Overfit: Good performance on training test and poor performance on test data
Underfit: Poor performance on training test and poor performance on test data

Degrees of Freedom

The problem is that when we have more parameters than observations, there is a risk of overfitting the training dataset. This is intuitive if we think of each coefficient in the model as a point of control. If we have more points of control in the model than we have observations, we can, in theory, configure the model to predict the training dataset correctly and exactly. Learning the details of the training dataset at the expense of performing well on new data is the definition of overfitting.

Parametric/Non-parametric

Most nonparametric tests don't have degrees of freedom.

Binomial Deviance

Deviance is a number that measures the goodness of fit of a logistic regression model. Think of it as the distance from the perfect fit — a measure of how much your logistic regression model deviates from an ideal model that perfectly fits the data. The smaller the number the better the model fits the sample data (deviance = 0 means that the logistic regression model describes the data perfectly). Higher values of the deviance correspond to a less accurate model.

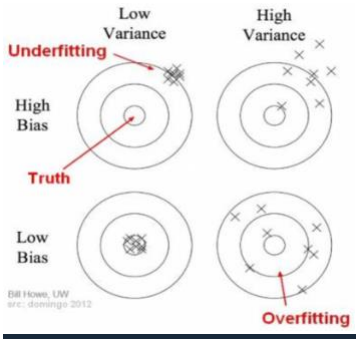
In fact, the model with the lowest deviance will certainly represent the sample data better than any other model. But the problem is that it may not generalize well. In this case we say that the model is overfitting the sample data. So optimizing for the smallest deviance on the sample data does not guarantee a small deviance on out-of-sample data. The problem is that the model with more predictors will ALWAYS have a lower deviance than the smaller model — i.e. you cannot lose accuracy by adding more variables (in the worst case, if the additional variables were not important at all, the model can always set their coefficients equal to 0). So if you choose the model with the lowest deviance, you will always end up picking a model that includes all the predictors under consideration. This is not desirable because not all the independent variables will be good predictors of the outcome. Some will seem good predictors just by chance (in fact, for each 20 predictors included in the model, 1 will have a p-value < 0.05 just by chance). Therefore, if you want to compare models of different sizes by using the deviance, you will also need to adjust for the number of predictors.

Bias-Variance Tradeoff

Bias is the difference between the average prediction of our model and the correct value which we are trying to predict. Model with high bias pays very little attention to the training data and oversimplifies the model. It always leads to high error on training and test data.

Variance is the variability of model prediction for a given data point or a value which tells us spread of our data. Model with high variance pays a lot of attention to training data and does not generalize on the data which it hasn't seen before. As a result, such models perform very well on training data but has high error rates on test data.

If our model is too simple and has very few parameters then it may have high bias and low variance. On the other hand if our model has large number of parameters then it's going to have high variance and low bias. So we need to find the right/good balance without overfitting and underfitting the data.



In statistics and machine learning, the bias–variance tradeoff is the property of a model that the variance of the parameter estimated across samples can be reduced by increasing the bias in the estimated parameters. The bias–variance dilemma or bias–variance problem is the conflict in trying to simultaneously minimize these two sources of error that prevent supervised learning algorithms from generalizing beyond their training set.

The bias error is an error from erroneous assumptions in the learning algorithm. High bias can cause an algorithm to miss the relevant relations between features and target outputs (underfitting).

The variance is an error from sensitivity to small fluctuations in the training set. High variance may result from an algorithm modeling the random noise in the training data (overfitting).

The bias–variance decomposition is a way of analyzing a learning algorithm's expected generalization error with respect to a particular problem as a sum of three terms, the bias, variance, and a quantity called the irreducible error, resulting from noise in the problem itself.

Another Cross-Entropy Function

```
cross_entropy <- function(p, actual){
        x <- 0
        for (i in 1:length(actual)){
                if (p[i, which(colnames(p) == actual[i])] == 0)
                p[i, which(colnames(p) == actual[i])] <- 10^(-15)
                x <- x + (log(p[i, which(colnames(p) == actual[i])]))
        }
        return(-x)
}
```

Appendix

```
```{r ref.label=knitr::all_labels(), echo=TRUE, eval=FALSE}
```
```