

STT3851 Homework 4

Dr. Hasthika Rupasinghe

Due – February 16

- 1) Explain what Bias-Variance trade off is.
- 2) How can you identify a High Bias model? How can you fix it?
- 3) Given the test MSE and the Training MSE, how can you tell if the model suffer from overfitting?
- 4) We have data from the questionnaires survey (to ask people opinion) and objective testing with two attributes (acid durability and strength) to classify whether a special paper tissue is good or not. Here is four training samples. Note that X_1 = Acid Durability (seconds), X_2 = Strength (kg/square meter) and Y = Classification.

X_1	X_2	Y
7	7	Bad
7	4	Bad
3	4	Good
1	4	Good

The factory produces a new paper tissue that pass laboratory test with $X_1 = 3$ and $X_2 = 7$. Without another expensive survey, use the following steps to find the classification of this new tissue.

- a) Suppose use $K = 3$
- b) Find the euclidean distance between the query-instance (3, 7) and all the training samples. A table might be useful.
- c) Rank the distances and figure out which points are included in 3-Nearest neighbors.
- d) Use simple majority of the category of nearest neighbors as the prediction value of the query instance.