Assignment 2

Kevin Puga

March 10 2025

Question 1a: **Linear Regression Assumptions:** Explain the key assumptions of linear regression and discuss how violating these assumptions can affect model performance.

- The key assumptions of linear regression are linearity, homoscedasticity, normal distribution, independence of errors, and no multicollinearity. Linearity refers to the relationship between dependent and independent variables. Homoscedasticity refers to the way in which the spread of errors should be uniform in relativity. The normal distribution refers to how the differences between the observed and predicted values should follow a normal distribution. Independence of errors refers to how the differences between the observed and predicted values are not correlated to each other, which ties into no multicollinearity.

- Violating these assumptions could affect the model performance by resulting in unreliable model predictions, biased coefficient estimates, and not being able to interpret the results of the analysis.

Question 1b: **Overfitting and Underfitting:** Suppose you implement a Linear Regression Model. How will you determine if the model is overfitting and underfitting. Provide at least one way you can fix underfitting if the model is underfitting, and provide one way to fix overfitting.

- A model is overfitting if it learns too much from the training data. The model could be too complex, the training data size could be too high, and the variance could be high as well as having a low bias.

- A model is underfitting if the model is too simple that it cannot understand what is happening in the data. The model's training data could be too low, unscaled features, and excessive regulations cause underfitting as well.

- To fix overfitting, we could reduce the complexity of the model or improve the quality of training data.

- To fix the underfitting, we could make the model more complex or increase the duration of training.

Question 2a: **Describe the concept of Gini Split and how it is used in decision trees. Why is it important for determining the best feature split?**

- The Gini Split is a way to determine how to split data into subsets using the Gini index. We use it in decision trees in order to split nodes.
- It is important in determining the best feature split because it leads to a more accurate classification model.

Question 2b: **Decision trees select features at each node based on a chosen metric. How does the tree determine which feature to split on at each step? Why might some features never be selected?**

- The tree determines the feature to split on by calculating the Gini impurity and choosing the feature that results in the most information gain. Features will not be chosen if they provide lower info gain.

Question 3a: **Curse of Dimensionality:** Explain how the curse of dimensionality (the performance degradation with higher dimensional features) affects the performance of nearest-neighbor classifiers.

- The curse of dimensionality affects the performance of nearest-neighbor classifiers by making the space between data points less meaningful, making it harder to identify the nearest neighbors. This leads to poor classification accuracy and increased computational complexity.

Question 3b: **Feature Normalization:** Why is feature scaling important for nearest-neighbor methods? Demonstrate with an example and provide examples of feature normalization methods.

- Feature scaling is important because nearest-neighbors methods depend on distance calculations.
- Ex:

Question 4: **Let us consider the Dice Roll problem.** Define the parameters of a dice roll process and specify the total number of parameters involved. Given N rolls, where each outcome i appears αi times, write the likelihood function for the multinomial model. Provide the Maximum Likelihood Estimation (MLE) for the parameters. Assume a Dirichlet prior for the dice probabilities. Derive the posterior distribution and compute the Maximum A Posteriori (MAP) estimate.

- **Parameters:** k is the number of sides on the dice, θi is the probability of the outcome i where i = 1, 2, ...k, so the parameters are: (θ_1, θ_2, ..., θ_k)
- **MLE for model:** L(θ| α) = P(θ| α)= $(\frac{N!}{\prod^k i{=}1\alpha_i!}) + (\prod^k i = 1)(\theta_i^{\alpha_i}))$

- **MLE for parameters:**

$$\begin{aligned}
log\,L(\theta|\alpha) &= log(\frac{N!}{\prod_{i=1}^{k}\alpha_i!}) + \sum_{i=1}^{k}\alpha_i log(\theta_i) \\
&= log\,L(\theta|\alpha) - \lambda(\sum_{i=1}^{k}\theta_i - 1) \\
&= \frac{\alpha_i}{\theta_i} - \lambda = 0 \\
&= \frac{\alpha_i}{\lambda} = \theta_1 \\
&= \theta_i^{MLE} = \frac{\alpha_i}{N}
\end{aligned}$$

- **MAP estimate:**

$$\begin{aligned}
p(\theta) &= (\frac{1}{B(\beta)})\prod_{i=1}^{k}\theta_i^{\beta_1 - 1} \\
p(\theta|\alpha) &= (\frac{1}{B(\alpha+\beta)})\Pi_{i=1}^{k}\theta_i^{\alpha_1+\beta_1-1}
\end{aligned}$$

$$\begin{aligned}
\theta_i^{MAP} &= (\frac{\alpha_i + \beta_i - 1}{\sum_{j=1}^{k}(\alpha_j + \beta_j - 1}) \\
&= (\frac{\alpha_i + \beta_i - 1}{N + \sum_{j=1}^{k}(\beta_j - k})
\end{aligned}$$