DNSC 6279 – Quiz 4

Name:

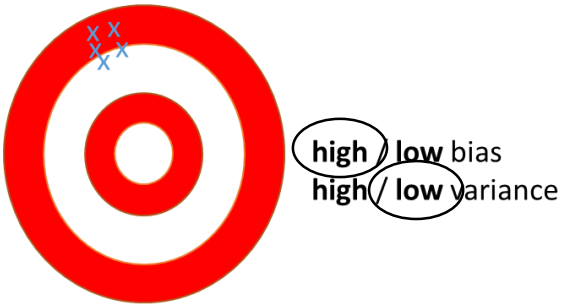
NetId/Email:

1.) **(2 pts.)** The bias-variance trade-off is a fundamental concept in predictive modeling and machine learning that helps us choose the right model for our data. Bias and variance are two different types of error. Circle the term on the left that matches the definition on the right.

Bias / **Variance** — Error due to a model's ability to produce differing predictions from the values in a data set, or the error defined as $\boldsymbol{E}[\hat{\boldsymbol{f}}(\boldsymbol{x}) - \boldsymbol{E}[\hat{\boldsymbol{f}}(\boldsymbol{x})]]^2$**,** where $\boldsymbol{E}$**[]** is the expected value, or average, operator**,** $\boldsymbol{f}(\boldsymbol{x})$, is the true value of the target variable for a data set, and $\hat{\boldsymbol{f}}(\boldsymbol{x})$ is the predicted value of the target variable for a data set.

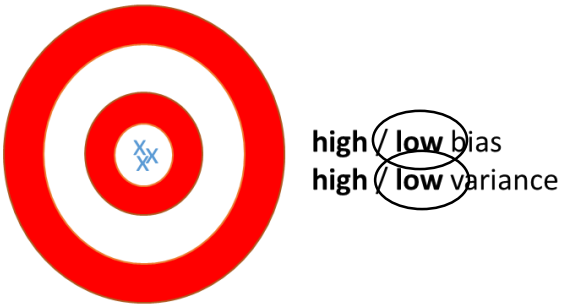
Bias / **Variance** — Error due to a model's inability to replicate the fundamental phenomena represented by a data set, or the error defined as: $\boldsymbol{E}\left[\hat{\boldsymbol{f}}(\boldsymbol{x})\right] - \boldsymbol{f}(\boldsymbol{x})$**,**

2.) **(4 pts.)** The bias-variance trade-off is often described using a bulls-eye analogy. For the charts below let the center of the bulls-eye represent $\boldsymbol{f}(\boldsymbol{x})$ and the x's represent $\hat{\boldsymbol{f}}(\boldsymbol{x})$ for a series of similar models. For each of the four bulls-eyes circle whether the $\hat{\boldsymbol{f}}(\boldsymbol{x})$ represent a high or low bias model and whether the $\hat{\boldsymbol{f}}(\boldsymbol{x})$ represent a high or low variance model.

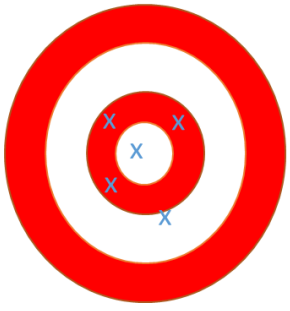


high / **low** bias
high / **low** variance

high / **low** bias
high / **low** variance



high / **low** bias
high / **low** variance



high / **low** bias
high / **low** variance

DNSC 6279 – Quiz 4

Name:

NetId/Email:

3.) **(3 pts.)** Related to the bias-variance tradeoff, "Honest Assessment" is a predictive modeling protocol used to prevent the all-too-common problem of over-fitting the training data. According to common nomenclature (and Professor Prasad's notes), circle the term on the left that matches the definition on the right.

Training / validation / test data	A partition of data used for monitoring and tuning the model to improve its generalization.
Training / validation / test data	A partition of data used only for the final, honest estimate of model performance.
Training / validation / test data	A partition of data used for fitting the model parameters or building the rules that define the model.

4.) **(1 pt.)** In the figure below, two partitions of data are used to fit and asses a predictive model. The variance of the model is increased, and the bias of the model is decreased, by adding predictor variables into the model. (SAS calls this process "complexity optimization".) At what number of variables does the model display the best generalization abilities given only the information in the figure? 11