Minería de Datos

Práctica 2. Evaluación de clasificadores

Ángel Ríos San Nicolás

19 de marzo de 2021

1. Buscar información sobre las siguientes medidas de evauación de clasificadores (las 3 primeras aparecen en Weka):

   - ROC area

     Es una medida de la calidad de un clasificador binario. Se obtiene calculando el área bajo una curva determinada por los falsos positivos frente a los verdaderos positivos. En algunos clasificadores como redes bayesianas o redes neuronales, la respuesta no es una clasificación de 0 o 1 si no una medida de probabilidad o distancia $x \in [0, 1]$ entre las clases de manera que la instancia x se clasifica como 0 o 1 dependiendo de el corte que se establezca. La medida del área bajo la curva ROC mide la bondad del clasificador teniendo en cuenta los falsos positivos y los positivos acertados relativos a todos los posibles cortes y se puede interpretar como la probabilidad de que a una instancia positiva se le asigne un valor más alto que a una negativa. Otra interpretación posible de la curva ROC se obtiene representando la sensibilidad (Recall) frente a la especificidad.

     Veamos un ejemplo del cálculo de esta medida. Supongamos que tenemos la siguiente clasificación:

Real	Clasificador
0	0.2
0	0.4
1	0.5
0	0.55
1	0.6

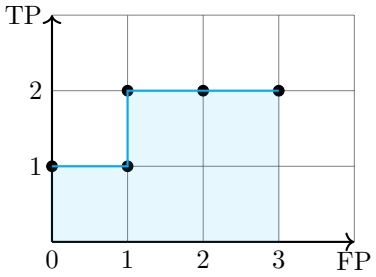
     Si denotamos FP a los falsos positivos y TP a los verdaderos positivos tenemos, tomando sucesivos cortes, lo siguiente:

Corte	FP	TP
0.1	3	2
0.25	2	2
0.45	1	2
0.525	1	1
0.555	0	1

     La curva ROC es la determinada por los segmentos que unen en orden los puntos

     $$[(3, 2), (2, 2), (1, 2), (1, 1), (0, 1)].$$

     Representamos la curva.



El área bajo la curva ROC es 5.

- PRC area

  En lugar de representar los verdaderos positivos frente a los falsos positivos (o la sensibilidad frente a la especificidad) como hace la curva ROC, la curva PRC representa la precisión (Precision) frente a la sensibilidad (Recall) relativos a todos los posibles cortes para la clasificación. La medida PRC area es el área bajo la curva PRC.

- Mathhews correlation coefficient (MCC)

  El coeficiente de correlación de Matthews está dado por la fórmula

$$\mathrm{MCC} = \frac{\mathrm{TP}\cdot\mathrm{TN} - \mathrm{FP}\cdot\mathrm{FN}}{\sqrt{(\mathrm{TP}+\mathrm{FP})(\mathrm{TP}+\mathrm{FN})(\mathrm{TN}+\mathrm{FP})(\mathrm{TN}+\mathrm{FN})}}.$$

  Es un coeficiente de correlación entre la verdadera clasificación y la dada por el modelo. El valor de MCC esta entre -1 y 1. Cuanto mayor sea el valor, mejor será la clasificación de manera que si $\mathrm{MCC} = 1$, la clasificación es perfecta, si $\mathrm{MCC} = 0$ no se diferencia de una clasificación aleatoria y si $\mathrm{MCC} = -1$, la clasificación es totalmente contraria a la verdadera.

- Informedness or Bookmaker Informedness

  La medida Informednsess o Bookmaker Informedness está dada por la fórmula

$$\mathrm{BM} = \mathrm{Recall} + \mathrm{Specificity} - 1.$$

  Esta medida es un estimador de que se tome una decisión informada. El valor de BM está entre -1 y 1. Si es 0, entonces el clasificador asigna la misma proporción de positivos a grupos positivos y negativos con lo que es inútil. Un valor de 1 indica que no hay falsos negativos ni falsos positivos con lo que el clasificador no se equivoca, es decir, el sobreajuste total. En general, cuanto mayor es BM, menores son las proporciones de falsos positivos y negativos y mejor el clasificador. Si BM es negativo, es porque el clasificador está asignando las clases de manera opuesta a como debería y entonces podemos cambiar todas las clasificaciones de manera que sea positivo y tenemos una clasificación mejor.

- Markedness

  La medida Markedness está dada por la fórmula

$$\mathrm{MK} = \mathrm{PPV} + \mathrm{NPV} - 1.$$

  Esta medida cuantifica de manera conjunta la fiabilidad de las predicciones positivas y negativas del modelo. El valor de MK está entre -1 y 1. Si $\mathrm{MK} = 1$ entonces todas las clasificaciones positivas eran realmente positivas y todas las negativas eran realmente negativas. Si $\mathrm{MK} = -1$ es porque todas las clasificaciones positivas son realmente negativas y las clasificaciones negativas son realmente positivas, es decir, todas las clasificaciones son incorrectas. Como antes, si el valor es negativo, podemos cambiar todas las clasificaciones para que sea positivo y la clasificación será mejor cuanto mayor sea MK.

- G-measure

  La G-measure es la media geométrica de la sensibilidad (Recall) y la precisión (Precision), es decir,

  $$G = \sqrt{\text{Recall} \cdot \text{Precision}}.$$

  Es una medida promedio de ambas medidas.

2. Buscar información sobre las siguientes técnicas de validación:

   - Stratified cross-validation

     La validación cruzada estratificada es una técnica de validación cruzada modificada. La diferencia es que en cada iteración, la partición del subconjunto de entrenamiento no se hace de manera totalmente aleatoria sino estratificada, es decir, de manera que cada parte sea estadísticamente representativa del total del conjunto. Esto significa que tiene aproximadamente la misma distribución de la clase (supervisión), misma media, varianza, etc. La validación cruzada estratificada tiene menor sesgo y varianza que la validación cruzada normal.

   - Bootstrap

     La técnica Bootstrap es una técnica de validación mediante muestreo con reemplazamiento. El procedimiento es el siguiente:

     1. Extraemos con reemplazamiento una muestra del mismo tamaño que el conjunto de datos.
     2. Ajustamos el modelo mediante el algoritmo de clasificación.
     3. Calculamos el error de validación en los datos que no se hayan extraído en el muestreo.
     4. Repetimos el proceso un número fijo de veces y calculamos la media de los errores de validación.
     5. Al finalizar las repeticiones, ajustamos el modelo final utilizando todas las observaciones de entrenamiento anteriores.

     Esta técnica es útil cuando se quiere comparar modelos más que estimar de manera precisa ya que tienen menor varianza.

   - Leave-one-out

     La técnica "'Leave-one-out cross-validation" consiste en considerar como conjunto de entrenamiento el total de los datos menos uno que se utiliza para validar el ajuste. Este procedimiento se repite para todos los datos, es decir, se selecciona un dato, se ajusta con el resto, se valida en el dato seleccionado y se calcula el error de validación. Cuando termina el proceso, se estima el error del clasificador promediando todos los errores calculados.

     Este procedimiento reduce la variabilidad porque se emplean todos los datos como entrenamiento y validación y no hay ningún proceso aleatorio. Sin embargo, requiere un elevado coste computacional porque es necesario aplicar el algoritmo de clasificación tantas veces como el número de datos.