RECONOCIMIENTO DE OBJETOS

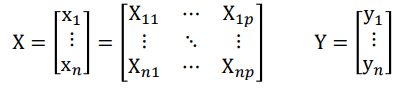
INTRODUCCIÓN

* Clasificación: enfoque basado en la teoría de la Decisión
  + Introducción a problemas de clasificación

Problema de aprendizaje basado en datos: los valores de la salida Y dependen de los valores de entrada de X.

Lo que queremos es encontrar la relación entre las variables de entrada y de salida, este depende de la función f. Por lo tanto, nuestro objetivo es buscar una función estimada de la función que relaciona ambas variables. Para ello necesitamos datos ya que lo estamos haciendo basado en datos.

Ambas variables se disponen en matricialmente:



donde n es el número de instancias y p el número de variables, en y tenemos el valor de salida de cada una de las instancias en vector columna.

Si la variable de salida representa una codificación de clases es un problema de clasificación, mientras que si es una variable numérica cuantitativa es un problema de regresión.

Para realizar un problema de clasificación:

1.- Debemos tener una fuente de datos

2.- Definimos y extraemos los atributos y características de los datos que tenemos para discriminar entre las diferentes clases que nos plantee el problema. Sacar las variables de entrada(predictores).

3.- Seleccionar un tipo de modelo y enseñarlo.

4.- Evaluar el modelo. Fase de test: comprobamos que nuestro algoritmo funciona con datos que están fuera del conjunto de datos del principio.

5.- Integración del modelo. Fase de funcionamiento: si los resultados del test son satisfactorios podemos probar a integrarlo en diferentes sitios.

Cuando sacamos las características debemos escoger las más relevantes, estos es el vector de características. El conjunto de datos lo dividimos en 2, uno para entrenamiento y otro para test.

De los datos de entrenamiento sacamos un clasificador atendiendo a las características explicadas antes y luego se comprueba mediante los datos que hemos dejado para realizar test.

Planteamiento Metodológico Común:

1.- Definición de la respuesta del modelo de predicción y sus posibles valores: se definen las diferentes clases a las que pueden pertenecer los datos.

2.- Creación del conjunto de datos:

2.1.- Extracción de atributos en cada instancia posible: se saca las características de los datos, pueden ser cuantitativa/ordinal/categórica. Si es de clasificación la variable de salida siempre es categórica.

2.2.- Selección de atributos – definición de un vector de atributos: de los atributos que hemos sacado de los datos tenemos que seleccionar mediante técnicas que están destinadas para ello los mejores atributos. Un mal descriptor hará que se de varios errores de clasificación ya que habrá solapamiento.

Requisitos de los atributos:

* Poder de discriminación: las características tienen que poder tomar valores distintos para las diferentes clases. Sensibles, debe haber diferencias para muestras similares de diferentes clases.
* Representatividad: valores similares para muestras de la misma clase.
* Nº de atributos: menor nº posible porque de tener muchos la complejidad puede aumentar.

Atributos de manera individual pueden no ser buenos clasificadores, pero de manera conjunto con otros sí.

3.- Clasificación:

3.1.- Fase de entrenamiento.

3.2.- Fase de test.

* Enfoque basado en la teoría de la Decisión

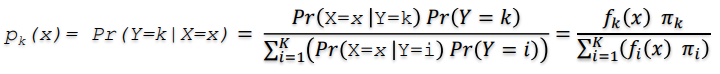
Problema de clasificación: Enfoque basado en la teoría de la decisión: cada clase debe tener una función de decisión propia. Solemos hacer que el dato a clasificar debe estar en la función de decisión cuyo valor sea máximo.

Análisis discriminante:

* Clasificación basada en el Teorema de Bayes: cambiamos d(x) por P(x), esta P proviene de la probabilidad. La suma de P1 y P2 será uno ya que cada uno es la probabilidad de que x pertenezca a la clase P en cuestión. Se asociará a la clase que tenga mayor probabilidad.



Teorema de Bayes:



Probabilidad a posteriori(condicional), se estima cuando sabemos los valores de X.



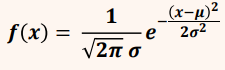
Probabilidad a priori(incondicional), se estima que puede pertenecer a una clase cualquiera:



Debemos de calcularlo tantas veces como clases tengamos. Para hacer un problema con los datos dados debemos sacar las diferentes variables de probabilidad dentro de la formula y así hallaremos la probabilidad que buscamos.

Se suele sacrificar el denominador ya que es el mismo para todas las funciones, por lo que no afecta al resultado. En caso de tener los mismos valores de atributos podemos obviar ese ya que tampoco va a influir al resultado.

Bayes Gaussiano: necesitamos la media y la desviación típica para calcularlo.



* Clasificación basada en distribución normal multivariante:

Análisis discriminante: mientras que antes el vector de predictores eran variables continuas, ahora el conjunto va a proceder de una gaussiana multivariante de tamaño p, también tendremos un vector de medias y una matriz de covarianzas.



Para obtener la media tenemos que sumar los datos y dividirlos entre el número de instancias que tenemos, las variables de manera independiente.

Para calcular las covarianzas de las distintas variables se debe calcular (punto-media)2 entre el número de instancias menos uno. Esto aparece en la diagonal de la matriz de covarianzas, la cual tiene pxp dimensiones.

Para calcular las covarianzas entre números se debe: (punto11-media1)\*(punto12-media2)+(punto21-media1)\*(punto22-media2)+…+(puntof1-media1)\*(puntof2-media2)

/\*Los datos se disponen en vector columna\*/

Coeficiente de correlación:



-1 es que tiene la máxima correlación negativa, 0 es que no tiene ninguna correlación y 1 es que tiene la máxima correlación positiva.

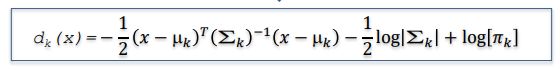
La covarianza es una medida de la relación lineal que hay entre dos variables.

Distancia Euclidea: de2 = (x-μ)’ \* (x-μ)

Distancia de Mahalanobis: dm2 = (x-μ)’ \* Σ-1 (x-μ) -> Matlab: pinv(mCovarianza)

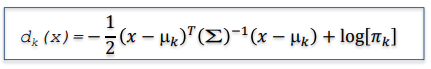
* Clasificador QDA: análisis discriminante cuadrático

Debemos de sacar la vector de medias, matriz de covarianzas y la probabilidad a priori, como esta antes marcado para sustituirlo en la función de densidad de una clase por QDA:



* Clasificador LDA: análisis discriminante lineal -> el problema del clasificador QDA es que se necesita una matriz de covarianzas para cada clase, por lo que de tener muchos atributos o muchas clases se complica mucho el problema. Para ello vamos a emplear este clasificador que lo que hace es suponer que todas las clases tienen la misma matriz de covarianzas.

Función de decisión:



Debemos de estimar la matriz de covarianza, para hacerlo en Matlab:

Necesitamos el número de datos de cada clase ->

mCov = ((numDatosC1-1)\*mCov1+(numDatosC2-1)\*mCov2) / numDatos-numClases

Utilizando la misma matriz de covarianza en la función de LDA obtenemos d1, d2, d12 = d1-d2(es una función que se devuelve de la forma AX1+BX2+C o AX1+BX2+CX3+D) y los coeficientes de dicha función: A, B, y C (y D).

* Casos particulares LDA: clasificadores mínima distancia ->

Clasificador mínima distancia de Mahalanobis: se asume clases equiprobables por lo que la probabilidad a priori es la misma para todas.

Clasificador mínima distancia Euclidea: se supone que lo mismo que en el clasificador mínima distancia de Mahalanobis, no hay correlación entre los predictores y las varianzas de estos son iguales. Si ocurre esto nos podemos olvidar de la matriz de covarianzas.