Gaussian Naive Bayes (GNB) Vs. Gaussian Discriminant Analysis (GDA)

Uzair Ahmad

Both Gaussian Naive Bayes (GNB) and Gaussian Discriminant Analysis (GDA) are probabilistic classifiers that make use of the Gaussian distribution. However, they approach the problem of classification differently, and this leads to different assumptions and properties for each model. Here's a comparison of the two:

1. **Model Assumption**:
    - **Gaussian Naive Bayes (GNB)**:
        - Assumes that features are conditionally independent given the class label.
        - For each class, it models the distribution of each feature as a univariate Gaussian distribution.
    - **Gaussian Discriminant Analysis (GDA)**:
        - Models the joint distribution of features and class labels together.
        - Assumes that the data from each class is generated from a multivariate Gaussian distribution.
2. **Parameters**:
    - **GNB**:
        - Because of the conditional independence assumption, it estimates a separate mean and variance for each feature in each class. This leads to $2 \times n \times k$ parameters for k classes and n features.
    - **GDA**:
        - Estimates a mean for each class and a shared covariance matrix (if using Linear Discriminant Analysis) or a separate covariance matrix for each class (if using Quadratic Discriminant Analysis). This means the model can capture correlations between features.
3. **Decision Boundary**:
    - **GNB**:
        - Due to its independence assumption, the decision boundary is linear.
    - **GDA**:
        - The decision boundary can be linear or quadratic, depending on whether the covariance matrices are assumed to be shared across classes (Linear Discriminant Analysis) or distinct for each class (Quadratic Discriminant Analysis).
4. **Use Cases**:
    - **GNB**:
        - Suitable when the dataset is large, and the conditional independence assumption is approximately valid.
        - Particularly popular in text classification problems where feature vectors are high-dimensional.
    - **GDA**:

        - More suitable when the number of training examples is large relative to the number of features and when features have some correlation.
5. **Computational Complexity**:
    - **GNB**:
        - Generally faster and requires less training data because it estimates fewer parameters due to its independence assumption.
    - **GDA**:
        - Requires the inversion of the covariance matrix which can be computationally expensive for a high number of features.
6. **Performance**:
    - **GNB**:
        - Can perform surprisingly well in practice despite its strong independence assumptions, especially when the dependency between features doesn't play a significant role in classification.
    - **GDA**:
        - Can achieve better performance when the assumptions hold, especially the Gaussian distribution assumption and, in the case of LDA, the shared covariance matrix assumption.

In summary, while both GNB and GDA make use of the Gaussian distribution, their core assumptions and the nature of their decision boundaries differ significantly. The choice between them depends on the dataset at hand and the specific problem requirements.