Polytechnique Montréal

INF8245AE Machine Learning

Assignment 2

Ali Abbas
Student ID: 2078188

K-Nearest Neighbors

2 (a) (b):
Validation Set Results:

Distance	k=1	k=2	k=3	k=4	k=5	k=10	k=20
Euclidean	97.12%	96.67%	97.20%	97.26%	97.19%	97.02%	96.41%
Cosine	97.50%	97.14%	97.55%	97.61%	97.60%	97.35%	97.06%

Test Set Results:

Euclidean Distance: Best $k = 4$, test accuracy: 96.58%
Cosine Distance: Best $k = 4$, test accuracy: 97.04%

Trend Commentary:

Cosine distance performs better than Euclidean distance with all k values. Both methods show a decline in accuracy when k increases beyond the optimal values of $k = 4$ on the validation set, likely due to noise from distant neighbors.

3 (b):
Time and space complexity of the building tree:
My implementation does the following:

1. Select the dimension to split based on the depth
2. Sort the points using numpy.argsort on the selected dimension
3. Recursively build left and right subtrees

Time Complexity: $O(n \log^2 n)$
At each level of recursion, I used `np.argsort` to find the median along the selected dimension, which requires sorting the points and takes $O(n \log n)$ time. Since the tree has a depth of $O(\log n)$ due to halving up to level k, the total time complexity becomes $O(n \log n) \times O(\log n) = O(n \log^2 n)$.
Space Complexity: $O(n)$
At each depth of recursion I store half of the points for left and right subtrees, $n, n/2, ..., n/2^k (k = depth)$ which sums to linear space O(n).

3 (c):
Time complexity of traversing the tree: $O(n)$
Assuming a worst case scenario where the algorithm would need to traverse from depth 0 to k the time complexity would be $O(k)$.

3 (d):
The test accuracy using the KD-tree is 100%

Question 4

(a) Show that for any classifier G, $R(G) = \mathbb{E}[\eta(X)1_{G(X)=0} + (1-\eta(X))1_{G(X)=1}]$ where $\eta(x) = P(g = 1|X = x)$.

$$\begin{aligned} R(G) &= \mathbb{E}_{X,g\sim v}[1_{G(X)\neq g}] \\ R(G) &= \mathbb{E}[\mathbb{E}[1_{G(X)\neq g}|X]] \text{ (Conditional Expectation)} \end{aligned} \tag{1}$$

$$\begin{aligned} 1_{G(X)\neq g} &= 1_{G(X)=0,g=1} + 1_{G(X)=1,g=0} \\ \mathbb{E}[1_{G(X)\neq g}|X = x] &= \mathbb{E}[1_{G(X)=0,g=1} + 1_{G(X)=1,g=0}|X = x] \\ &= P(G(X) = 0, g = 1|X = x) + P(G(X) = 1, g = 0|X = x) \\ &= P(g = 1|X = x)1_{G(X)=0} + P(g = 0|X = x)1_{G(X)=1} \\ &= \eta(x)1_{G(X)=0} + (1-\eta(x))1_{G(X)=1} \end{aligned}$$

$$R(G) = \mathbb{E}[\eta(X)1_{G(X)=0} + (1-\eta(X))1_{G(X)=1}] \text{ (using equation (1))}$$

(b) Show that R$^* := R(G^*) = \mathbb{E}[\min(\eta(X), 1-\eta(X))]$

$R^* = \mathbb{E}[\min(\eta(X), 1-\eta(X))]$
$R^* = \mathbb{E}[\eta(X)1_{G^*(X)=0} + (1-\eta(X))1_{G^*(X)=1}]$

$$G^*(X) = \begin{cases} 0 & \text{if } \eta(X) < 1/2 \\ 1 & \text{elsewhere} \end{cases}$$

Suppose $\eta(X) \geq \frac{1}{2}$, then $\eta(X)\,\mathbf{1}_{G^*(X)=0} = 0$
and $(1-\eta(X))\,\mathbf{1}_{G^*(X)=1} = 1-\eta(X) \leq \eta(X)$
So, $\min(\eta(X), 1-\eta(X)) = (1-\eta(X))\ and\ R^* = \mathbb{E}[(1-\eta(X))]$

Suppose $\eta(X) < \frac{1}{2}$, then $\eta(X)\,\mathbf{1}_{G^*(X)=0} = \eta(X) < 1-\eta(X)$
and $(1-\eta(X))\,\mathbf{1}_{G^*(X)=1} = 0$
So, $\min(\eta(X), 1-\eta(X)) = \eta(X)\ and\ R^* = \mathbb{E}[\eta(X)]$

Thus, in both cases, we have :
$R^* = \mathbb{E}[\min(\eta(X), 1-\eta(X))]$

Show that for any classifier G, $R(G) - R^* = \mathbb{E}[|2\eta(X)-1|1_{G(X)\neq G^*(X)}]$

$1_{G(X)=0} = 1 - 1_{G(X)=1}$ and $1_{G^*(X)=0} = 1 - 1_{G^*(X)=1}$

$$1_{G(X)=0} - 1_{G^*(X)=0} = 1 - 1_{G(X)=1} - 1 + 1_{G^*(X)=1} = 1_{G^*(X)=1} - 1_{G(X)=1}$$

$$\begin{aligned}
R(G) - R^* &= \mathbb{E}[|2\eta(X) - 1|1_{G(X)\neq G^*(X)}] \\
\\
&= \mathbb{E}[\eta(X)1_{G(X)=0} + (1-\eta(X))1_{G(X)=1}] \\
&\quad - \mathbb{E}[\eta(X)1_{G^*(X)=0} + (1-\eta(X))1_{G^*(X)=1}] \\
\\
&= \mathbb{E}[\eta(X)1_{G(X)=0} - \eta(X)1_{G^*(X)=0} + (1-\eta(X))1_{G(X)=1} \\
&\quad - (1-\eta(X))1_{G^*(X)=1}] \\
\\
&= \mathbb{E}[\eta(X)(1_{G(X)=0} - 1_{G^*(X)=0}) + (1-\eta(X))(1_{G(X)=1} - 1_{G^*(X)=1})] \\
&= \mathbb{E}[\eta(X)(1_{G^*(X)=1} - 1_{G(X)=1}) + (1-\eta(X))(1_{G(X)=1} - 1_{G^*(X)=1})] \\
&= \mathbb{E}[(2\eta(X) - 1)(1_{G^*(X)=1} - 1_{G(X)=1})]
\end{aligned}$$

Case 1: suppose $1_{G^*(X)\neq G(X)} = 0$, then $1_{G^*(X)} = 1_{G(X)}$; $1_{G^*(X)} - 1_{G(X)} = 0$

Case 2: suppose $1_{G^*(X)\neq G(X)} = 1$
if $G^*(X) = 1$ and $G(X) = 0$, then $\text{G}^*(X) = 1\, and\, \eta \geq 1/2$
$\hookrightarrow 2\eta(X) - 1 \geq 0$
$\hookrightarrow (2\eta(X) - 1)(1 - 0) = 2\eta(X) - 1$
since $\eta(X) \geq 1/2,\, then\, 2\eta(X) - 1 = |2\eta(X) - 1|$

if $G^*(X) = 0$ and $G(X) = 1$, then $\text{G}^*(X) = 0\, and\, \eta < 1/2$
$\hookrightarrow 2\eta(X) - 1 < 0$
$\hookrightarrow (2\eta(X) - 1)(0 - 1) = -(2\eta(X) - 1) = 1 - 2\eta(X)$
since $\eta(X) < 1/2,\, then$ 1 - $2\eta(X) = |2\eta(X) - 1|$

Thus, in both cases, we have :
$R(G) - R^* = \mathbb{E}[|2\eta(X) - 1|1_{G(X)\neq G^*(X)}]$

(c) Show that $R(\hat{G}_n) - R^* \leq 2\mathbb{E}[|\eta(X) - \hat{\eta}_n(X)|]$.

$$\begin{aligned}
R(\hat{G}_n) - R^* &= \mathbb{E}[|2\eta(X) - 1|1_{\hat{G}_n(X)\neq G^*(X)}] \\
&= 2\mathbb{E}[|\eta(X) - 1/2|1_{\hat{G}_n(X)\neq G^*(X)}]
\end{aligned}$$

$$\text{G}^*(X) = \begin{cases} 1 & \text{if } \eta(X) \geq 1/2 \\ 0 & \text{elsewhere} \end{cases}$$

$$\hat{G}_n(X) = \begin{cases} 1 & \text{if } \hat{\eta}_n(X) \geq 1/2 \\ 0 & \text{elsewhere} \end{cases}$$

Case 1: suppose $\text{G}^*(X) = 1\, and\, \hat{G}_n(X) = 0$, then $\eta(X) \geq 1/2$ and $\hat{\eta}_n(X) < 1/2$
$\hookrightarrow \hat{\eta}_n(X) < 1/2 \leq \eta(X) \to |\eta(X) - \hat{\eta}_n(X)| \geq |\eta(X) - 1/2|$

Case 2: suppose $\mathrm{G}^*(X) = 0 \, and \, \hat{G}_n(X) = 1$, then $\eta(X) < 1/2$ and $\hat{\eta}_n(X) \geq 1/2$
$\hookrightarrow \eta(X) < 1/2 \leq \hat{\eta}_n(X) \to |\eta(X) - \hat{\eta}_n(X)| \geq |\eta(X) - 1/2|$

Thus, in both cases, we have : $|\eta(X) - \hat{\eta}_n(X)| \geq |\eta(X) - 1/2|$

$$\begin{aligned} R(\hat{G}_n) - R^* &= 2\mathbb{E}[|\eta(X) - 1/2| 1_{\hat{G}_n(X) \neq G^*(X)}] \\ &\leq 2\mathbb{E}[|\eta(X) - \hat{\eta}_n(X)| 1_{\hat{G}_n(X) \neq G^*(X)}] \\ &\leq 2\mathbb{E}[|\eta(X) - \hat{\eta}_n(X)|] \end{aligned}$$

Question 5

(a) Show that $\mathbb{E}[(\eta(x) - \hat{\eta}_n(X))^2] \leq 2\mathbb{E}[(\eta(x) - \tilde{\eta}_n(X))^2] + 2\mathbb{E}[(\tilde{\eta}_n(x) - \hat{\eta}_n(X))^2]$

$$\begin{aligned} \eta(x) - \hat{\eta}_n(x) &= \eta(x) - \tilde{\eta}_n(X) + \tilde{\eta}_n(x) - \hat{\eta}_n(x) \\ &= (\eta(x) - \tilde{\eta}_n(x)) + (\tilde{\eta}_n(x) - \hat{\eta}_n(x)) \\ (\eta(x) - \hat{\eta}_n(x))^2 &= (\eta(x) - \tilde{\eta}_n(x))^2 + (\tilde{\eta}_n(x) - \hat{\eta}_n(x))^2 \\ &\quad + 2(\eta(x) - \tilde{\eta}_n(x))(\tilde{\eta}_n(x) - \hat{\eta}_n(x)) \end{aligned}$$

(Using $2ab \leq a^2 + b^2$)

$$\begin{aligned} 2(\eta(x) - \hat{\eta}_n(x))(\hat{\eta}_n(x) - \tilde{\eta}_n(x)) &\leq (\eta(x) - \tilde{\eta}_n(x))^2 + (\tilde{\eta}_n(x) - \hat{\eta}_n(x))^2 \\ 2(\mathbb{E}[(\eta(x) - \hat{\eta}_n(X))(\tilde{\eta}_n(X) - \hat{\eta}_n(X))]) &\leq \mathbb{E}[(\eta(x) - \tilde{\eta}_n(X))^2] + \mathbb{E}[(\tilde{\eta}_n(X) - \hat{\eta}_n(X))^2] \\ \mathbb{E}[(\eta(x) - \hat{\eta}_n(X))^2] &\leq \mathbb{E}[(\eta(x) - \tilde{\eta}_n(X))^2] + \mathbb{E}[(\tilde{\eta}_n(X) - \hat{\eta}_n(X))^2] \\ &\quad + \mathbb{E}[(\eta(x) - \tilde{\eta}_n(X))^2] + \mathbb{E}[(\tilde{\eta}_n(X) - \hat{\eta}_n(X))^2] \\ &= 2\mathbb{E}[(\eta(x) - \tilde{\eta}_n(X))^2] + 2\mathbb{E}[(\tilde{\eta}_n(X) - \hat{\eta}_n(X))^2] \end{aligned}$$

(b) Show that $\mathbb{E}[(\eta(X) - \tilde{\eta}_n(X))^2] \leq \varepsilon + \mathbb{E}[\sum_{i=1}^n w_{n,i}(X) 1_{\|X_i - X\| \geq \delta}]$.

$$\begin{aligned} \eta(X) - \tilde{\eta}_n(X) &= \eta(X) - \sum_{i=1}^n w_{n,i}(X)\eta(X_i) \\ &= \sum_{i=1}^n w_{n,i}(X)\eta(X) - \sum_{i=1}^n w_{n,i}(X)\eta(X_i) \\ &= \sum_{i=1}^n w_{n,i}(X)(\eta(X) - \eta(X_i)) \end{aligned}$$

Since $\sum_{i=1}^n w_{n,i}(X) = 1$, Jensen's inequality:

$$
\begin{aligned}
(\eta(X) - \tilde{\eta}_n(X))^2 &= \Big(\sum_{i=1}^{n} w_{n,i}(X)(\eta(X) - \eta(X_i))\Big)^2 \\
&\leq \sum_{i=1}^{n} w_{n,i}(X)(\eta(X) - \eta(X_i))^2 \\
\mathbb{E}[(\eta(X) - \tilde{\eta}_n(X))^2] &\leq \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X)(\eta(X) - \eta(X_i))^2] \\
\mathbb{E}[(\eta(X) - \tilde{\eta}_n(X))^2] &\leq \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X)(\eta(X) - \eta(X_i))^2 1_{\|X_i - X\| < \delta}] \\
&\quad + \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X)(\eta(X) - \eta(X_i))^2 1_{\|X_i - X\| \geq \delta}]
\end{aligned}
$$

(ε, δ)-definition of limit: If η is continuous, then for every $\varepsilon > 0$, there exists $\delta > 0$ such that for all $p :\|x - p\| < \delta \Rightarrow |\eta(x) - \eta(p)| < \epsilon$

$$
\begin{aligned}
\mathbb{E}[(\eta(X) - \tilde{\eta}_n(X))^2] &\leq \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X)\varepsilon^2 1_{\|X_i - X\| < \delta}] + \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X) \cdot 1 \cdot 1_{\|X_i - X\| \geq \delta}] \\
&\leq \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X)\varepsilon^2] + \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X) 1_{\|X_i - X\| \geq \delta}] \\
&\leq \varepsilon^2 \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X)] + \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X) 1_{\|X_i - X\| \geq \delta}] \\
\mathbb{E}[(\eta(X) - \tilde{\eta}_n(X))^2] &\leq \varepsilon^2 + \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X) 1_{\|X_i - X\| \geq \delta}] \\
&= \varepsilon + \mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X) 1_{\|X_i - X\| \geq \delta}] \text{ (since } \varepsilon \text{ can be close to null)}
\end{aligned}
$$

(c) Show that $\mathbb{E}[(\tilde{\eta}_n(X) - \hat{\eta}_n(X))^2] = \mathbb{E}[\sum_{i=1}^{n} w_{n,i}^2(X)(g_i - \eta(X_i))^2]$

$$
\begin{aligned}
\tilde{\eta}_n(X) - \hat{\eta}_n(X) &= \sum_{i=1}^{n} w_{n,i}(X)\eta(X) - \sum_{i=1}^{n} w_{n,i}(X) g_i \\
&= \sum_{i=1}^{n} w_{n,i}(X)(\eta(X) - g_i) \\
(\tilde{\eta}_n(X) - \hat{\eta}_n(X))^2 &= \Big(\sum_{i=1}^{n} w_{n,i}(X)(\eta(X) - g_i)\Big)^2 \\
\mathbb{E}[(\tilde{\eta}_n(X) - \hat{\eta}_n(X))^2] &= \mathbb{E}\Big[\Big(\sum_{i=1}^{n} w_{n,i}(X)(\eta(X) - g_i)\Big)^2\Big]
\end{aligned}
$$

Because $g_i \in [0,1]$, we have $(g_i - \eta(x_i))^2 \leq 1$

$$\mathbb{E}[\sum_{i=1}^{n} w_{n,i}^2(x)(g_i - \eta(x_i))^2] \leq \mathbb{E}[\sum_{i=1}^{n} w_{n,i}^2(x)].$$

with the weights, $w_{n,i}(x) \in \{1/k, 0\}$

$$\sum_{i=1}^{n} w_{n,i}^2(x) = k(\frac{1}{k})^2 = \frac{1}{k}.$$

Since $\sum_i w_{n,i}(x) = 1$,

$$\sum_i w_{n,i}^2(x) \leq \max_i w_{n,i}(x) = \frac{1}{k}.$$

Therefore,

$$\mathbb{E}[(\tilde{\eta}(x) - \hat{\eta}_n(x))^2] \leq \mathbb{E}[\max_i w_{n,i}] = \frac{1}{k}.$$

(d) Combining parts (a) and (c):

$$\mathbb{E}[(\eta(x) - \hat{\eta}_n(X))^2] \leq 2\varepsilon^2 + 2\mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X) 1_{\|X_i - X\| \geq \delta}] + \frac{2}{k}$$

$k \to \infty$ and $\frac{k}{n} \to 0$ as $n \to \infty$, then: $\frac{2}{k} \to 0$, $\mathbb{E}[\sum_{i=1}^{n} w_{n,i}(X) 1_{\|X_i - X\| \geq \delta}] \to 0$ and ε can be close to null.

Therefore, $\mathbb{E}[(\eta(x) - \hat{\eta}_n(X))^2] \to 0$.

Logistic Regression

(a) **Gradient with respect to W:**

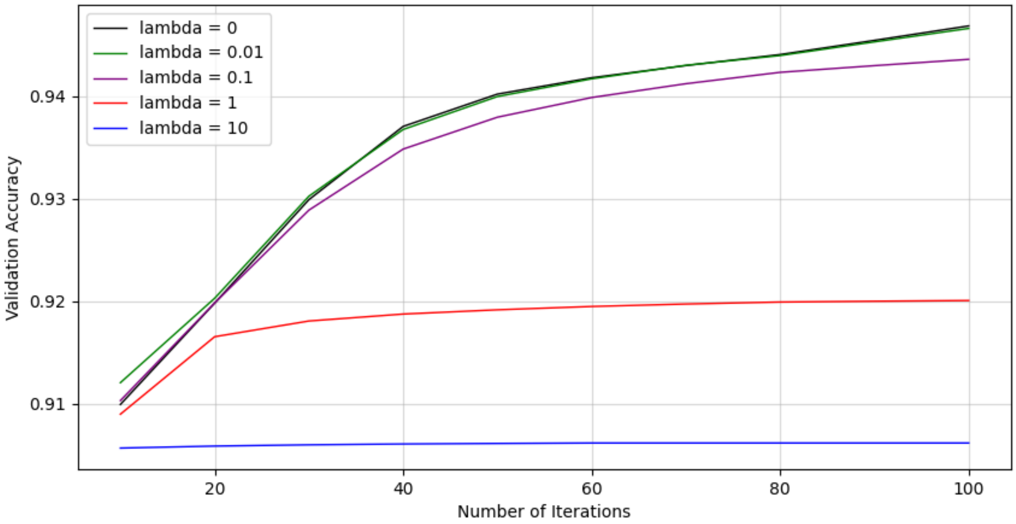
$$\begin{aligned}
\frac{\partial L}{\partial W} &= -\frac{1}{N} \sum_{i=1}^{N} \frac{\partial}{\partial W} \log P(g_i|X_i) \\
&= -\frac{1}{N} \sum_{i=1}^{N} \frac{1}{P(g_i|X_i)} \frac{\partial P(g_i|X_i)}{\partial W} \\
&= \frac{1}{N} \sum_{i=1}^{N} X_i \left(\frac{e^{W^T X_i + b}}{\sum_{j=1}^{K} e^{W_j^T X_i + b_j}} - \mathbf{1}_{g_i} \right)^T \\
&= \frac{1}{N} X^T \left(\frac{e^{WX+b}}{\sum_{j=1}^{K} e^{W_j^T X + b_j}} - Y \right)
\end{aligned}$$

Gradient with respect to b:

$$
\begin{aligned}
\frac{\partial L}{\partial b} &= -\frac{1}{N}\sum_{i=1}^{N}\frac{\partial}{\partial b}\log P(g_i|X_i) \\
&= \frac{1}{N}\sum_{i=1}^{N}\left(\frac{e^{W^T X_i + b}}{\sum_{j=1}^{K} e^{W_j^T X_i + b_j}} - \mathbf{1}_{g_i}\right) \\
&= \frac{1}{N}\left(\frac{e^{WX+b}}{\sum_{j=1}^{K} e^{W_j^T X + b_j}} - Y\right)
\end{aligned}
$$

(Y is the target)

(e) Test Results: The best lambda is 0, and the test accuracy is: 0.9482
Plot: evolution of the accuracy on the validation set



The lambda hyperparameter controls the strength of regularization. A higher lambda value increases the penalty on large wights, which reduces complexity and overfitting. According to the test results the best lambda is 0, indicating that no regularization is needed for this dataset.

Gaussian Naive Bayes

(a)

$$
\begin{aligned}
P(g = k|X) &= \frac{P(X|g = k)P(g = k)}{P(X)} \\
\log P(g = k|X) &= \log\left(\frac{P(X|g = k)P(g = k)}{P(X)}\right) \\
&= \log P(X|g = k) + \log P(g = k) - C(X) \\
\text{where } C(X) &= \log P(X)
\end{aligned}
$$

(b) The log-likelihood for class k is:

$$\begin{aligned}\log L &= \sum_{i|g_i=k} \log P(X_i|g=k)\\ &= \sum_{i|g_i=k} \log \mathcal{N}(X|\mu_k, \Sigma_k)\\ &= \sum_{i|g_i=k} \left[-\frac{d}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(X_i-\mu_k)^T\Sigma_k^{-1}(X_i-\mu_k)\right]\end{aligned}$$

With diagonal $\Sigma_k = \text{diag}(\sigma_1^{(k)} \ldots \sigma_d^{(k)})$:

$$\log L_k = \sum_{i|g_i=k}\left[-\frac{d}{2}\log(2\pi) - \frac{1}{2}\sum_{j=1}^{d}\log(\sigma_j^{(k)})^2 - \frac{1}{2}\sum_{j=1}^{d}\frac{(X_{ij}-\mu_{kj})^2}{(\sigma_j^{(k)})^2}\right]$$

Optimizing for μ_k:

$$\begin{aligned}\frac{\partial \log L_k}{\partial \mu_{kj}} &= \sum_{i|g_i=k}\frac{X_{ij}-\mu_{kj}}{(\sigma_j^{(k)})^2} = 0\\ \mu_{kj} &= \frac{1}{N_k}\sum_{i|g_i=k} X_{ij}\\ \mu_k &= \frac{1}{N_k}\sum_{i|g_i=k} X_i\end{aligned}$$

Optimizing for Σ_k/σ_j:

$$\begin{aligned}\frac{\partial \log L_k}{\partial (\sigma_j^{(k)})^2} &= \sum_{i|g_i=k}\left[-\frac{1}{2(\sigma_j^{(k)})^2} + \frac{(X_{ij}-\mu_{kj})^2}{2((\sigma_j^{(k)})^2)^2}\right] = 0\\ (\sigma_j^{(k)})^2 &= \frac{1}{N_k}\sum_{i|g_i=k}(X_{ij}-\mu_{kj})^2\\ \Sigma_k &= \frac{1}{N_k}\sum_{i|g_i=k}(X_i-\mu_k)(X_i-\mu_k)^T\end{aligned}$$

(e) **GNB Test Results:**

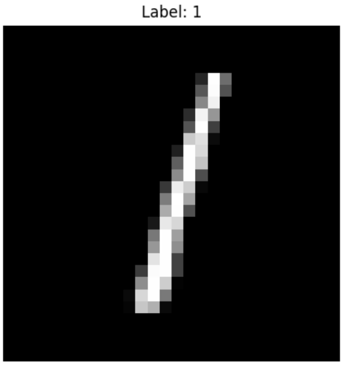
MNIST Dataset: accuracy: 74.69%

Iris Dataset: accuracy: 100.00%

Per class error rates for MNIST:

Class	0	1	2	3	4
Error Rate	22.76%	7.22%	32.95%	54.46%	10.59%

Class	5	6	7	8	9
Error Rate	31.05%	11.69%	23.74%	11.29%	48.46%



Class 1 has the lowest error rate (7.22%) (highest accuracy) due to its simple vertical shape. GNB works better on irises (100% accuracy) because its features are independent, while MNIST's classes (or shapes) are most likely correlated which does not fit well with GNB's assumptions. KNN outperforms GNB on MNIST by considering local patterns without distributional constraints, in other words it works best given that it has no assumptions about the distribution only about local smoothness (or the local neighborhood of points).