able to understand how each compnent X_j affects Y, in what particular fashion. Depending on the task (target variable Y) will vary in weight of importance between the interpretation and accuracy. Hence, the phase we are with predicting or defining the task it is important to take this in account before designing and selecting a model.

The notation for ideal Regression function:

$$f(x) = E(Y|X = x)$$

Immediately it emphasis that this regression function gives conditional expectation of $Y|X$. Is there an ideal $f(X)$?

1A.1 Ideal regression function

The ideal regression function means the expected value (average) of Y given X. In section 1B we will explore more on this conditional expectation, its useful properties, and geometric interpretation.

For example, if X had three components $x \in R^3$ It is going to be a conditional expectation of Y given three particular instances of these three components of X.

$$f(x) = f(x_1, x_2, x_3) = E(Y|X_1 = x_1, X_2 = x_2, X_3 = x_3)$$

The question given to the function is that at particular point X with three coordinates, X_1, X_2, X_3, what is good value for the function at that point of instances. We assume that there is some relationship between Y and covariates X.

Meaning of the ideal or optimal regression function:

- Conditional Average - $E(Y)$ would be the **averages** of $Y's$ at these coordinates, and the regression function would do that at all points in the plane.
- Ideal means with regard to a loss function, the particular choice of the function $f(x)$ will minimise the sum of squared errors.
- $Y = f(X) + \epsilon$ where ϵ is a random error term (also referred as irreducible error), which is independent of X and has mean of zero.
  - linear regression assumption: $\epsilon \sim$ i.i.d. N(0, σ^2)

$f(x) = E(Y|X = x)$ *is the function that minimises* $E((Y - g(X))^2|X = x)$ *over all functions* g *at all points* $X = x$

1A.2 Reducible and irreducible error

At each point X , there will be mistakes. Even with what we call ideal or optimal predictor of Y with regard for the function, by nature of real world data, there will be a certain amount of noise presence on the ϵ distribution. We must remember that the function learns and predict basis of conditional averages of all $Y's$ given each X coordinates ($E(Y - \hat{Y})^2$ represents average of squared difference between predicted and actual value of Y). By means, despite the fit being ideal function f to the regression model, it does not mean that prediction will be perfect.

The expected squared error at a point x is:

$$Err(x) = E[(Y - \hat{f}(x))^2]$$

The Err(x) can be further decomposed as

$$Err(x) = (E[\hat{f}(x)] - f(x))^2 + E[(\hat{f} - E[\hat{f}(x)])^2] + \sigma_e^2$$

$$Err(x) = (f(x) - \hat{f}(x))^2 + Var(\epsilon)$$

$$\therefore \; Err(x) = Bias^2 + Variance + Irreducible\ Error$$

The accuracy of $\hat{Y}$ as a prediction for Y depends on the two quantities: reducible and irreducible error.

In general, $\hat{f}$ will not be a perfect estimate for f, and this inaccuracy will introduce reducible error. This part of error we can potentially improve by uising the most appropriate statistical learning technique to estimate f.

However, even we our best estimate, there will be some left out residuals, irreducible error.

"Y is also function of ϵ, which by definition can not be predicted using X." This irreducible error can be larger than zero, as it may contain unmeasured variables that are useful in predicting Y. The only option to reduce such unmeasured quantity is by adding external variables. Establishing the independency of $E(Y - \hat{Y})^2$ with predicting Y can be a useful guideline.

That said, we will be focusing on estimating f with the aim of minimizing the reducible error. It is important to keep in mind that the irreducible error will always provide an upper bound on the accuracy of our prediction for Y. This bound is almost always unknown in practice.

1B Conditional Expectation (Harvard Open Source Lecture)

Q: Why do we care about conditional probability?

We gather evidence and condition on evidence to predict Y.

Useful properties of conditional estimation are as follow:

$$E(h(X)Y|X) = h(X)E(Y|X) \ [Taking\ out\ what\ is\ known]$$

- If the X is given (known) so we know any function h of X, hence it is a constant from our view.

$$E(Y|X) = E(Y)\ if\ X,\ Y\ are\ independent$$

- If X, Y are independent, conditional distribution of Y given X is no different to non-conditional Y given X.

*In this case the X being conditional does not help at all on predicting Y.

$$E(E(Y|X)) = E(Y)\ more\ like\ E(Y) = E(E(Y|X))$$

* Interated Expectation

- E(Y) is essentially what we want to find out;
- Cleverly choose appropriate X to make the problem simpler

$$E((Y - E(Y|X))h(X)) = 0$$

- i.e. Y - E(Y|X) (residuals) is uncorrelated with h(X) (any function of X)
- How far are we with prediction
- $ cov(Y-E(Y|X), h(X)) = E((Y-E(Y|X))h(X)) - E(Y-E(Y|X))Eh(X) $
- Uncorrelated means in geometrically, the residuals E(Y - E(Y|X)) is orthogonal to the any function X
- Geometric intuition still applies to multiple dimensional space.
- When we do geometric conditional expectation, it is the projection!

$$Var(Y) = E(Var(Y|X)) + Var(E(Y|X))$$

- Unconditional variance of Y = conditional variance of Y given X + conditional expectation of Y given X
- EVE's Law

1C Dimensionality and Structured Models

1C.1 Nearest Neighbouring Averaging

The conditional averaging approach on regression can have its limitation, when the data is not sufficient in amount to average every point Y. Also, referred as k-Nearest Neighbour (k-NN) Regression.

Intuition: An estimate is formed by averaging over the 'k' nearest data points which are defined by a function called *neighborhood* (N)

- $E(Y|X = x)$, Not enough data points to conditionally estimate the Y.
- Relax the definition (parametric -> non-parametric) and let $\hat{f}(x) = Ave(Y|X \in N(x))$, where $N(x)$ is some *neighborhood* of x.
- The degree of relaxing/smoothing can be represented by local size/amount of the neighbourhood acceptance boundary.
- Non-parametric methods, in general, more likely to come closer to estimate the f and with an optimal parameter search, can outperform traditional methods. That said, the setting of parameters in k-NN regression is limited to only k, which may pave way for parametric approach of solving complexed model.

1C.2 Curse of High Dimensionality

There are some pitfall to nearest neighbour averaging. Nearest neighbour averaging is known to be most effective under these circumstances:

- Small number of variables, small P, i.e. p =< 4 and large N to supply enough points in each neighbour to average for our estimate.

In high dimension (large P), the nearest neighbour methods can be lousy, this is often referred as curse of dimensionality.

> Claim 1: "Generalizing correctly becomes exponentially harder as the dimensionality grows because fixed-size training sets cover a dwindling fraction of the input space."

The expected edge length is $e_D(r) = r^{1/D}$, e.g. $e_{10}(0.01) = 0.63$, $e_{10}(0.1) = 0.80$

i.e. to capture 1% or 10% of the data, we must cover 63% or 80% of the range of each input variable.

> Claim 2: In high-dimension, data-points are far from each other. Consequently, "as the dimensionality increases, the choice of nearest neighbor becomes effectively random."

Consider N data points uniformly distributed in a D-dimensional unit ball centered at the origin. We consider a nearest-neighbor estimate at the origin. The median distance from the origin to the closest data point is:

For N = 500, D = 10, this number is 0.52, more than halfway to the boundary.

- We need to get a reasonable fraction of the N values of Y_i to average to bring the variance down - e.g. 10% (10% of the data points to be in each interval). So that our estimate has got a reasonable small variance.
- A 10% neighbourhood in high dimensions need no longer be local, so we lose the spirit of estimating $E(Y|X = x)$ by local averaging.