

Regresión logística

Víctor Eduardo
Acevedo Vitvitskaya

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 validarla.

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 email 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 precisión total (o, siendo pesimistas, el ratio de error).
- Este patrón es un indicador de que el modelo tiene una pobre calibración y también desempeño.

Regresión Logística I

- En vez de modelar directamente la respuesta Y , los modelos de regresión logística modelan la probabilidad de que Y pertenezca a una categoría en particular.
- Para la data Default, la regresión logística modela la probabilidad de que un cliente incumpla con el pago de la tarjeta de crédito (moroso).
- Por ejemplo, la probabilidad de que sea moroso dado balance puede ser escrita como

$$Pr(default = Yes | balance)$$

- Los valores de $Pr(default = Yes | balance)$, que puede abreviarse como π , se encontrarán en el rango entre 0 y 1.
- Por ejemplo, es posible predecir default=yes para aquellos individuos en que $\pi > 0.5$

- Alternativamente, si una compañía desea ser más flexible en predecir a los individuos que están en riesgo de ser morosos, es posible elegir un umbral más pequeño, como por ejemplo $\pi > 0.1$

El Modelo Logístico I

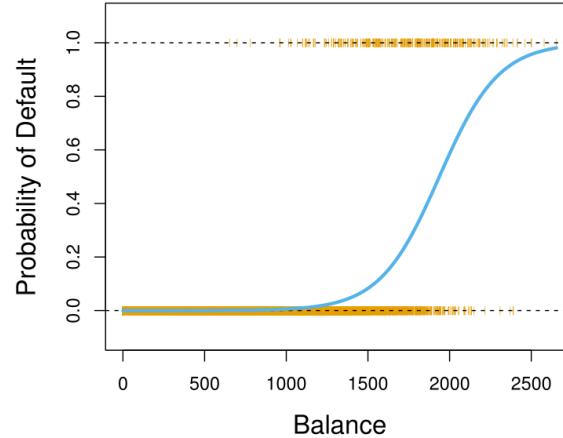
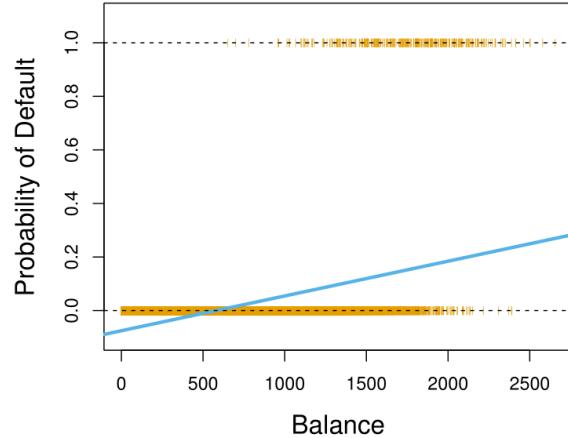
- ¿Cómo debería de modelarse la relación entre $\pi = Pr(Y = 1|X)$ y X ?
- Si se eligiera usar una regresión lineal simple para representar esas probabilidades

$$\pi = \beta_0 + \beta_1 X$$

Se obtendría un modelo estimado similar al mostrado en el lado izquierdo de la siguiente figura.

- Es posible ver el problema con este enfoque: Para valores de balance cercanos a cero se predicen valores de probabilidad de ser moroso negativos. Si se realizan predicciones para valores muy altos de balance se pueden obtener probabilidades mayores a 1.

El Modelo Logístico II



- Para evitar este problema, debemos modelar π usando una función que de como salida un valor que se encuentre entre 0 y 1 para todos los valores de X .

El Modelo Logístico III

- En la regresión logística, es usada la función logística

$$\pi = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

- La gráfica del lado derecho de la figura anterior muestra el ajuste a un modelo de regresión logística para el conjunto de datos Default.
- Se puede observar que el modelo logístico captura mejor el rango de probabilidades que el modelo de regresión lineal mostrado en el lado izquierdo.
- La probabilidad ajustada promedio en ambos casos es 0.0333, la cual es la misma que la proporción total de morosos en la data.
- Con alguna manipulaciones básicas se puede obtener

$$\frac{\pi}{1 - \pi} = e^{\beta_0 + \beta_1 X}$$

El Modelo Logístico IV

- El valor $\frac{\pi}{1-\pi}$ es conocido como odds y puede tomar cualquier valor entre 0 y ∞ .
- Los valores de odds cercanos a 0 o a ∞ indican probabilidades muy bajas o muy altas de ser morosos, de manera respectiva.
- Por ejemplo, en promedio 1 de 5 personas con un odd de $1/4$ será morosa dado que $\pi = 0.2$ implica que los odds son de $\frac{0.2}{1-0.2} = 1/4$
- Del mismo modo, en promedio nueve de cada diez personas con odds de 9 será morosa, dado que $\pi = 0.9$ implica un odds de $\frac{0.9}{1-0.9} = 9$
- Tomando logaritmos a la anterior ecuación se obtiene

$$\log \left(\frac{\pi}{1 - \pi} \right) = \beta_0 + \beta_1 X$$

La expresión del lado izquierdo es conocido como el logit.

El Modelo Logístico V

- A diferencia de la regresión lineal simple, la cantidad que π cambie ante un cambio unitario de X depende del nivel actual de esta variable.
- Si β_1 es negativo, el incremento de X estará asociado con un incremento de π .
- Si β_1 es positivo, el incremento de X estará asociado con una disminución de π .

Regresión Binaria: Definición I

- **Componente aleatorio:** Sean Y_1, \dots, Y_n v.a. dicotómicas independientes. Asumiendo que $y_i = 1$ tiene probabilidad π_i y $y_i = 0$ con probabilidad $1 - \pi_i$:

$$y_i \sim Bernoulli(\pi_i)$$

- **Componente sistemático:**

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} = \mathbf{x}_i^T \boldsymbol{\beta}$$

donde η_i es denominado como predictor lineal y $x_i = (x_{i1}, \dots, x_{ip})'$ es un vector de covariables, donde x_{i1} igual a 1 corresponde al intercepto.

- **Función de Enlace:**

$$g(\pi_i) = \eta_i$$

donde $g(\cdot)$ es una función monótona y diferenciable.

Regresión Binaria: Definición II

- **Función de Respuesta:**

$$\pi_i = \pi(x_i) = h(\eta_i) = h(x_i' \beta), i = 1, \dots, n$$

donde $h(\cdot)$ es una función de distribución acumulativa monótona estrictamente creciente sobre la recta de los números reales. Esto asegura que $h(\eta) \in [0, 1]$ y $g = h^{-1}$

Enlaces Comunes I

- Enlace Logit

$$\log \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\} = \beta_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

- Enlace Probit

$$\Phi(\pi(x))^{-1} = \beta_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

donde $\Phi(\cdot)$ es la f.d.a. de la normal estándar.

- Enlace log-log complementario (cloglog)

$$\log \{-\log(1 - \pi(X))\} = \beta_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

Definición del Modelo I

El modelo de regresión logística múltiple está expresado por:

$$\log \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\} = \beta_1 + \beta_2 x_2 + \dots + \beta_p x_p$$

donde $x = (1, x_2, \dots, x_p)^T$ contienen los valores observados de las variables explicativas. Esto es, se tiene un modelo para el logaritmo de los odds (log-odds) $\log \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\}$. Usando la transformación con la función exponencial se obtiene:

$$\frac{\pi(X)}{1 - \pi(X)} = \exp(\beta_1) \cdot \exp(\beta_2 x_2) \cdot \dots \cdot \exp(\beta_p x_p)$$

lo cual implica que los efectos de las covariables afectan los odds $\frac{\pi(X)}{1 - \pi(X)}$ en una forma exponencial multiplicativa.

Interpretación de los Coeficientes I

Basados en el predictor lineal:

$$\eta_i = \beta_0 + \beta_1 x_{1i} + \dots + \beta_p x_{pi} = \mathbf{x}_i^T \boldsymbol{\beta}$$

Los odds (ventajas):

$$\frac{\pi_i}{1 - \pi_i} = \frac{P(y_i = 1 | x_i)}{P(y_i = 0 | x_i)}$$

siguen el modelo multiplicativo

$$\frac{P(y_i = 1 | x_i)}{P(y_i = 0 | x_i)} = \exp(\beta_0) \cdot \exp(\beta_1 x_1) \cdot \dots \cdot \exp(\beta_p x_p)$$

Si, por ejemplo, x_1 se incrementa en una unidad , los siguientes cambios se aplican a la razón de odds:

$$\frac{P(y_i = 1 | x_{i1}, \dots)}{P(y_i = 0 | x_{i1}, \dots)} / \frac{P(y_i = 1 | x_{i1} + 1, \dots)}{P(y_i = 0 | x_{i1} + 1, \dots)} = \exp(\beta_1)$$

Interpretación de los Coeficientes II

De acuerdo a esto:

- $\beta_i > 0$: $P(y_i = 1)/P(y_i = 0)$ se incrementa.
- $\beta_i < 0$: $P(y_i = 1)/P(y_i = 0)$ disminuye.
- $\beta_i = 0$: $P(y_i = 1)/P(y_i = 0)$ se mantiene constante.
- A partir de lo anterior podemos dar una interpretación a los parámetros del modelo:
 - β_0 es el valor del logit cuando las variables predictoras son nulas.
 - β_j es la variación del logit cuando x_j se incrementa en una unidad y las demás variables se mantienen constantes.
- Alternativamente, podemos también interpretar a e^{β_j} como la variación porcentual del riesgo relativo cuando x_j se incrementa en una unidad y las demás variables se mantienen constantes.

Interpretación de los Coeficientes III

- Nóte que si x_j es una variable dummy asociada a alguna categoría de una variable categórica, entonces esta variación debe interpretarse como un cambio porcentual del riesgo relativo cuando pasamos de la categoría base a la categoría representada por x_j , manteniéndose las demás variables constantes.

Criterios de Ajuste I

- Dado que potencialmente existe una gran cantidad de variables que podrían ser usadas para clasificar Y , es importante tener algunos criterios que nos midan el ajuste de estos modelos.
- Supongamos un modelo de regresión logística (que llamaremos saturado) con k variables predictoras. El modelo más simple que podría considerarse sería uno con sólo el intercepto. Si el modelo saturado brinda información útil para estimar la probabilidad de éxito, entonces este debería ajustar mucho mejor que tal modelo, lo que se vería reflejado en una mayor función de verosimilitud L estimada. Denotando por L_0 y L_S , respectivamente, a las funciones de verosimilitud estimadas bajo los modelos sólo con intercepto y saturado, Cox y Snell propusieron el siguiente coeficiente R^2 de ajuste:

$$R_{CS}^2 = 1 - \left(\frac{L_0}{L_S}\right)^{\frac{2}{n}}.$$

Criterios de Ajuste II

- Claramente mientras más cercano a 1 esté este coeficiente, mejor ajuste tendrá el modelo saturado.
- Nótese sin embargo, que la cota 1 no se alcanza, pues en el mejor de los casos; es decir, en un ajuste perfecto, se tendrá que $L_S = 1$; así el máximo valor que podrá tomar el coeficiente anterior será

$$R_{max}^2 = 1 - L_0^{\frac{2}{n}}.$$

- En tal sentido Nagelkerke propuso corregir el pseudo coeficiente de determinación anterior, para acotarlo al intervalo $[0, 1]$, tomando

$$R^2 = \frac{R_{CS}^2}{R_{max}^2}.$$

Contraste de Hosmer y Lemeshow I

- Cuando el número k de grupos es fijo en un experimento binomial y $\frac{n_i}{n} \rightarrow a_i > 0$ cuando $n \rightarrow \infty$, el desvío $D(\mathbf{y}; \hat{\boldsymbol{\mu}})$ sigue una distribución chi cuadrado con $k-p$ grados de libertad, bajo el hecho de que la hipótesis del modelo adoptado sea verdadera.
- Este resultado no es válido cuando $n \rightarrow \infty$ y $n_i\pi_i(1 - \pi_i)$ queda limitado.
- En ese caso, Hosmer y Lemeshow (1989) sugieren una estadística alternativa para la evaluación de la calidad del ajuste.
- Esa estadística es definida comprando el número observado con el número esperado de éxitos de g grupos formados.
- El primer grupo deberá contener n'_1 elementos correspondientes a las n'_1 menores probabilidades ajustadas, las cuales serán denotadas por $\hat{\pi}_{(1)} \leq \hat{\pi}_{(2)} \leq \dots \leq \hat{\pi}_{(n'_1)}$

Contraste de Hosmer y Lemeshow II

- El segundo grupo deberá de contener los n'_2 elementos correspondientes a las siguientes probabilidades ajustadas $\hat{\pi}_{(n'_1+1)} \leq \hat{\pi}_{(n'_1+2)} \leq \dots \leq \hat{\pi}_{(n'_1+n'_2)}$.
- Los demás grupos se conforman de manera similar, hasta el último grupo que deberá contener las n'_g mayores probabilidades ajustadas $\hat{\pi}_{(n'_1+\dots+n'_g+1)} \leq \hat{\pi}_{(n'_1+\dots+n'_g+2)} \leq \dots \leq \hat{\pi}_{(n)}$.
- El número observado de sucesos en el primer grupo formado será dado por $O_1 = \sum_{j=1}^{n'_1} y_{(j)}$, donde $y_j = 0$ si el elemento correspondiente es un fracaso y $y_j = 1$ si es un éxito.
- Generalizando, obtenemos $O_i = \sum_{j=n'_1+\dots+n'_{i-1}+1}^{n'_1+\dots+n'_i} y_{(j)}$, $2 \leq i \leq g$.

Contraste de Hosmer y Lemeshow III

- El estadístico es definido por:

$$\hat{C} = \sum_{i=1}^g \frac{(O_i - n'_i \bar{\pi}_i)^2}{n'_i \bar{\pi}_i (1 - \bar{\pi}_i)}$$

donde

$$\bar{\pi}_1 = \frac{1}{n'_1} \sum_{j=1}^{n'_1} \hat{\pi}_{(j)} \quad y \quad \bar{\pi}_i = \frac{1}{n'_i} \sum_{j=n'_1+\dots+n'_{i-1}+1}^{n'_1+\dots+n'_i} \hat{\pi}_{(j)}$$

para $2 \leq i \leq g$.

Contraste de Hosmer y Lemeshow IV

- Hosmer y Lemeshow sugieren la formación de $g = 10$ grupos del mismo tamaño (aproximadamente), de modo que el primero grupo contenga n'_1 elementos correspondientes a las $[n/10]$ menores probabilidades ajustadas y así hasta el último grupo con n'_{10} elementos correspondientes a las $[n/10]$ mayores probabilidades ajustadas.
- Cuando no hay empates, esto es, $n_i = 1, \forall i$, es relativamente más fácil formar los 10 grupos con tamaños aproximadamente iguales.
- Por otro lado, cuando hay empates, puede ser necesario que dos individuos con la misma configuración de covariables sean colocados en grupos adyacentes a fin de que los grupos formados no tengan tamaños muy desiguales.
- Hosmer y Lemeshow verificaron a través de simulaciones que la distribución nula asintótica de \hat{C} puede ser bien aproximada por una distribución chi-cuadrado con $(g - 2)$ grados de libertad.

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 (sensitividad) 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 óptimo debería ser desplazado hacia la esquina superior izquierda de la gráfica.

Una curva ROC se construyen en base a:

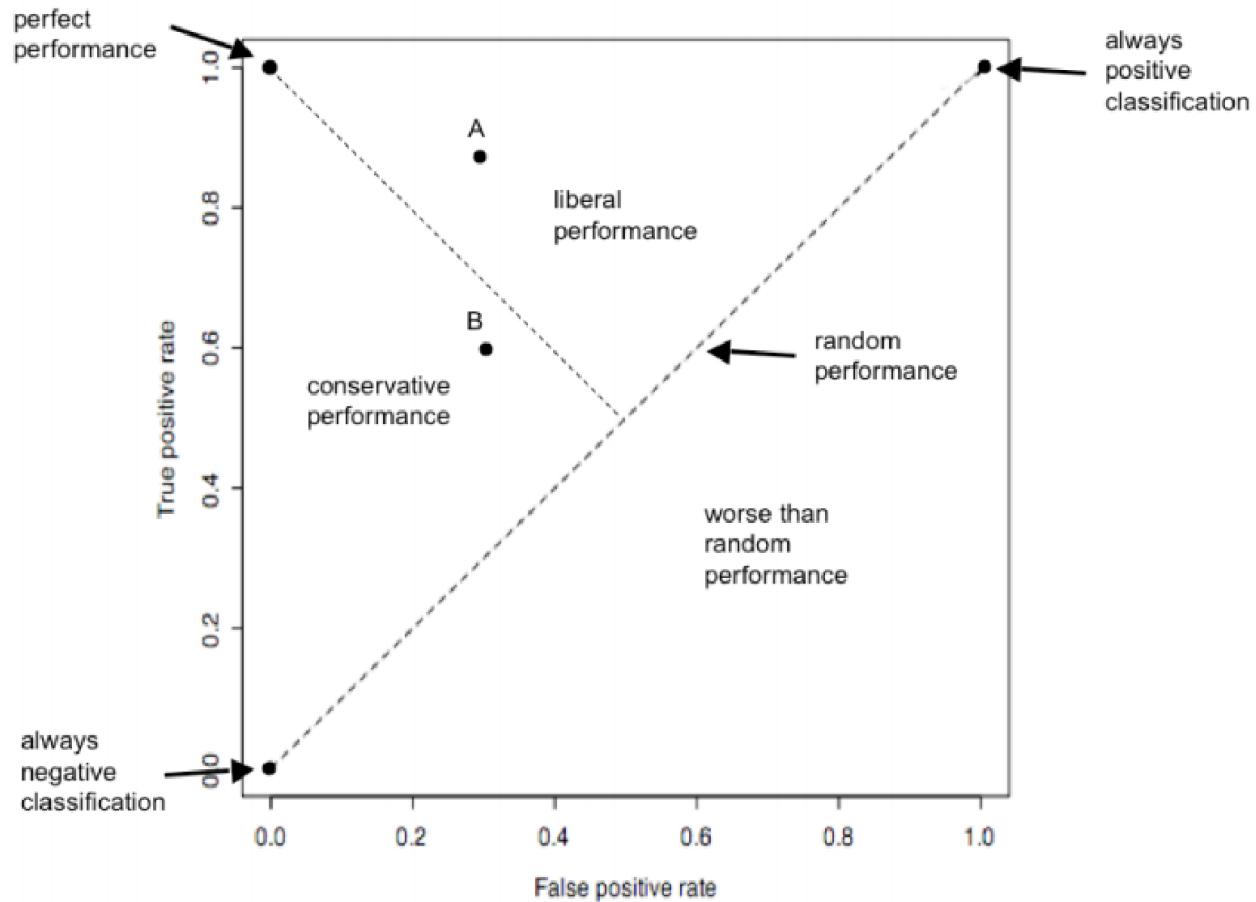
- La sensibilidad (S), definida como $S = \frac{TP}{TP+FN}$; es decir, la proporción de objetos correctamente clasificados como éxitos e, informalmente, conocidos como la proporción de verdaderos positivos.
- La especificidad (E), definido como $E = \frac{TN}{FP+TN}$; es decir, la proporción de objetos 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 $c \in [0, 1]$.

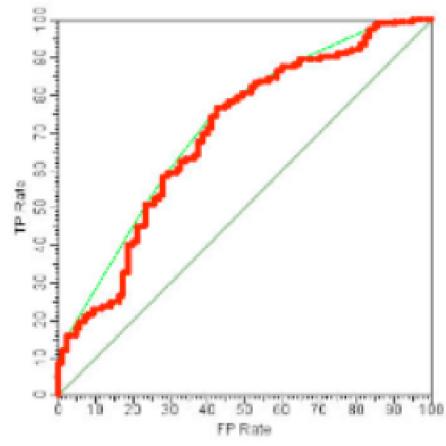
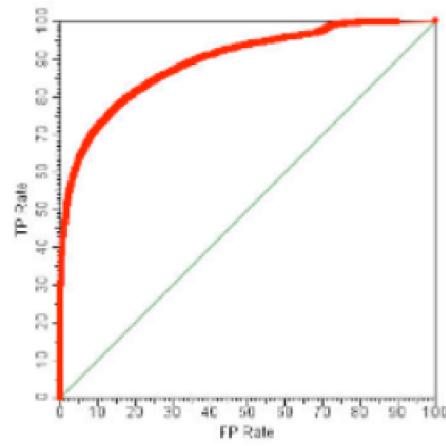
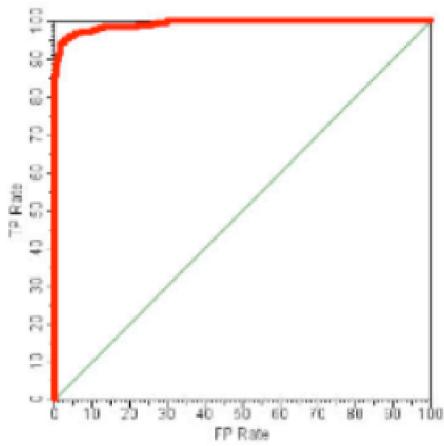
Un modelo ideal sería aquel que tuviera 100 % de sensibilidad y 100 % de especificidad, situándose en el margen superior izquierdo de la gráfica. Y el peor modelo, sería aquel que viniera representado por una línea diagonal desde el margen inferior izquierdo hasta el margen superior derecho. En este 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.

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



Ejemplo: Admisión I

Un investigador está interesado en estudiar el efecto de ciertas variables, tales como el puntaje alcanzado en la prueba GRE (Graduate Record Exam scores), GPA (Grade Point Average) y el prestigio de la institución a nivel de pregrado, en el hecho de que un estudiante sea admitido en un programa de postgrado. La variable respuesta es del tipo binario y tiene los valores de admitido/no admitido (1 y 0 respectivamente). Los datos del ejemplo se encuentran en la dirección:

<http://www.ats.ucla.edu/stat/data/binary.csv>