5.

5.1. When we discuss prediction models, prediction errors can be decomposed into two main subcomponents we care about: error due to "bias" and error due to "variance". There is a tradeoff between a model's ability to minimize bias and variance. Understanding these two types of error can help us diagnose model results and avoid the mistake of over- or under-fitting.

5.1.1. Error due to Bias: The error due to bias is taken as the difference between the expected (or average) prediction of our model and the correct value which we are trying to predict. Of course you only have one model so talking about expected or average prediction values might seem a little strange. However, imagine you could repeat the whole model building process more than once: each time you gather new data and run a new analysis creating a new model. Due to randomness in the underlying data sets, the resulting models will have a range of predictions. Bias measures how far off in general these models' predictions are from the correct value.

5.1.2. Error due to Variance: The error due to variance is taken as the variability of a model prediction for a given data point. Again, imagine you 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.

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

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

This error may then be decomposed into bias and variance components:

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

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

5.2. As the amount of training data decreases, the amount of variation that we can explain also reduces. But if we decrease the dimensionality of the model proportional to the reduction in training data, we can keep the bias constant or even reduce it further.

If we have enough data ($n>>p$), then we can afford to have huge test dataset. Moreover, it will have acceptable variance. But as the size of available data decreases ($n>p$), variance increases. To reduce this variance, we forcefully add bias in our model to reduce the variance. But if size of data is very small ($n<p$), even adding bias might not be enough. In that case, we need to use approaches like PCA.

5.3. This is the classic bias-variance trade-off problem: Too few bins (i.e. k is small) result into too much bias (i.e., deviation from the true but unknown underlying density) but low variability (of

the histogram) across different data realizations i.e. samples; whereas too many bins (i.e. k is large) lead to little bias but too much variability.

Let **B**= $[t_k; t_k+1)$ denote the k_{th} bin. Suppose the histogram has fixed bin width h.

$$h = (b-a)/k$$

A frequency histogram is built using blocks of height 1 and width h stacked in the appropriate bins. The integral of such a figure is clearly equal to nh. Since this is normalized histogram, it uses building block of (1/nh). Let **V** denote the bin count of the k_{th} bin, that is, the number of sample points falling in bin **B**.

Assuming Binomial distribution and f(x) be the true but unknown pdf function:

$$E[\mathbf{V}] = n\mathbf{P}$$

$$Var[\mathbf{V}] = n\mathbf{P}(1-\mathbf{P})$$

Using above two formulas,

$$Var(pdf) = Var[\mathbf{V}]/(nh)^2 = \mathbf{P}(1-\mathbf{P})/nh^2$$

and,

$$Bias(pdf) = (E[\mathbf{V}]/nh) - f(x) = (\mathbf{P}/h) - f(x)$$

Using mean value theorem,

$$Var(pdf) < (\mathbf{P}/nh^2) - (\text{some term})$$

$$|Bias(pdf)| < \mathbf{C}h \text{ where, } \mathbf{C} \text{ is some constant changing with number of bins}$$

These two equations summarizes the recurring trade-off between bias and variance as determined by the choice of h (which is inversely proportional to k). The variance may be controlled by making h large (i.e. k small) so that the bins are wide and of relatively stable height; however, the bias is large. On the other hand, the bias may be reduced by making h small (i.e. k large) so that the bins are narrow; however, the variance is large. Note that the bias can be eliminated by choosing h= 0 ,but this very rough histogram is exactly the empirical probability density function f(x), which has infinite (vertical) variance. The bias and variance may be controlled simultaneously by choosing an intermediate value of the bin width, and allowing the bin width to slowly decrease as the sample size increases.

Note: All the values in bold are value for k_{th} bin. They will be different for different bins.