Implementing Bayesian Classifiers for MNIST Digit Recognition: From Naive Bayes to KDE

The code is a step-by-step example of how to apply different Bayesian classifiers to the MNIST dataset, which consists of images of handwritten digits.

The presentation also includes the real world use case of ground truth annotation in the health sector (detection of brain tumor) using the techniques which have been learned through the MNIST dataset analysis.

### **1. Loading and Preparing Data:**

* The MNIST data is loaded.
* A small portion of the data is visualized using Matplotlib to see how the digits look.
* Unnecessary columns are removed, and the pixel values are scaled between 0 and 1 to make them easier for the model to work with.

### **2. Naive Bayes Classifier:**

* The code builds a Naive Bayes classifier from scratch.
* It calculates the average (mean), variance, and prior probability for each digit class in the training data.
* Using these values, the model predicts the class of new data based on which one has the highest probability.
* The model is then tested on the test data to see how accurate it is.

### **3. Checking Misclassified Digits:**

* The code finds the digits where the model’s predictions don’t match the actual labels.
* It visualizes some of these misclassified digits to understand where the model made mistakes.

### **4. Dimensionality Reduction:**

* **t-SNE Visualization:** The test data is reduced to two dimensions and plotted, helping to visualize how well the different digits are separated.
* **PCA (Principal Component Analysis):** PCA reduces the number of features from 784 (one per pixel) to 50, making the data easier to manage and improving the model's performance.

### **5. Multivariate Gaussian Bayes Classifier:**

* A more advanced Bayesian model is implemented that takes into account the relationship between features (pixels).
* This model is trained on the reduced data and its accuracy is measured.

### **6. KDE (Kernel Density Estimation) Bayes Classifier:**

* A KDE-based Bayesian classifier is implemented, which uses a technique called Kernel Density Estimation to model the probability distribution of the data.
* The model is trained on the reduced data and tested for accuracy.

### **7. Final Results:**

* The code prints the accuracy of the KDE-based classifier, showing how well it performed on the test data. (97%).

Overall, the code notebook is an educational guide to implementing different types of Bayesian classifiers, using the MNIST dataset as an example. It covers basic to advanced techniques, from simple Naive Bayes to more complex models that consider feature relationships and distributions.

Use Case: Ground Truth Annotation

**Ground Truth Annotation** plays a vital role in the development and training of machine learning models, particularly in the context of medical image analysis. Here’s a summary of the key points:

1. **Purpose of Ground Truth Annotation**:
   * Ground truth annotation is the process of labeling data accurately by experts to serve as a reference for training machine learning models. This annotated data represents the "truth" that the model needs to learn from.
2. **Use Case in Medical Imaging**:
   * We demonstrated an example related to medical imaging, where ground truth annotation is used to label MRI scans. Experts manually annotate regions in these scans to indicate the presence of tumors or other abnormalities.
   * This annotated data is then used to train models, such as Naive Bayes classifiers, to recognize patterns and make predictions about unseen data.
3. **Steps Involved**:
   * **Data Collection**: Thousands of brain scans are collected from patients, including those with and without tumors.
   * **Annotation**: Medical experts manually label these scans, marking areas where tumors are present. This labeled data constitutes the ground truth.
   * **Preprocessing**: The MRI scans are preprocessed to normalize the data and augment it for variations, aiding the model in better generalization.
   * **Feature Extraction and Labeling**: Key features like texture and intensity are extracted and labeled based on the ground truth annotations.
   * **Model Training**: Initial models, such as a simple Naive Bayes classifier, are trained on this well-labeled data. More complex models, like Gaussian Naive Bayes, are then developed to handle more intricate patterns.
4. **Impact of Automated Ground Truth Annotation**:
   * **Efficiency**: Automating the annotation process with machine learning models speeds up the identification of tumors, reducing the need for manual labor.
   * **Accuracy**: Continuous learning and feedback loops help improve the model's accuracy in detecting even small or early-stage tumors.
   * **Scalability**: The techniques used can be applied across different types of medical imaging, broadening their impact in healthcare.

The presentation also illustrates how ground truth annotation is foundational to the effective training of machine learning models in real-world applications, particularly in critical fields like healthcare.