**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 evaluará 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 gaussianas 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 como conjunto de datos de prueba se selecciona un subconjunto distinto en cada iteración. La elección de qué datos hay en cada subconjunto se va a hacer de manera aleatoria (usando la función *randperm* de MATLAB). El último paso será clasificar un conjunto de datos nuevos para comprobar la eficiencia de clasificación con ese parámetro seleccionado. Con el objetivo de comparar esta eficiencia entre los distintos métodos se va a devolver el error y la matriz de confusión obtenida de evaluar estos nuevos datos.

Clasificador no-paramétrico basado en histogramas.

Este método de clasificación consiste en hacer una división del espacio de trabajo en volúmenes constantes. A continuación se crea un histograma de este espacio de trabajo teniendo en cuenta el número de datos por clase que hay en cada volumen. Para normalizar estos valores se suelen dividir entre el número total de datos de la clase. Es por ello que, para este método, es especialmente interesante que la distribución de datos sea equiprobable entre las clases. Finalmente, dado un nuevo dato perteneciente a un volumen determinado se clasificará en la clase cuyo valor de histograma, o probabilidad, sea mayor en ese volumen.

En este método el parámetro a estimar por validación cruzada será el tamaño de los volúmenes para crear el histograma. Suponiendo conocidos los valores mínimo y máximo de cada dimensión, el problema es análogo al de seleccionar el número de divisiones que vamos a hacer en cada una de estas dimensiones del espacio de trabajo.

En esta práctica se va a acotar los posibles valores de cada dimensión a valores entre 0 y 50. Dicho acotamiento supone que cuando se recibe un dato con valores menor que el mínimo o mayor que el máximo se acumulará a los valores del histograma para los volúmenes del mínimo o máximo respectivamente. Por otro lado, como posibles números de divisiones por dimensión se van a elegir los siguientes:

DIVISIONES\_HISTOGRAMA = 3:30;

La validación cruzada en este caso se encarga de encontrar, entre estos posibles valores en los que se dividen las dimensiones del histograma, él que mejor rendimiento obtenga a la hora de clasificar. Para facilitar los cálculos a realizar y ahorrar memoria, en lugar de crear un histograma de probabilidades por clase, se ha construido un solo histograma donde cada celda contendrá la etiqueta de la clase que mayor probabilidad tenga en ese volumen del espacio de trabajo. Para saber el valor de estas probabilidades es necesario calcular la probabilidad a priori de cada clase, que en el caso de este estudio es igual en cada caso por tener una distribución equiprobable de los datos en cada clase. En caso de empate de probabilidades se clasifica la división con la etiqueta más frecuente entre las en las divisiones vecinas. Dicho algoritmo de construcción de histogramas se emplea tanto cuando se construye a partir del conjunto de datos de entrenamiento en la validación cruzada como cuando se crea para clasificar los datos nuevos y evaluar el clasificador.

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 muy grande de posibles valores de k podría ser muy costoso, limitamos esta búsqueda a unos valores concretos:

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

Sobre estos valores de k se realiza una la validación cruzada según se ha explicado previamente. A continuación se ordenan los valores k según la eficiencia de clasificación para seleccionar el que mejor rendimiento haya obtenido.

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 este problema 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.

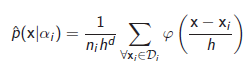
Otro carácter importante a tener en cuenta con este algoritmo es la medida de distancia seleccionada para determinar quiénes son los k vecinos. En este trabajo se ha seleccionado la norma o distancia Euclídea entre dos datos (usando la función *norm* de MATLAB).

Clasificador no-paramétrico basado en ventanas de Parzen

Este método de clasificación consiste en dividir el espacio de trabajo en regiones que no están en posiciones fijas. Usando una función kernel dada, esta técnica aproxima la distribución de un conjunto de entrenamiento a través de una combinación de kernels centrados en los puntos a observar. El ancho de la ventana de estos kernels será el parámetro a optimizar en este algoritmo. Al igual que en los clasificadores anteriores, los valores posibles de ancho de ventana en este algoritmo vendrán determinados de antemano y se corresponderán con:

VALORES\_H\_EN\_PARZEN = [0.5, 1, 2];

La estimación de la función de probabilidad asociada a una determinada clase se calcula usando técnicas de kernel. Esto se traduciría a que, dado un ancho de ventana h, esta función de probabilidad sería:



Donde la aproximación más sencilla de la función sería contar los puntos que están de x a un radio de distancia menor que h, lo cual vendría determinado por:



Sin embargo, es probado que una aproximación que suaviza más la función de probabilidad y, en este caso, da mejores resultados es suponer la gaussianidad de los datos. Esto sería equivalente a definir:



La búsqueda del valor de h óptimo, como en los casos anteriores, se realiza a través de validación cruzada de 5 grupos. En este caso lo que se buscará es el valor de h que satisfaga:



No obstante, es importante remarcar que, para ese trabajo, en este productorio sólo se incluyen los valores de la función de probabilidad en los que la estimación de la clase de un determinado dato de test coincidía con la clase real del dato. Dicho de otro modo, sólo se incluyen los valores en el productorio cuando, para un determinado dato de test su función de probabilidad sea mayor en su clase real. Es por ello que a la hora de elegir el mejor h lo primero que se tendrá en cuenta es la tasa de acierto. En caso de que la mayor tasa de acierto se obtenga por más de un valor se seleccionaría el valor de h cuyo productorio total sea mayor.

EVALUACIÓN

Para realizar la evaluación de los algoritmos se han hecho pruebas con 2 tipos de inicialización de pruebas que se han llamado gaussianas 1 y gaussianas 2. Estas pruebas se han dividido dependiendo del número de datos iniciales. De este modo se han ejecutado los métodos de clasificación para 50, 200 y 1000 datos iniciales. Para obtener resultados más precisos se han realizado un número de tandas de ejecuciones para cada caso y se han calculado las medias aritméticas de los valores devueltos.

GAUSSIANAS 1

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AVERAGE |  | 50 DATOS |  |  |
|  | KNN | HIST | GAUSS | Parzen |
| Elapsed time | 2,864 | 5,749 | 1,324 | 46,311 |
| Error | 0,217 | 0,239 | 0,035 | 0,041 |
| ParameterValue | 3 | 11,667 |  | 1 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AVERAGE |  | 200 DATOS |  |  |
|  | KNN | HIST | GAUSS | Parzen |
| Elapsed time | 11,966 | 5,271 | 1,329 | 248,263 |
| Error | 0,194 | 0,231 | 0,037 | 0,044 |
| ParameterValue | 3 | 12 |  | 0,5 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AVERAGE |  | 1000 DATOS |  |  |
|  | KNN | HIST | GAUSS | Parzen |
| Elapsed time | 174,268 | 7,033 | 1,340 | 3040,363 |
| Error | 0,193 | 0,193 | 0,021 | 0,035 |
| ParameterValue | 19,6667 | 20,3333333 |  | 0,83333333 |

GAUSSIANAS 2

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AVERAGE |  | 50 DATOS |  |  |
|  | KNN | HIST | GAUSS | Parzen |
| Elapsed time | 3,210 | 4,220 | 1,410 | 47,942 |
| Error | 0,053 | 0,081 | 0,050 | 0,063 |
| ParameterValue | 7 | 16,000 |  | 1,16666667 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AVERAGE |  | 200 DATOS |  |  |
|  | KNN | HIST | GAUSS | Parzen |
| Elapsed time | 12,782 | 3,293 | 1,398 | 255,345 |
| Error | 0,044 | 0,067 | 0,032 | 0,037 |
| ParameterValue | 21,6667 | 14,3333333 |  | 0,83333333 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| AVERAGE |  | 1000 DATOS |  |  |
|  | KNN | HIST | GAUSS | Parzen |
| Elapsed time | 181,930 | 4,863 | 1,358 | 3210,131 |
| Error | 0,037 | 0,047 | 0,031 | 0,029 |
| ParameterValue | 17,6667 | 16,3333333 |  | 1,5 |

CONCLUSIONES

Como se puede ver en los gráficos anteriores para ambos tipos de datos gaussianos el algoritmo que mejores resultados devuelve es el basado en gaussianas. Dicho algoritmo optimiza mucho la clasificación al suponer la gaussianidad de los datos, lo cual se corresponde con la realidad. Con respecto a los otros tres clasificadores en el primer conjunto de datos se puede observar que el error es mucho menor para las ventanas de Parzen. Esto se debe a que los otros dos métodos de clasificación no son capaces de estimar que los datos de la primera de las clases están divididos en dos conjuntos en el espacio de trabajo. Sin embargo. hay que remarca el coste computacional del algoritmo de Parzen, tardando más de 3040 segundos de media cuando tenemos más de 1000 datos de entrenamiento. Si lo que se busca es más rapidez una buena opción sería el clasificador basado en histogramas. Este clasificador, tardando siempre menos de 10 segundos de media, da porcentajes de error medio parecidos al knn, incluso cuando se dispone de muchos datos lo llega a igualar. Por otro lado, para el conjunto de datos de entrenamiento obtenidos por la segunda gaussiana, no teniendo en cuenta el clasificador gaussiano, una buena elección sería el knn. En este caso el knn da resultados parecidos a los obtenidos por Parzen, pero con la ventaja de ser bastante más rápido.