Computing for Data Sciences

Lecture 17

Model

In statistical modelling, a model can be represented in a simplified real world form as

Data = Model + Error

Here, in the above equation, “Model” part represents the relationships among variables. According to machine learning vocabulary most of the models can be categorized as being explanatory or supervised. The sole aim of any model is to give a good fit to observed data, so that the error term of the equation might be considered as a white noise with a minimal variance. Conventionally models are a subject of theory (biology, economics, physics etc.) and the role of statistics consists in:

- Estimating model parameters, usually done by Maximum Likelihood (ML) which has displaced other techniques like moments, minimum chi-square.
- Checking if the data is in agreement with the model (and vice-versa)

Model checking is frequently omitted in too many publications, even when included models are used to assess the influence of variables or risk factors on a response rather than to predict individual behaviors. This might be in contradiction with the scientific exigency of having falsifiable models.

Model selection and Assessment

While solving a prediction problem, the approach is divided into two parts:

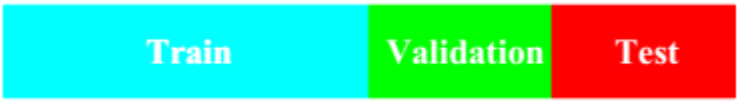
1. Model selection
2. Model assessment

Model selection involves estimating the performance of different models in order to choose the best one while in model assessment involves estimating its prediction error (generalization error) on new data.

In a data-rich situation, the best approach for both of the problems is to randomly divide the dataset into three parts:

- Training set
- Validation set
- Test set.

The training set is used to fit the models, the validation set is used to estimate prediction error for model selection, the test set is used for assessment of the generalization error of the final chosen model.



Ideally, the test set should be kept hidden and be brought out only at the end of the data analysis. Suppose instead that we use the test-set repeatedly, choosing the model with smallest test-set error. Then the test set error of the final chosen model will underestimate the true test error, sometimes substantially.

The steps involved in the process are:

1. Model selection, the validation model is treated as the test data. We train all competing model on the train data and define the best model as the one that predicts best in the validation set. We could re-split the train/validation data, do this many times, and select the method that, on average, best performs.
2. Because we chose the best model among many competitors, the observed performance will be a bit biased. Therefore, to appropriately assess performance on independent data we look at the performance on the test set.
3. Finally, we can re-split everything many times and obtain average results from steps 1 and 2.

It is difficult to give a general rule on how to choose the number of observations in each of the three parts, as this depends on the signal-to-noise ratio in the data and the training sample size. A typical split might be 50% for training, and 25% each for validation and testing. We use the train and validation data to select the best model and the test data to assess the chosen model.

There are two common problems

1. When the amount of data is limited, the results from fitting a model to 1/2 the data can be substantially different to fitting to all the data.
2. Model fitting might have high computational requirements.

A model with the lowest Expected Prediction Error (EPE) can be termed as the best model:

$$EPE\ (\lambda) = E\ [\ L\left\{\ Y - \hat{f}_{\lambda}(X)\right\}\ ]$$

Here Y and X are drawn at random from the population and the expectation averages anything that is random.

Typical loss function are squared error, $L\left(\ Y, \hat{f}(X)\right) = \ (Y - \ \hat{f}(X))^2$, and absolute error, $L\left(\ Y, \hat{f}(X)\right) = \ \left|Y - \ \hat{f}(X)\right|$

We define training error as the observed average loss,

$$\frac{1}{N}\sum_{i=1}^{N} L\{y_i\,, \hat{f}(x_i)\}$$

With squared error loss this is the residual sum of squares divided by N, which is termed as the Average Squared Error (ASE).
For categorical data, using square loss doesn't make much sense. Typical loss functions are 0–1 loss, $L\left(G, \hat{G}(X)\right) = 0$ if $G = \hat{G}(X)$, 0 otherwise, and the log likelihood:
$L\left(G, \hat{G}(X)\right) = -2\sum_{k=1}^{K} I\ (G = k)\log_{\widehat{p_k}}(X) = \ -2\log_{\widehat{p_G}}(X)$

The latter is also known as *cross-entropy*. Here '–2' is used so that in case of a normal error it becomes equivalent to the loss function. For 0–1 loss it is simple the percentage of times we are wrong in the training data. For the likelihood loss we simply use the observed log-likelihood times –2/N:

$$-2\sum_{i\ =1}^{N}\log_{\widehat{p_{g_i}}}(x_i)$$

Need to understand Bias and Variance.

Prediction errors can be decomposed into two main subcomponents

1. Error due to "bias"
2. Error due to "variance"

Having a clear understanding of these two types of errors help us diagnose model results and help us in avoiding over- or under-fitting.

Bias and Variance

Conceptual Definition

Error due to Bias:

The error due to bias is the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. Having one model and talking about expected or average prediction values might seem a little absurd.

Imagining that one can repeat the whole model building process more than once and by gathering new data each time and running a new analysis consequentially creating a new model. Given the credit to the randomness in underlying data sets, resulting models would have a range of predictions. Bias measures how far off these models' predictions are from the correct value.

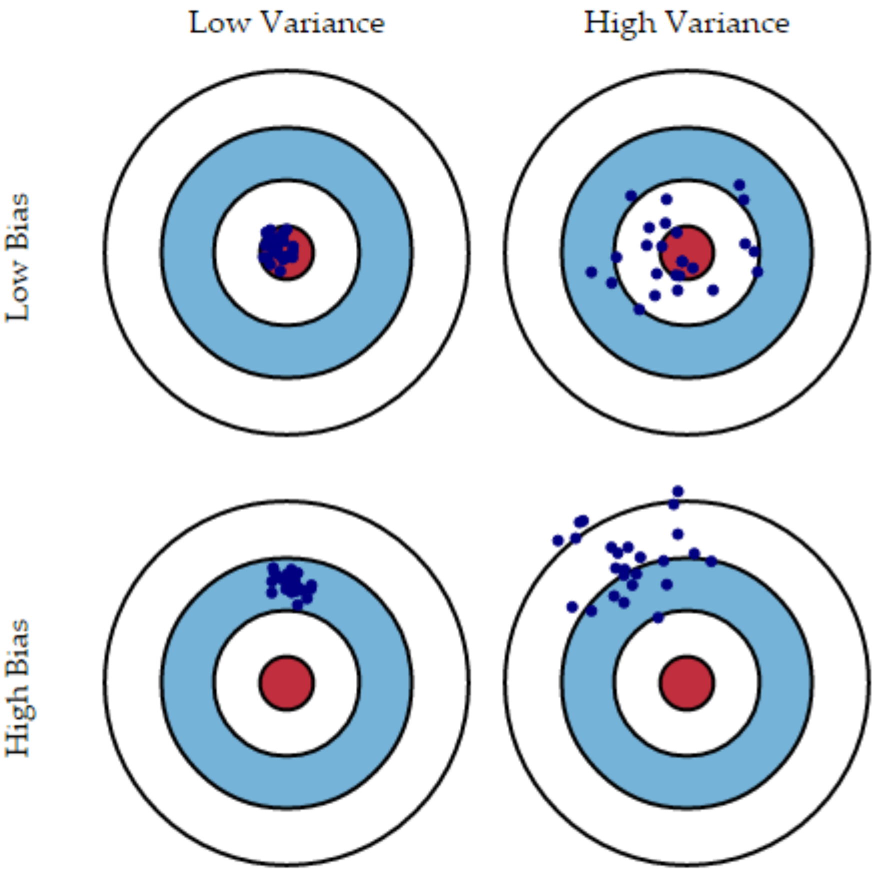
Error due to Variance:

The error due to variance is the variability of a model prediction for a given data point. Imagining that one can repeat the entire model building process multiple times. The variance is how much the predictions for a given point vary between different realizations of the model.

Graphical Definition

A graphical visualization of bias and variance using a bulls-eye diagram. Imagine that the center of the target is a model that perfectly predicts the correct values. As we move away from the bulls-eye, our predictions get worse and worse. We have to repeat our entire model building process to get a decent number of separate hits on the target. Each hit represents an individual realization of our model, given the chance variability in the training data we gather.

Sometimes we will get a good distribution of training data so we predict very well and we are close to the bulls-eye, while sometimes our training data might be full of outliers or non-standard values resulting in poorer predictions. These different realizations result in a scatter of hits on the target.



Graphical illustration of bias and variance.

Mathematical Definition

Let Y denote the variable which we are trying to predict and X denote the covariates. Assume that there is a relationship relating one to the other such as

$$Y = f(X) + \epsilon$$

Here the error term "ϵ" is normally distributed with a mean of zero i.e. $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon)$.

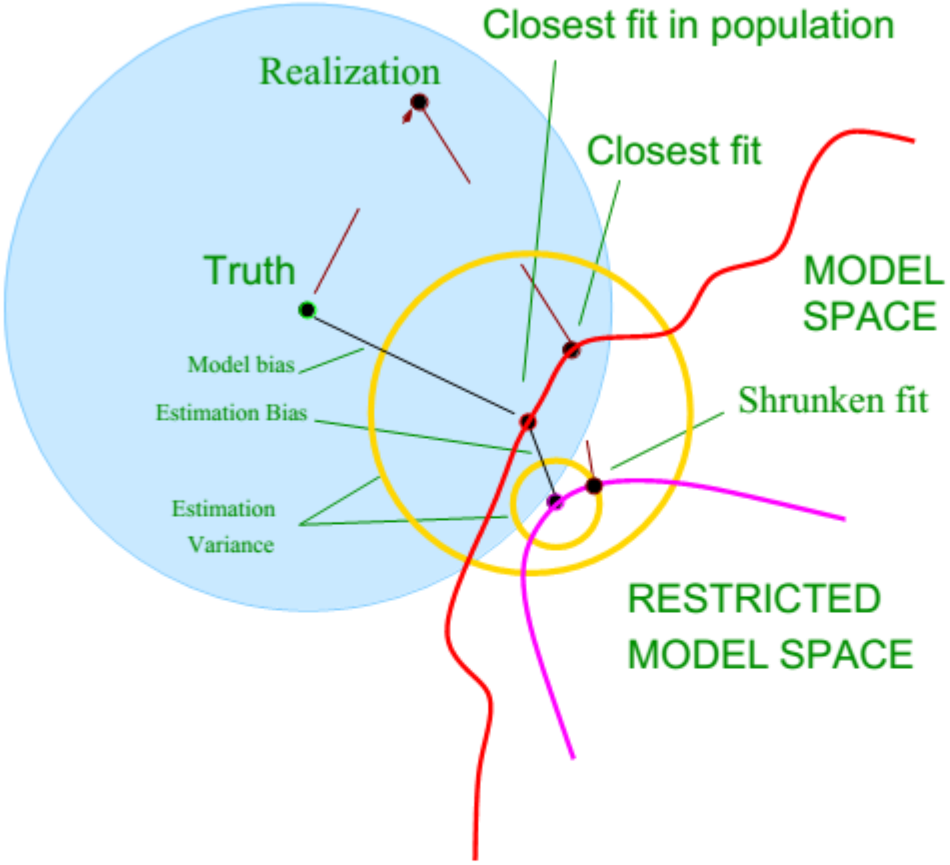
Let us estimate a model $\hat{f}(X)$ of $f(X)$ using linear regressions or any other modeling technique. In this case, the expected squared prediction error at a point x is:

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

This error can be decomposed into bias and variance components as

$$Err(x) = \left(E\left[\hat{f}(x)\right] - f(x)\right)^2 + \ E\left[\hat{f}(x) - E\left[\widehat{f}(x)\right]\right]^2 + \ \sigma_\epsilon^2$$

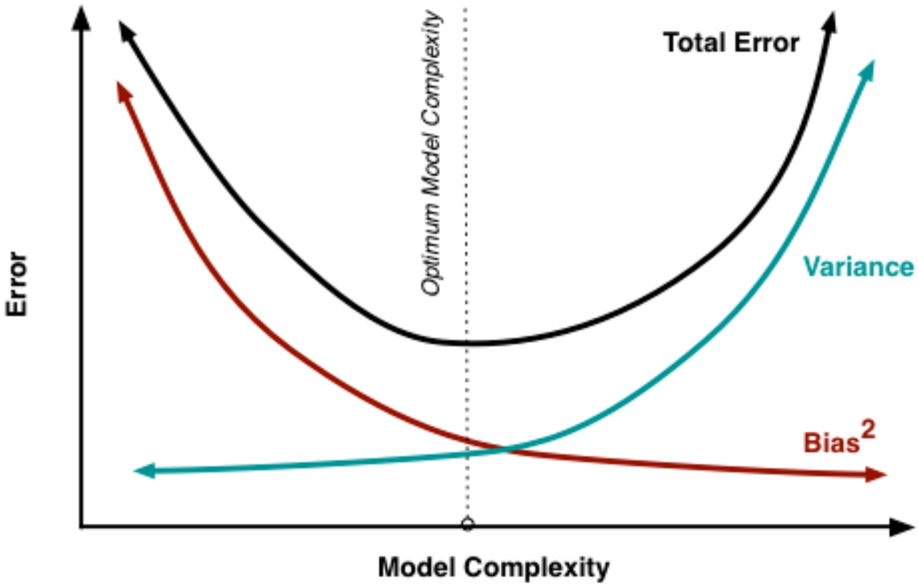
$$Err(x) = Bias^2 + Variance + Irreducible\ Error$$



The third term, irreducible error, is the noise term in the true relationship that cannot fundamentally be reduced by any model. Given the true model and infinite data to calibrate it, we should be able to reduce both the bias and variance terms to 0. However, in a world with imperfect models and finite data, there is a tradeoff between minimizing the bias and minimizing the variance.

Understanding Over- and Under-Fitting

Fundamentally, dealing with bias and variance is all about dealing with over- and under-fitting. Bias is reduced and variance is increased in relation to model complexity. As more number of parameters are added to the model, the complexity of the model rises and variance becomes our primary concern while bias steadily falls.



Bias and variance contributing to total error.

Understanding bias and variance is critical for understanding the behavior of prediction models, but generally what we really care about is overall error, not the specific decomposition. The sweet spot for any model is the level of complexity at which the increase in bias is equivalent to the reduction in variance.

Mathematically: $$\frac{dBias}{dComplexity} = -\frac{dVariance}{dComplexity}$$

If our model complexity exceeds this sweet spot, we are in effect over-fitting our model; while if our complexity falls short of the sweet spot, we are under-fitting the model.

In practice, there is not an analytical way to find this location. So, it is recommended to use an accurate measure of prediction error and explore differing levels of model complexity and then choose the complexity level that minimizes the overall error.

The selection of an accurate error measure is the key to this process, as, often grossly inaccurate measures are used which can be deceptive.

References:

- The Elements of Statistical Learning - Data Mining, Inference, and Prediction By Trevor Hastie, Robert Tibshirani, Jerome Friedman ISBN: 978-0-387-84857-0 (Print)
- http://scott.fortmann-roe.com/docs/BiasVariance.html
- http://www.souravsengupta.com/cds2015/lectures/Gonzalez_Slides.pdf
- Model assessment - by Gilbert Saporta, Ndèye Niang
- Lecture 5: Model selection and assessment by Hector Corrada Bravo and Rafael A. Irizarry February, 2010