Subject: Probability and Statistics

Class XVI: Evaluating Classification Models

UdeA

Bioengineering

Francisco José Campuzano Cardona

Bioengineerer, MSc in Engineering

Evaluación de modelos de Clasificación

UdeA

Los modelos de clasificación tienen diferentes métricas para su evaluación:

Estas métricas se basan en la cuantificación de los siguientes casos:

Verdaderos Positivos (VP): el caso donde se asigna la clase de interés correctamente.
Verdaderos Negativos (VN): el caso donde se asignan las demás clases bien (las otras diferente a la de interés)
Falsos Positivos (FP): el caso donde se asigna la clase de interés pero no era esta la clase correcta.
Falsos Negativos (FN): el caso donde se aginan las demás clases pero no correspondía.

Exactitud: (*Accuracy):* Es una medida general del desempeño del modelo, en general qué tan bien clasifica.

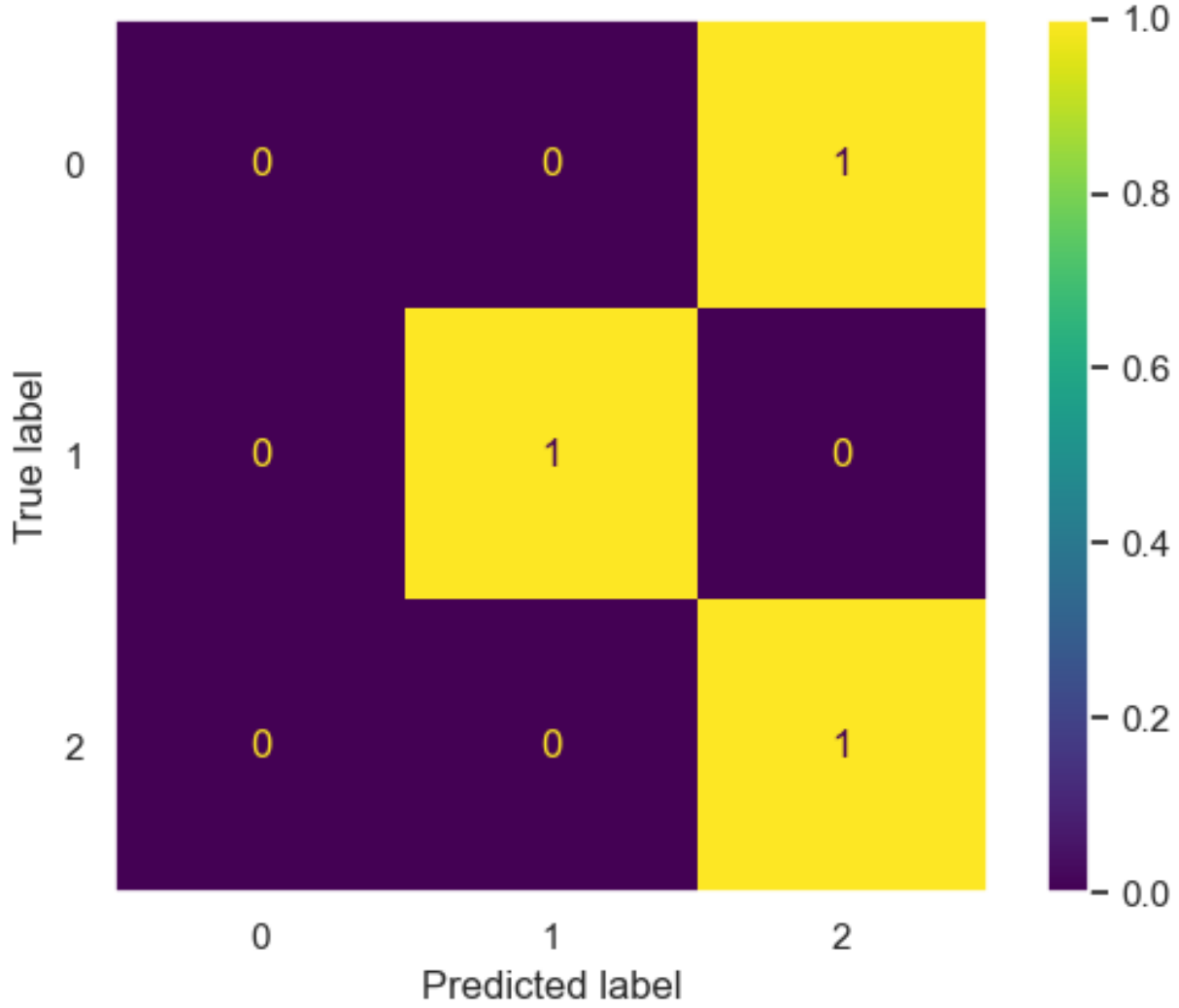
$$accuracy = \frac{VP + VN}{n}$$

Permite decir, en qué porcentaje de las veces el modelo hace una asignación correcta de la clase.

UdeA

La Matriz de confusión:

Es una matriz que permite evidenciar los aciertos y fallos en la clasificación. En el eje y tenemos las etiquetas reales, y en x las etiquetas asignadas según el modelo. La diagonal presenta el numero de etiquetas correctamente asignadas.
Es una matriz simétrica.



Las siguientes métricas se miden para cada clase:

Precisión: En este caso se mide que tan exacto es el modelo para predecir una clase de interés en particular. Es la proporción de instancias clasificadas como positivas que efectivamente lo son. En otras palabras, de los casos positivos que clasificó, qué proporción está bien clasificada.

$$precision = \frac{VP}{VP + FP}$$

Sensibilidad *(recall):* En este caso se mide la fuerza con la que el modelo predice una clase en particular. Es la proporción de casos positivos reales que el modelo identifica como tales.
En otras palabras, de los casos positivos, qué proporción identifica el modelo como tal.

$$recall = \frac{VP}{VP + FN}$$

Especificidad: En este caso se mide la habilidad del modelo para predecir las otras clases, diferentes a la de interés. Es la proporción de instancias diferentes a la de interés que clasificó como tal, y que efectivamente lo eran.

$$especificidad = \frac{VN}{VN + FN}$$

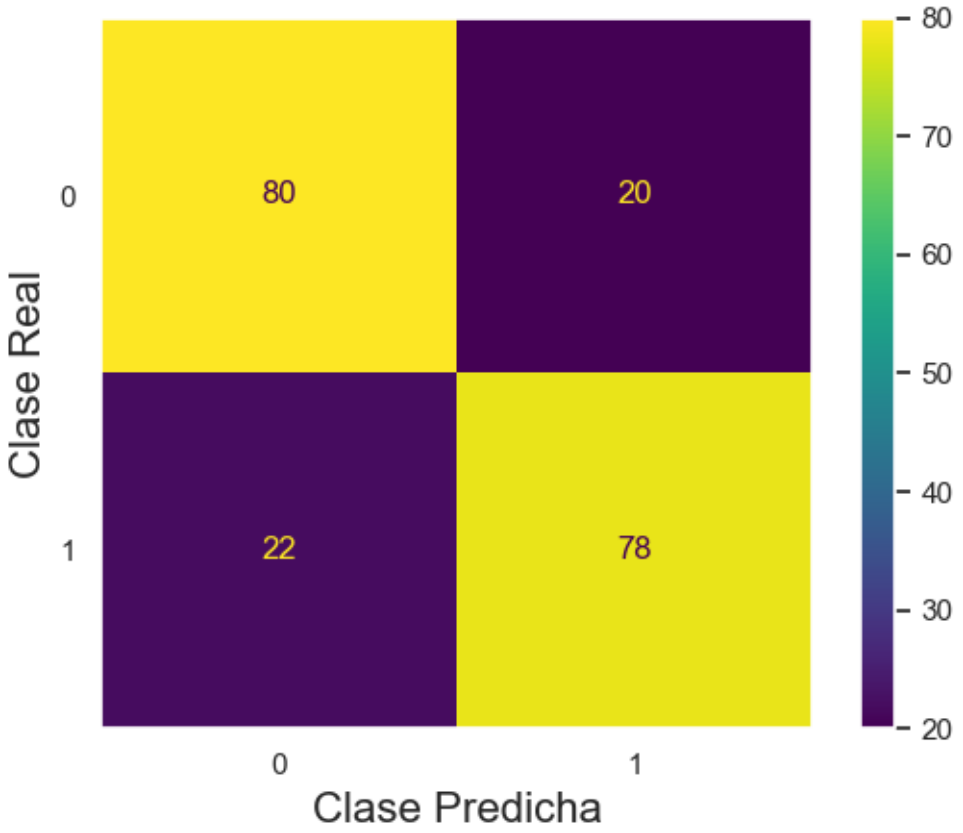
F1-Score: es una métrica que de alguna forma promedia, el valor de *recall*, con el valor de precisión. Siendo 1 un modelo perfecto y 0 uno muy malo. Esta métrica se conoce como promedio armónico entre la precisión y el *recall* y se usa porque la precisión y el *recall* se compensan, si uno crece el otro decrece y viceversa, pero el F1 da una medida 'promedio'.

$$F1 = \frac{2VP}{2VP + FP + FN} = 2 \times \frac{precisión \times recall}{precisión + recall}$$

Evaluación de modelos de Clasificación

UdeA

Calcule las métricas



$$accuracy = \frac{VP + VN}{n}$$

$$precision = \frac{VP}{VP + FP}$$

$$recall = \frac{VP}{VP + FN}$$

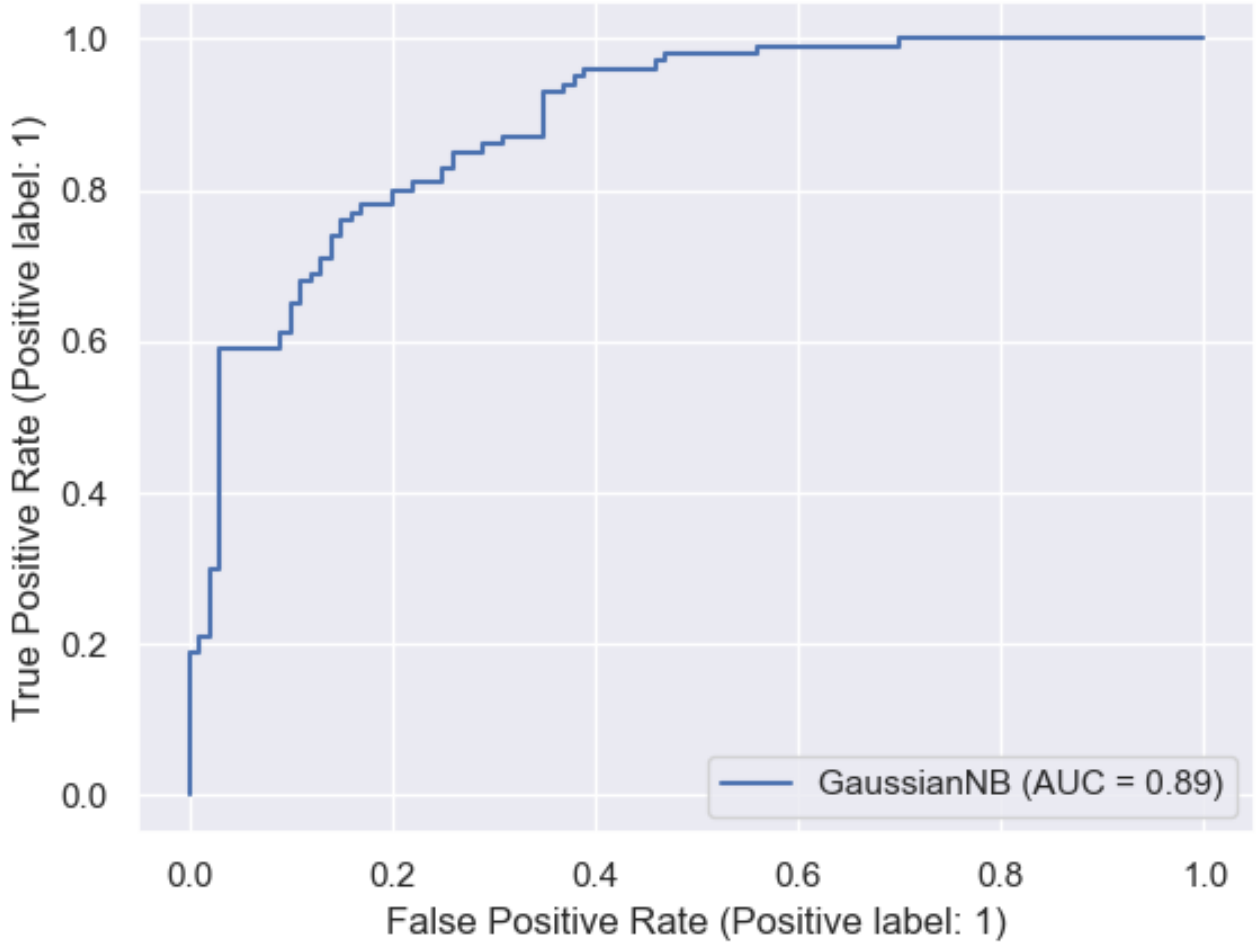
$$especificidad = \frac{VN}{VN + FN}$$

$$F1 = \frac{2VP}{2VP + FP + FN}$$

Evaluación de modelos de Clasificación

UdeA

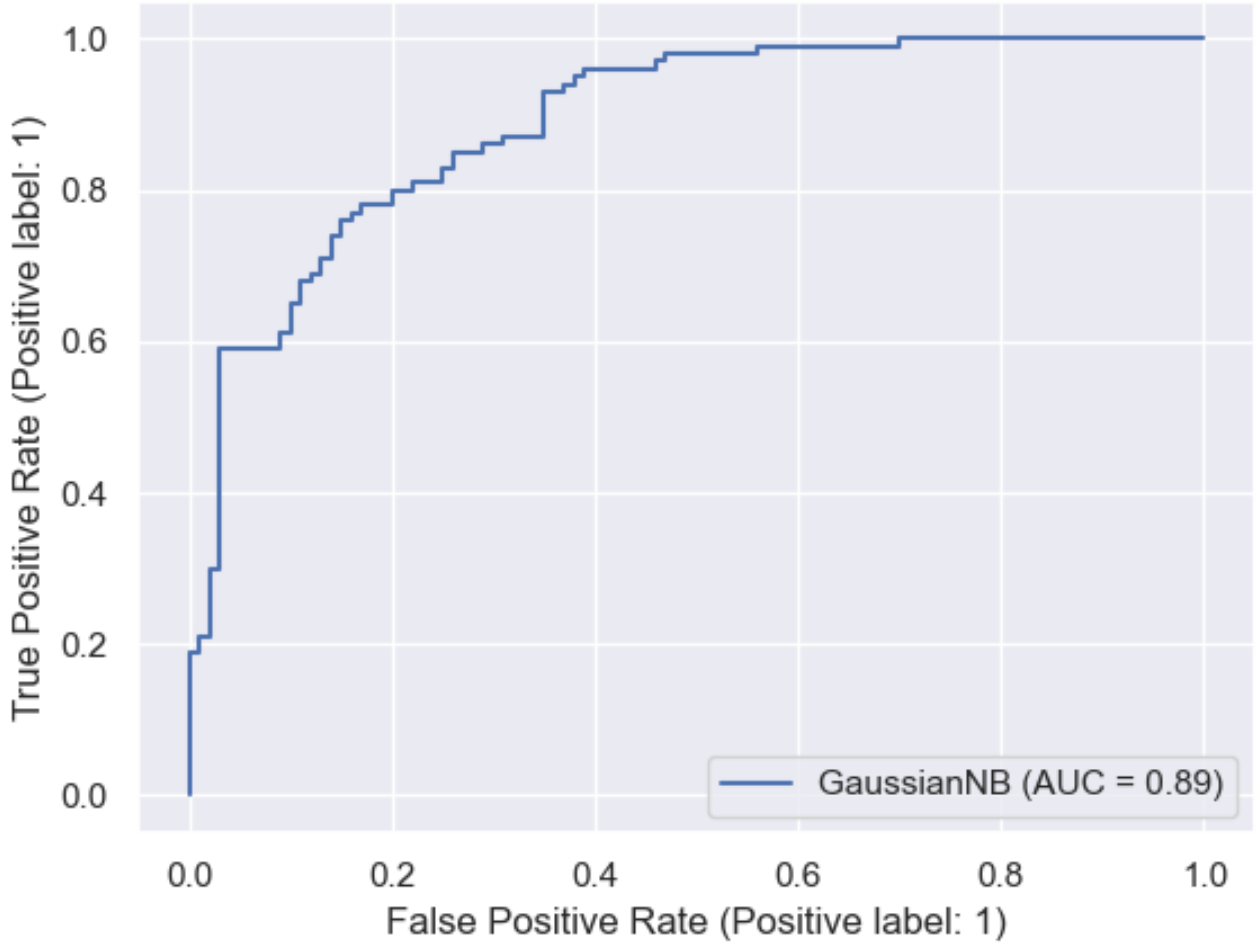
Curva ROC (Receiver Operating Characteristics), y Curva Recall-Precisión. Estas curvas están restringidas a problemas de dos clases, y capturan la compensación entre el Recall y la precisión. La primera hace una curva de la proporción de falsos positivos vs. el Recall, y la un gráfico de Precisión vs. Recall, al variar el umbral de clasificación.



Evaluación de modelos de Clasificación

UdeA

De estos gráficos también se obtiene el área bajo la curva (AUC), que puede variar entre 0.5 y 1. Con 0.5. un clasificador que no es mejor que el asar, y 1 un clasificador perfecto.



UdeA

El problema de casos Raros (*Rare-case problem*).

En ocasiones, un problema de clasificación, puede tener una clase muy poco frecuente, y esta puede ser la clase de interés.

Por ejemplo: suponga que un algoritmo de clasificación quiere clasificar a partir de cierta información clínica, la presencia de un tumor maligno en pacientes. La gran mayoría de la población no tiene tumores malignos, y un pequeño porcentaje, sí. Aquí hay desequilibrio grande entre las clases. Y la clase minoritaria resulta ser la más importante y la llamaremos 1, y el caso donde no hay tumor, la clase 0.

UdeA

El problema de casos Raros (*Rare-case problem*).

Continuando el ejemplo:
La clase de mayor importancia, lo es porque sería muy grave un falso negativo, es decir, clasificar como 0, un paciente que sí era 1. Mientras que clasificar un paciente 0, como 1, no es tan grave, porque seguramente habrá alguna confirmación posterior que lo descartará.

Entonces, en estos casos, el *accuracy* más alto se obtendría en un modelo que clasifique todo como 0, recordemos la fórmula, y pensemos un caso donde 1 sea el 0.1% y 0 el 99.9%:

$$accuracy = \frac{VP + VN}{n}$$

UdeA

El problema de casos Raros (*Rare-case problem*).

Entonces en estos casos, se debe modificar el umbral de probabilidad, que usulamente es 0.5, pero aquí podría ser 0.4, 0.3, o incluso menos. El modelo tendería a clasificar algunos ceros como unos, pero aseguraría que todos los unos sean bien clasificados. Existen métodos para modificar este valor del umbral, que veremos luego.

Subject: Probability and Statistics

UdeA

Bioengineering

¡Thanks!

Francisco José Campuzano Cardona

Bioengineering. MSc in Engineering