Clasificador y generador de dígitos utilizando el modelo de clasificación Naive Bayes

Introducción

El modelo de Naive Bayes es un algoritmo de aprendizaje automático supervisado utilizado en tareas de clasificación y clasificación de texto. Se basa en el teorema de Bayes, que es un principio de probabilidad condicional que describe cómo se pueden actualizar las creencias o probabilidades sobre un evento dado la evidencia observada.

La "ingenuidad" en el nombre del modelo proviene de la suposición de independencia condicional, que es una simplificación fuerte pero útil que se hace en el modelo. En otras palabras, el modelo asume que todas las características utilizadas para la clasificación son independientes entre sí, lo que es poco realista en la mayoría de las aplicaciones del mundo real. A pesar de esta suposición simplificada, el modelo de Naive Bayes a menudo funciona sorprendentemente bien en muchas tareas de clasificación, como la clasificación de spam, la clasificación de documentos, la detección de sentimientos y más.

El modelo de Naive Bayes se basa en el teorema de Bayes y utiliza la siguiente fórmula para realizar la clasificación:

$$\Pr(\mathrm{C}|\mathrm{X}) = \frac{\Pr(\mathrm{C})\Pr(\mathrm{X}|\mathrm{C})}{\Pr(\mathrm{X})}$$

donde:

- $\Pr(\mathrm{C}|\mathrm{X})$ es la probabilidad condicional de que un ejemplo pertenezca a la clase C dado un conjunto de características X.
- $\Pr(\mathrm{C})$ es la probabilidad apriori de la clase C.
- $\Pr(\mathrm{X}|\mathrm{C})$ es la probabilidad de que las características X se observen dado que el ejemplo pertenece a la clase C.
- $\Pr(\mathrm{X})$ es la probabilidad marginal de observar las características X.

El modelo de Naive Bayes calcula $\Pr(\mathrm{C}|\mathrm{X})$ para cada clase posible y asigna el ejemplo a la clase con la probabilidad condicional más alta.

Las tres variantes más comunes del modelo de Naive Bayes son:

1. Naive Bayes Gaussiano: Asume que las características siguen una distribución normal (gaussiana).
2. Naive Bayes Multinomial: Se utiliza comúnmente en tareas de clasificación de texto, como clasificación de documentos y detección de spam, donde las características son recuentos de palabras o frecuencias de términos.
3. Naive Bayes Bernoulli: Se utiliza para características binarias, como la presencia o ausencia de ciertos términos en un documento.

Ventajas y Desventajas del Clasificador Naive Bayes

Ventajas:

1. Simplicidad y facilidad de implementación: El modelo de Naive Bayes es fácil de entender y de implementar. Su simplicidad lo hace adecuado para problemas de clasificación rápidos y sencillos.
2. Eficiencia computacional: Naive Bayes es un modelo computacionalmente eficiente, lo que significa que puede funcionar rápidamente en grandes conjuntos de datos. Es especialmente útil en aplicaciones en tiempo real.
3. Manejo de datos de alta dimensionalidad: El modelo es robusto en entornos con un gran número de características o dimensiones, como en la clasificación de texto con un gran vocabulario.
4. Requiere menos datos de entrenamiento: En comparación con algunos otros modelos de aprendizaje automático, Naive Bayes puede funcionar bien incluso con conjuntos de datos más pequeños.
5. Buen rendimiento en clasificación de texto: Naive Bayes se utiliza comúnmente en tareas de procesamiento de lenguaje natural, como la clasificación de documentos y la detección de spam, donde ha demostrado un buen rendimiento.

Desventajas:

1. Suposición de independencia condicional: La principal desventaja del modelo es la suposición de independencia condicional entre las características, que rara vez se cumple en situaciones del mundo real. Esto puede llevar a resultados subóptimos en problemas donde las características están altamente correlacionadas.
2. Sensible a características irrelevantes: Naive Bayes puede verse afectado por la presencia de características irrelevantes o ruido en los datos, ya que asume que todas las características son igualmente importantes.

3. No maneja bien datos numéricos continuos: El modelo de Naive Bayes no funciona bien con datos numéricos continuos, a menos que se realice una discretización previa de los mismos.
4. Baja capacidad de representación: Naive Bayes tiene una capacidad de representación limitada y no puede capturar relaciones complejas en los datos. Por lo tanto, puede no ser adecuado para problemas donde se requiere un alto nivel de complejidad en la clasificación.
5. Necesita suficientes ejemplos de cada clase: El rendimiento del modelo depende de la disponibilidad de ejemplos suficientes para cada clase en el conjunto de entrenamiento. Si una clase tiene pocos ejemplos, el modelo puede tener dificultades para hacer predicciones precisas para esa clase.

Objetivo

Utilizar el método de clasificación Naive Bayes y ejemplificar todo el procedimiento para hacer un buen uso de este por medio de la base de datos MINST que contiene un conjunto de 70,000 imágenes en escala de grises, cada una de 28x28 píxeles. Estas imágenes representan dígitos escritos a mano del 0 al 9 y cada imagen tiene una etiqueta asociada que indica qué número representa (0, 1, 2, ..., 9).

Los pasos del método de clasificación utilizando Naive Bayes son los siguientes:

1. Recopilación de datos de entrenamiento.
2. Preprocesamiento de datos.
3. Separación de datos.
4. Estimación de las probabilidades.
5. Entrenamiento del modelo.
6. Clasificación de nuevos ejemplos.
7. Evaluación del modelo.
8. Ajuste y optimización.
9. Despliegue.

Es importante recordar que el éxito de un modelo de Naive Bayes depende en gran medida de la calidad de los datos y de la idoneidad de la suposición de independencia condicional para el problema en cuestión. En muchos casos, Naive Bayes puede ser una opción efectiva y eficiente para tareas de clasificación.

1. Adquisición de datos

Reúne un conjunto de datos de entrenamiento que contenga ejemplos etiquetados, donde cada ejemplo pertenezca a una de las clases de interés. Cada ejemplo debe

estar representado por un conjunto de características que describan las propiedades relevantes del objeto a clasificar.

2. Preprocesamiento de los datos

La etapa de limpieza de datos representa un paso fundamental en el proceso de minería de datos, ya que es en este punto donde surgen las primeras problemáticas relacionadas con la calidad de los datos. La corrección de estos problemas es crucial para asegurar la confiabilidad de los resultados que el modelo producirá en el futuro. En este proceso, se inicia verificando la coherencia de los tipos de datos en cada columna y se comprueba si cumplen con las expectativas. Se examina la presencia de valores nulos, celdas vacías o duplicados, y se realiza un análisis para determinar si los datos se encuentran dentro de los rangos apropiados. En el contexto de los píxeles, por ejemplo, se espera que todos los valores sean números enteros y se sitúen en el intervalo de 0 a 255.

3. Separación de Datos

En el archivo colab se muestra como se trabajó en el tratamiento de datos para construir el modelo de Naive Bayes con las base de datos pedida.

4. Estimación de las probabilidades

Las estimaciones de las probabilidades se realizarán utilizando el modelo de clasificador de Naive Bayes, lo cual cumple con el objetivo de nuestro proyecto.

5. Entrenamiento del Modelo

Se crean los dataframes X_entreno, X_prueba, y_entreno, y_prueba, donde X_entreno es el conjunto de entrenamiento de las características, y_entreno el conjunto de entrenamiento de las categorías, X_prueba el conjunto de testeo de las características y y_prueba es el conjunto de testeo de las categorías.

6. Clasificación de nuevos ejemplos

El proceso de clasificación se llevará a cabo en dos conjuntos de datos: el conjunto de entrenamiento original y el conjunto derivado a través de la técnica PCA. En el caso de los datos transformados mediante PCA, se empleará un clasificador de tipo Gaussian Naive Bayes, ya que las variables presentan un comportamiento continuo

tras la mencionada transformación. Por otra parte, en los valores originales, que consisten en números enteros, se aplicará un modelo de clasificación basado en Bernoulli Naive Bayes.

7. Evaluación del Modelo

Para evaluar el rendimiento del modelo se utilizan las métricas precisión, recall, f1-score y accuracy.

8. Ajuste y Optimización del Modelo

El rendimiento del modelo es satisfactorio y para efecto de este trabajo no se realizará ningún ajuste al modelo.

9. Predicciones

Se realizarán predicciones por medio de los dos modelos que se estimaron al conjunto de datos de testeo.

Despliegue del Modelo

Debido a que el objetivo del proyecto es hacer una aplicación del método de clasifaición Naive Bayes entonces el despliege no es un paso que se va tener encuenta.

Monitoreo y Mantenimiento

Igualmente al paso anterior este paso no es necesario para los objetivos del presente notebook.

Resultado

El resultado se muestra a continuación:

Digitos generados de la base de datos MNIST con Naive Bayes