Clasificación: Medidas de Evaluación

1EST17 - Aprendizaje Estadístico I

Mg. Enver Gerald Tarazona Vargas

enver.tarazona@pucp.edu.pe

Maestría en Estadística

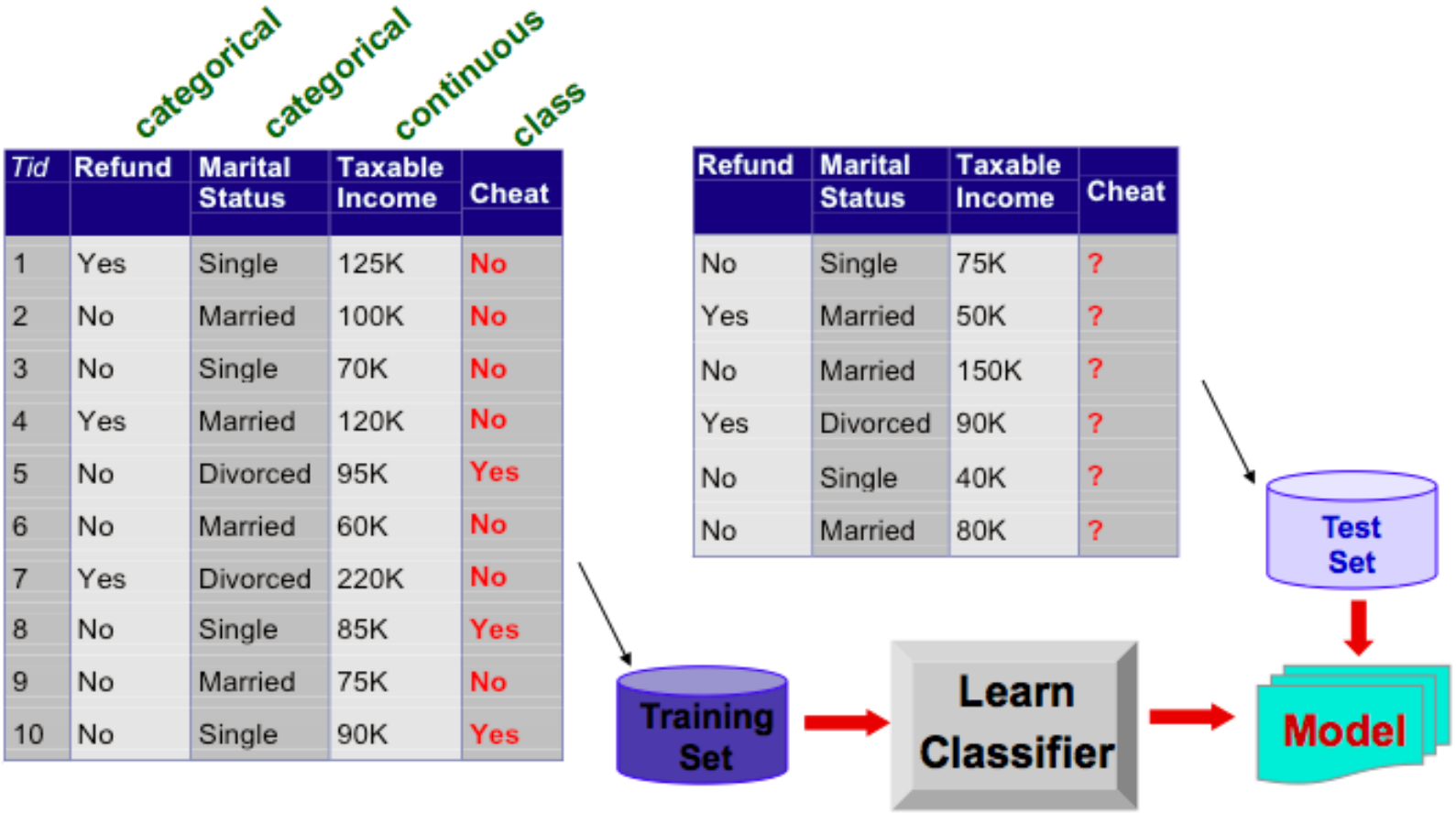
Escuela de Posgrado

Resumen I

Clasificación: Definición

- Dado un conjunto de registros (conjunto de entrenamiento)
  - Cada registro contiene un conjunto de **atributos**, donde uno de ellos es la **clase**.
- Encontrar un modelo para el atributo de clase en función de los valores de los demás atributos.
- Objetivo: Nuevos registros sean asignados a una clase con la mayor precisión posible.
  - Un conjunto de prueba es usada para determinar la precisión del modelo. Usualmente, el conjunto de datos original es dividido en un conjunto de prueba y de entrenamiento, donde el conjunto de entrenamiento es usado para construir el modelo y el de prueba para validarlo.

Clasificación: Ejemplo



Predicción I

- Los modelos de clasificación generan dos tipos de predicciones:
  - Continuas: Usualmente en la forma de una probabilidad (los valores predichos de pertenencia a una clase para un individuo está entre 0 y 1).
  - Categóricas (discretas): Clase predicha.
- Para la mayoría de las aplicaciones prácticas, la predicción de una categoría discreta es necesaria para poder tomar una decisión y es el objetivo de la predicción. Ejemplo: Filtro automático de spam.
- La probabilidad estimada para cada clase puede ser muy útil para medir el ajuste del modelo sobre la clasificación predicha: Un mensaje por e-mail con una probabilidad de ser spam de 0.51 puede ser clasificado de manera similar que otro mensaje con una probabilidad de 0.99
- En algunas aplicaciones el resultado deseado es la probabilidad de pertenecer a una clase, la que será usada como entrada para otros cálculos.

Predicción II

- Ejemplo 1: Una compañía de seguros desea descubrir y procesar reclamos fraudulentos. Usando datos históricos, se puede construir un modelo para predecir la probabilidad de un reclamo fraudulento. Esta probabilidad podría combinarse con los costos de investigación de la compañía y la pérdida monetaria potencial para determinar si la investigación tiene un interés financiero para la institución.
- Ejemplo 2: El CLV (Customer Life Value) está definido como el monto del beneficio asociado con un cliente sobre un periodo de tiempo (Gupta et al. 2006). Para estimar el CLV, varias cantidades son requeridas, incluyendo el monto pagado por un cliente sobre el tiempo en estudio, el costo de mantenimiento del cliente, y la probabilidad de que el cliente realice una nueva compra durante ese tiempo.

Predicción III

- Algunos modelos usados para clasificación, como las redes neuronales y mínimos cuadrados parciales, producen predicciones continuas que no siguen la definición de un valor de probabilidad predicho (los valores no están en la escala de 0 a 1, o no suman 1 las probabilidades de todas las categorías).
- En situaciones como la descrita anteriormente, es posible usar una transformación para coercer las predicciones en un tipo de escala de probabilidad, de tal forma que puedan ser interpretados y usados para la clasificación. Uno de los principales métodos usados para esta finalidad es la transformación SoftMax (Bridle 1990)

$$\hat{p}_l^* = \frac{e^{\hat{y}_l}}{\sum_{l=1}^{C} e^{\hat{y}_l}}$$

donde $\hat{y}_l$ es la predicción numérica para la l-ésima clase y $\hat{p}_l^*$ es el valor transformado entre 0 y 1.

Evaluación de las Clases Predichas I

- Un método común para describir la performance de la clasificación es la matriz de confusión.
- Esta es un simple tabulación cruzada para las clases observadas y predichas.

Evaluación de las Clases Predichas II

Predichos	Observados	
	Eventos	No Eventos
Eventos	TP	FP
No Eventos	FN	TN

Figura: Matriz de confusión para un problema de clasificación con dos clases (eventos y no eventos). Las celdas de la tabla indican el número de verdaderos positivos (TP), falsos positivos (FP), verdaderos negativos (TN) y falsos negativos (FN)

Evaluación de las Clases Predichas III

- La métrica más simple es el ratio de la exactitud total.

$$acc(d) = \frac{TP + TN}{P + N}$$

- o, siendo pesimistas, el ratio de error: $1 - acc(d)$
- Este patrón es un indicador de que el modelo tiene una pobre calibración y también desempeño.
- Existen ciertas desventajas al usar estas estadísticas:
  - En primer lugar, la exactitud no realiza distinciones entre el tipo de error que fue cometido. Ejemplo: En filtros de spam, el costo de borrar erróneamente un e-mail importante es mucho mayor que los costos de permitir que un emailspam pase el filtro. Una discusión sobre este problema puede ser revisado en Provost et al. (1998).
  - En segundo lugar, es importante considerar la frecuencia total de cada clase.

Evaluación de las Clases Predichas IV

- ¿Cuál es el ratio de exactitud que debe ser utilizado como *benchmark* para determinar si un modelo se desempeña adecuadamente?. Se puede usar un ratio no informativo.
- Un ratio no informativo es el ratio de exactitud que se podría alcanzar sin usar un modelo.
- Existen varias formas para definir este ratio:
  - Para un conjunto de datos con C clases, la definición más simple, basada solamente en el azar, es $1/C$. Esta definición no toma en cuenta las frecuencias relativas de las clases en el conjunto de entrenamiento.
  - Una definición alternativa es el porcentaje de la clase de mayor frecuencia en el conjunto de entrenamiento.Modelos con una precisión mayor a este ratio pueden considerarse como razonables.
  - Sobre el efecto severo de clases no balanceadas y posibles medidas remediales ver Kuhn y Johnson (2013).

Evaluación de las Clases Predichas V

- El coeficiente Kappa fue originalmente diseñado para evaluar la concordancia entre dos evaluadores (Cohen, 1960).
- Kappa toma en cuenta la precisión que sería generada por causas aleatorias. Se encuentra definido como:

$$Kappa = \frac{O - E}{1 - E}$$

  donde O es la precisión observada y E la precisión esperada basada sobre los totales marginales de la matriz de confusión.
- El estadístico puede tomar valores entre -1 y 1. Un valor de 0 indica que no hay concordancia entre las clases observadas y las pronosticadas, mientras un valor de 1 indica un perfecta concordancia de la predicción del modelo con las clases observadas.
- Valores negativos indican que las predicciones están en una dirección opuesta, aunque esto raramente ocurre.

Evaluación de las Clases Predichas VI

- Dependiendo del contexto, valores de Kappa entre 0.30 a 0.50 indican una razonable concordancia. Suponga que la precisión para un modelo es alta (90 %) pero la precisión esperada es también alta (85 %), el estadístico Kappa podría mostrar una concordancia moderada (Kappa=1/3) entre las clases observadas y pronosticadas.
- El estadístico Kappa puede ser también extendido para evaluar la concordancia en problemas con más de dos clases. Cuando hay un ordenamiento natural de las clases (p. ej. bajo, medio y alto), una forma alternativa es el estadístico Kappa ponderado puede ser usado para representar sanciones más severas en errores que se alejen más del verdadero resultado. Revisar Agresti (2002) para mayores detalles.

Clasificación Binaria I

- Para dos grupos existen estadísticas adicionales que pueden ser relevantes cuando una clase es interpretada como un evento de interés.
- La sensibilidad o recuperación de un modelo es el ratio en que el evento de interés es predicho correctamente para todas las muestras que contienen el evento.

$$sensibilidad(d) = \frac{TP}{TP + FN}$$

- La sensibilidad es usualmente considerada como el ratio de verdaderos positivos dado que mide la precisión en los eventos de la población.

Clasificación Binaria II

- De manera inversa, la especificidad es definida como el ratio de observaciones que no son los eventos que son predichos como no eventos (ratio de verdaderos negativos).

$$Especificidad(d) = \frac{TN}{FP + TN}$$

- La falsa alarma o ratio de falsos positivos es definido como uno menos la especificidad.

$$Falarm(d) = 1 - Especificidad(d) = \frac{FP}{FP + TN}$$

- La precisión es la comparación entre los verdaderos positivos con las instancias predichas como positivas:

$$Precision(d) = \frac{TP}{TP + FP}$$

Clasificación Binaria III

- Basado en la precisión y la sensibilidad, la medida-F (F1-score) puede ser usada, la cuál es la media armónica entre la precisión y la sensibilidad:

$$F_1(d) = 2\frac{precision * sensibilidad}{precision + sensibilidad}$$

Equilibrio entre sensibilidad y la especificidad I

- Asumiendo un nivel fijo de exactitud para el modelo, normalmente hay un equilibrio que debe hacerse entre la sensibilidad y la especificidad.
- Intuitivamente, incrementando la sensibilidad de un modelo es probable en incurrir en una pérdida de especificidad, debido a que más observaciones son predichas como eventos.
- Equilibrios potenciales entre la sensibilidad y especificidad pueden ser apropiados cuando hay diferentes penalidades asociadas con cada tipo de error. En el caso de filtro de spam, usualmente el foco es la especifidad, debido a que la mayoría de personas está dispuesta a tolerar algunos spams si es que los emails de familiares y colaboradores no van a ser borrados.

Equilibrio entre sensibilidad y la especificidad II

- Usualmente, es de interés in tener una medida única que refleja los ratios de falsos positivos y los de falsos negativos. El índice de Youden (Youden, 1950) definido como:

$$J = sensibilidad + Especificidad - 1$$

  mide las proporciones de observaciones correctamente predichas tanto para los eventos como para los no eventos. En algunos contextos esto puede ser un método apropiado para resumir la magnitud de ambos tipos de errores.
- La curva ROC (receiver operating characteristic) es una de las técnicas más comunes para evaluar la combinación de la sensibilidad y la especificidad dentro de un único valor.

Curvas ROC I

- Los valores pronosticados de la variable dicotómica dependiente Y_i para cada sujeto i son obtenidos en base a las probabilidades estimadas de éxito
- Aquí se predice un éxito si $\hat{\pi} \geq c$, donde c es un punto de corte que a priori puede tomarse como 0.5.
- Sin embargo, dependiendo del contexto y las precisiones que uno quisiera obtener un punto de corte de $c = 0.5$ resulta arbitrario y puede no ser óptimo en casos donde por ejemplo las probabilidades de Éxito son muy extremas.
- La pregunta es entonces, ¿cómo determinar este punto?
- Las curvas ROC (de Receiver Operating Characteristic) nos proveen de una herramienta útil para tal propósito.

Curvas ROC II

- Las curvas ROC (Altman y Bland, 1994; Brown y Davis, 2006; Fawcett, 2006) fueron diseñadas como un método general para que, dado un conjunto de datos, determinar un umbral efectivo tal que los valores sobre el umbral son indicadores de un evento específico.
- Las curvas ROC pueden ser usada para determinar puntos de corte alternativo para las probabilidades de las clases.
- Para cada umbral candidato, el ratio de verdaderos positivos resultante (sensibilidad) y de falsos positivos (uno menos la especificidad) son graficados uno contra el otro.
- Este gráfico es útil para encontrar un umbral que apropiadamente maximice el equilibrio entre la sensitividad y la especificidad.

Curvas ROC III

- También se puede usar como evaluación cuantitativa para contrastar dos o más modelos con diferentes predictores (mismo modelo) o clasificadores distintos (comparación entre modelos), calculando el área debajo de la curva.
- El modelo más optimo debería ser desplazado hacia la esquina superior izquierda de la gráfica.

Una curva ROC se construyen en base a:

Una curva ROC se construyen en base a:

- La sensibilidad (S), definida como $S = \frac{TP}{TP+FN}$; es decir, la proporción de objetos correctamente clasificados como éxitos e, informalmente, conocidos como la proporción de verdaderos positivos.

Una curva ROC se construyen en base a:

- La sensibilidad (S), definida como $S = \frac{TP}{TP+FN}$; es decir, la proporción de objetos correctamente clasificados como éxitos e, informalmente, conocidos como la proporción de verdaderos positivos.
- La especificidad (E), definido como $S = \frac{TN}{FP+TN}$; es decir, la proporción de objetos correctamente clasificados como fracasos.

Una curva ROC se construyen en base a:

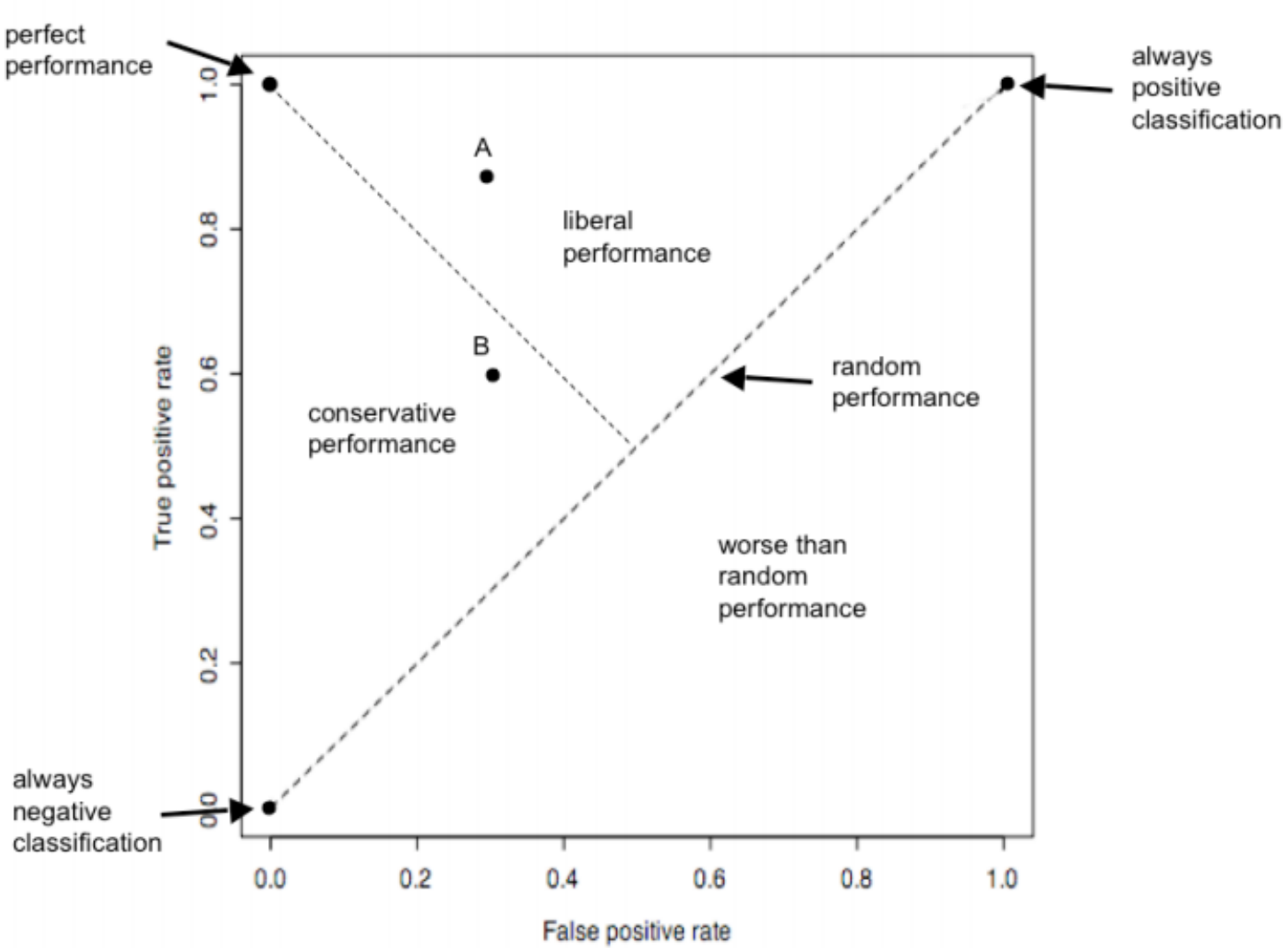
- La sensibilidad (S), definida como $S = \frac{TP}{TP+FN}$; es decir, la proporción de objetos correctamente clasificados como éxitos e, informalmente, conocidos como la proporción de verdaderos positivos.
- La especificidad (E), definido como $S = \frac{TN}{FP+TN}$; es decir, la proporción de objetos correctamente clasificados como fracasos.

La curva ROC no es sino la gráfica de $1 - E = \frac{n_{12}}{n_{.2}}$ o proporción de falsos positivos en el eje de las abscisas frente a la sensibilidad S o proporción de verdaderos positivos en el eje de las ordenadas, para diferentes valores del punto de corte $c \in [0, 1]$.

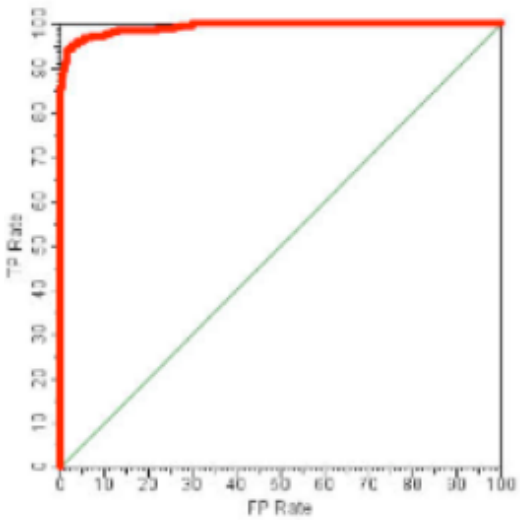
Un modelo ideal sería aquel que tuviera 100 % de sensibilidad y 100 % de especificidad, situándose en el margen superior izquierdo de la gráfica. Y el peor modelo, sería aquel que viniera representado por una línea diagonal desde el margen inferior izquierdo hasta el margen superior derecho. En este último caso, cada incremento en la sensibilidad, vendría asociado a un incremento de igual magnitud en la proporción de falsos positivos. Es obvio, que la mayoría de los modelos se encuentran entre estos dos extremos, y que aquellos modelos que tengan una buena predicción, obtendrán una curva que se alejará de la diagonal para aproximarse hacia el vértice superior izquierdo.

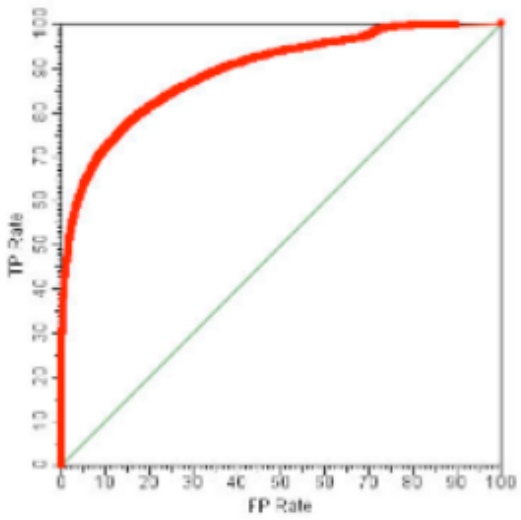
Un modelo ideal sería aquel que tuviera 100 % de sensibilidad y 100 % de especificidad, situándose en el margen superior izquierdo de la gráfica. Y el peor modelo, sería aquel que viniera representado por una línea diagonal desde el margen inferior izquierdo hasta el margen superior derecho. En este último caso, cada incremento en la sensibilidad, vendría asociado a un incremento de igual magnitud en la proporción de falsos positivos. Es obvio, que la mayoría de los modelos se encuentran entre estos dos extremos, y que aquellos modelos que tengan una buena predicción, obtendrán una curva que se alejará de la diagonal para aproximarse hacia el vértice superior izquierdo.
Esta curva nos sirve para objetivar como varían conjuntamente la sensibilidad y la especificidad y comprobar la exactitud del pronóstico en distintos puntos de corte. Por lo general, el mejor punto de corte se sitúa en la zona donde *tuerce la curva*. Una vez obtenido el mejor punto de corte, acorde a los objetivos del estudio, podremos finalmente realizar la clasificación.

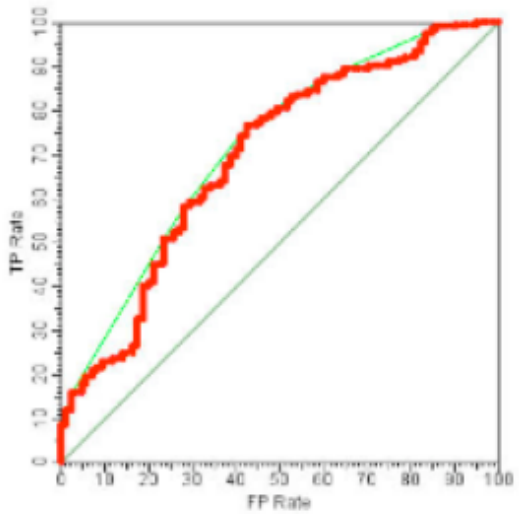
Regiones de la Curva ROC



Ejemplos de Curvas ROC







Equilibrio entre sensibilidad y la especificidad I

- Un aspecto que se suele pasar por alto al evaluar la sensibilidad y la especificidad es que son medidas condicionales.
- La sensibilidad es el ratio de precisión solo para los eventos de la población (y la especificidad para los no eventos)
- Usando la sensibilidad y la especificidad una obstetriz puede hacer afirmaciones como "asumiendo que el feto no tenga síndrome de Down, la prueba tiene una precisión de 95 %".
- Sin embargo, estas afirmaciones pueden ser de poca ayuda para un paciente dado que, para una nueva observación, todo lo que se sabe es la predicción. La persona que usa un modelo predictivo está típicamente interesado en preguntas no condicionales tales como "¿Cuál es la posibilidad de que el feto tenga un desorden genético?".

Equilibrio entre sensibilidad y la especificidad II

- Esto depende de tres valores: la sensibilidad y la especifidad de la prueba diagnóstica y la prevalencia del evento en la población. De manera intuitiva, si el evento es raro, esto debería ser reflejado en la respuesta.
- La prevalencia está definida como

$$Prevalencia = \frac{TP + FN}{TP + FN + FP + TN}$$

- Tomando la prevalencia en consideración, el análogo a la sensibilidad es el valor positivo pronosticado (PPV).

$$PPV = \frac{sensibilidad * Prevalencia}{(Sentitividad * Prevalencia) + ((1 - Especificidad) * (1 - Prevalencia))}$$

- El PPV responde a la pregunta "¿Cuál es la probabilidad de que esta observación sea un evento?"

Equilibrio entre sensibilidad y la especificidad III

- Del mismo modo, el análogo a la especificidad es el valor negativo pronosticado (NPV).

$$NPV = \frac{Especificidad * (1 - Prevalencia)}{(Prevalencia * (1 - sensibilidad) + (Especificidad * (1 - Prevalencia))}$$

- En relación a la estadística Bayesiana, la sensibilidad y la especificidad son las probabilidades condicionales, la prevalencia es la priori, y los valores pronosticados positivos/negativos las probabilidades a posteriori.

Selección de Medidas I

- No existe una respuesta sencilla a la pregunta relacionada a que medida de evaluación usar.
- En general, ningún clasificador es óptimo para todas las métricas de evaluación.
- Cuando se evalúa un problema de clasificación en general, la exactitud es más que suficiente, junto con el análisis del coeficiente Kappa de Cohen.
- Por supuesto, si existen problemas de categorías no balanceadas, uno debería tener en cuenta las medidas F para verificar si existe un buen balance entre la precisión y la sensibilidad.
- En problemas específicos, hay que tener mucho cuidado con la selección de la medida.

Selección de Medidas II

- Por ejemplo, cuando existe un alto costo relacionado a la clasificación de la clase negativa, una falsa alarma muy alta es problemática.
- En algunos casos es recomendable usar múltiples medidas, y buscar un balance entre ellas.