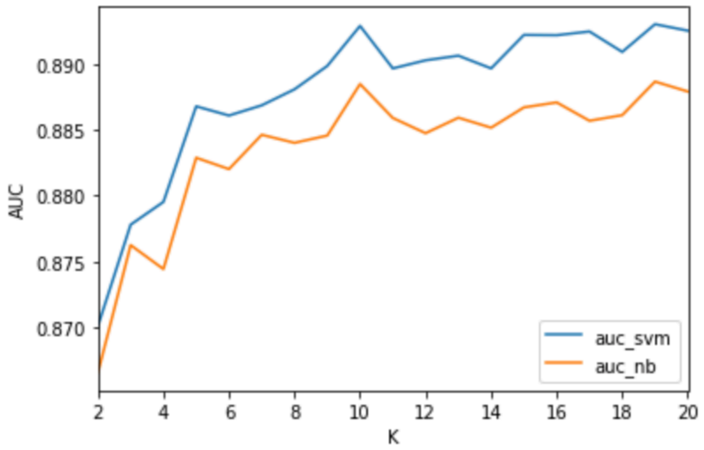
Q2



•

Write your analysis in a separate pdf file (not in code) on the following:
- How does k affect model performance on evaluation sets?
- By varying k, you also change sample size for training. How can the sample size affect model performance?

From the curve showing above, we find that when k is extremely small, the performance of the model is the worst. As the k becomes larger, the performance becomes better, and after that, the performance tend towards balance.

This is reasonable. AUC is the area under the ROC curve. It tells the ability of the classifier that distinguishing between classes. The larger of AUC, means this ability is better. So we usually use AUC to describe the performance of a classifier. For k-folder CV, increasing k decreases the bias because the training set better represents the data, while increasing k also increases the variance of the estimator because the training data sets are becoming more similar. So, at first, the AUC would increase as the k increase. When k is large enough, the classifier is stable and the AUC is stable.

Moreover, as the k becomes larger, the model has a risk of over-fitting, because the training set is much larger than the test set. In this case, the performance of the classifier decreases. The over-fitting problem is typical in machine learning. The cross validation is a way to avoid over-fitting because it makes full use of the training set. As the result, it is not the bigger the k, the better.