**Cheat Sheet – Cross Validation**

|  |  |  |
| --- | --- | --- |
| **Type** | **Details** | **Pros and cons** |
| Splitting data into training and test | Data is split into training and test datasets. The model is constructed using the training set, and then tested using the test set | *Pros:* Easy to understand, computationally easy  *Cons:* The validation estimate of the test error rate can be highly variable; because statistical models tend to perform worse when trained on fewer observations, the validation test error rate may tend to *overestimate* the test error rate for the model fit on the entire dataset. |
| Leave one out cross validation (LOOCV) | A single observation is used for the validation set, and the | *Pros*: Far less bias than splitting; always yields the same results; very general method that can be used with any type of model  *Cons*: Expensive to implement |
| K-folds | Works by randomly dividing a set of observations into *k* groups, or folds. Most popular is 5 or 10. The MSE is then calculated on the observations in the held-out fold. | *Pros:* K folds is the most popular method of cross validation; often gives a more accurate estimate of the test error rate than does LOOCV |
| Bootstrap | Works by repeatedly sampling from the same dataset. |  |