**Máster de Visión Artificial**

**Asignatura: Reconocimiento de Patrones**

**Práctica 1: Clasificadores generativos**

**Orión García Gallardo**

**DNI: 48330747L**

INTRODUCCIÓN

El objetivo principal de esta práctica es entender los clasificadores paramétricos y no paramétricos. Para ello se va evaluar el rendimiento de diferentes clasificadores:

1. Clasificador paramétrico basado en Gaussianas.
2. Clasificador no-paramétrico basado en los k vecinos más próximos (K-NN).
3. Clasificador no-paramétrico basado en histogramas.
4. Clasificador no-paramétrico basado en ventanas de Parzen.

Para realizar esta evaluación se van a suponer únicamente dos dimensiones en el vector de características de los datos. Se utilizarán tres clases equiprobables para evitar descompensación entre ellas y los datos tendrán una distribución gaussiana y conocida. Con el objetivo de optimizar cada clasificador se van a entrenar con conjuntos de 50, 200 y 1000 datos por clase con dos conjuntos de gausianas diferentes.

Este documento se divide en varias secciones. En la primera de ellas se describe el método desarrollado. A continuación se explica las pruebas realizadas sobre la muestra y se evalúan los resultados obtenidos. Y por último se comenta las conclusiones obtenidas y los flancos que quedan abiertos después de la finalización de esta práctica.

MÉTODO DESARROLLADO

Cada uno de los métodos mencionados tiene algún parámetro que ha de ser introducido previamente, aunque algunos se llamen no paramétricos. Con el objetivo de maximizar el rendimiento de cada clasificador se va a realizar una búsqueda del valor óptimo de los parámetros. La técnica que se va a emplear para realizar esta búsqueda va a ser una validación cruzada de 5 grupos (5-fold). Dicha técnica consiste en dividir los datos de muestra en K subconjuntos. Uno de los subconjuntos se utiliza como datos de prueba y el resto (K-1) como datos de entrenamiento. Dicho proceso se repite durante k iteraciones, donde el conjunto de datos de prueba será un subconjunto distinto en cada iteración. La elección de qué datos habrá en cada subconjunto se realiza de manera aleatoria (usando la función *randperm* de MATLAB).

Clasificador no-paramétrico basado en los k vecinos más próximos (K-NN).

Este método de clasificación consiste en clasificar cada dato teniendo en cuenta la clase de los k datos más cercanos. El número de vecinos k a tener en cuenta será el parámetro a optimizar, con el objetivo de que la clasificación final sea lo más eficiente posible. Como probar con un número de posibles valores de k muy grande podría ser muy costoso limitamos esta búsqueda a unos valores posibles:

VALORES\_K\_EN\_KNN = [1, 3, 5, 7, 11, 13, 17, 19, 21, 23, 29, 31];

Sobre estos valores de k se realizara una la validación cruzada según como hemos explicado previamente. A continuación se ordenará los valores k según la eficiencia de clasificación obtenida para quedarnos con el que mejor clasifique de todos ellos. El último paso será clasificar un conjunto de datos nuevos para comprobar la eficiencia de clasificación obtenida con el k seleccionado.

Uno de los casos excepcionales de este método se puede dar cuando hay un empate de etiquetas de los k vecinos. Es decir, si k = 5, y se tienen 2 etiquetas de la clase 1, 2 etiquetas de la clase 2 y una etiqueta de la clase 3 no se sabría si clasificar el dato en la clase 1 o 2. Para resolver esto se han almacenado también las distancias de los k vecinos al dato dado. De tal modo que en caso de empate se selecciona la clase del dato más próximo. Se ha escogido este entre los múltiples criterios de desempate que existen porque es sencillo, rápido y fácil de implementar.

Otro carácter importante a tener en cuenta con este algoritmo es la medida de distancia seleccionada para determinar quiénes son los k vecinos. En este trabajo se ha seleccionado la norma o distancia Euclídea entre dos datos (usando la función *norm* de MATLAB).

Clasificador no-paramétrico basado en histogramas.

DIVISIONES\_HISTOGRAMA = 3:30;

Calcular aprioris de cada clase

Creacion de histogramas – Datos menores del minimo se ponen en el primer rango y mayores del rango en el último

Entrenar

%crear el histograma de etiquetas en lugar de 3 histogramas con probabilidades para los datos de entrenamiento

%Evaluar los datos de test

encontrar n optima

%crear el histograma de etiquetas con todos los datos de entrenamiento en lugar de 3 histogramas con probabilidades

Ventanas de Parzen

Suponemos una distribución normal gaussiana

Calculamos h con validación cruzada con h = arg max…

EVALUACIÓN

CONCLUSIONES