**1.Naïve Bayes**

1. To discrete Naïve Bayes here we would use discretization (or binning)

* First, we need to convert the continuous features into categorical values by spliting its range into bins
* Replace each continuous value with its bin used for categorization
* For each class y ε Y:
  + Prior: P(y) = count(y) / N
  + Likelihood: P(xj = bink | y) = count(xj = bink , y) / count(y)
* We assign each new instance of x to its corresponding bin
* Posterior:
  + P(y|x) ∝ P(y) P(xj = bink | y)

1. Deriving the expression for the continuous Naïve Bayes

* Assume features are conditionally independent:
  + P(xj | y) =
* Prior:
  + P(y)
* Mean:
* Variance:
* MLE estimation:
  + Prior:
    - P(y=c) =

For each feature and class y = c

* + Mean:
  + Variance:
* Prediction:
  + Posterior for each class y = c:
  + Gaussian:
  + Prediction for highest posterior:

**2. Implement Logic Regression**

\*Code provided

**3. Clustering**

\*Code provided

**4. Variance/Bias Tradeoff:**

• Logistic Regression vs Decision Tree (depth 5)

* Decision Tree has a lower bias, higher variance than logistic regression as DT’s can fit more complex decision boundaries than a linear model which introduces more variance due to its flexibility

• Decision Tree vs Random Forest vs Gradient Boosted Tree

* For variance, DT had the highest of the three, then GBT followed by RF. For bias, Dt has the highest, followed by RF, then GBT. This is because RF reduces variance using averaging while GBT reduces bias by using sequential corrections but is more likely to overfit if not regularized

• Logistic Regression vs 1NN classifier

* 1NN has a lower bias, higher variance than logistic regression as 1NN memorizes the training data which reduces bias but makes it more sensitive to noise

• 5NN Classifier vs 1NN Classifier

* 5NN has a higher bias, lower variance than 1NN as 5NN has smoother predictions by averaging over more points which reduces sensitivity to noise

• Depth 5 Decision Tree vs Depth 10 Decision Tree

* Depth 10 DT has a lower bias, higher variance than a Depth 5 DT because deeper trees fit more training data more closely but are more likely to overfit

• Random Forest with 100 Trees vs Random Forest with 10 Trees

* 100 Trees has a lower variance, but the same bias as 10 Trees as averaging more trees stabilizes predictions which reduces variance

• AdaBoost with 10 Decision Stumps vs AdaBoost with 100 Decision Stumps.

* 100 Stumps have lower bias, higher variance than 10 Stumps as more stumps allow better fitting training data but have a risk of overfitting

• Linear Regression vs Quadratic (Polynomial degree 2) Regression.

* Quadratic Regression has lower bias, higher variance than linear regression as the quad model captures curvature reducing bias but makes it more sensitive to noise as the model is more complex

**5. Ensemble Models and Bias / Variance Tradeoff**

\*Code provided

* The model that had the lowest bias was Gradient Boosting
* The model that had the lowest variance was Decision Tree
* The model that achieved the lowest MSE was Gradient Boosting

Bagging (RFs) reduces variance significantly in compared to single trees while boosting (GB) reduces bias but may increase variance as well. It would be best to use bagging on smaller, noisier datasets while boosting would be better suited for more complex datasets.

**6. Gradient Boosting**

* GB with Logistic Loss
* Logistic Loss:
* Residuals:
* Sigmoid Func.:
* Algorithm:
  + Pseudo-residuals:

Final:

* GB with Hinge Loss
* Hinge Loss:
* Sub-gradient:
* Algorithm:
  + Pseudo-residuals:

Final:

**7. Neural Networks**

* Loss Function:
* Derivatives:
* Gradient:
* Chain Rule for z:
* Gradients: