Problem 5

Naïve Bayes and Logistic regression are two popular models used to solve numerous machine learning problems, in many ways the two algorithms are similar, but at the same time very dissimilar.

Naïve Bayes is a classification method based on Bayes' theorem that derives the probability of the given feature vector being associated with a label. Naïve Bayes has a naive assumption of conditional independence for every feature, which means that the algorithm expects the features to be independent which not always is the case.

Logistic regression is a linear classification method that learns the probability of a sample belonging to a certain class. Logistic regression tries to find the optimal decision boundary that best separates the classes.

1. Both algorithms are used for classification problems

The first similarity is the classification use case, where both Naive Bayes and Logistic regression are used to determine if a sample belongs to a certain class, for example, if an e-mail is spam or ham.

2. Algorithm's Learning mechanism

The learning mechanism is a bit different between the two models, where Naive Bayes is a generative model and Logistic regression is a discriminative model.

Generative model: Naive Bayes models the joint distribution of the feature X and target Y, and then predicts the posterior probability given as $P(y|x)$

Discriminative model: Logistic regression directly models the posterior probability of $P(y|x)$ by learning the input to output mapping by minimizing the error.

Posterior probability can be defined as the probability of event A happening given that event B has occurred, in more layman terms this means that the previous belief can be updated when we have new information. For example, let's say we think the stock market will go up by 50% next year, this prediction can be updated when we get new information such as updated GPD numbers, interest rates etc.

3. Model assumptions

Naïve Bayes assumes all the features to be conditionally independent. So, if some of the features are in fact dependent on each other (in case of a large feature space), the prediction might be poor.

Logistic regression splits feature space linearly, and typically works reasonably well even when some of the variables are correlated.

4. Approach to be followed to improve model results

Naïve Bayes: When the training data size is small relative to the number of features, the information/data on prior probabilities help in improving the results

Logistic regression: When the training data size is small relative to the number of features, including regularization such as Lasso and Ridge regression can help reduce overfitting and result in a more generalized model.