Fundamentos de Análisis de Datos
Modelos Predictivos

Dra. Andrea Alejandra Rey

Especialización en Ciencia de Datos - ITBA


¿Qué es probable que ocurra en el futuro?
Examinación de datos

Pregunta

¿Cómo podemos hacer predicciones usando un modelo de regresión lineal?

1. Recolectar datos.
2. Ajustar el modelo de regresión a los datos.
3. Verificar la bondad de ajuste del modelo.
4. Usar la ecuación del modelo para predecir el valor de nuevas observaciones.

Regresión lineal

¿Cuándo se pueden hacer prediciones?

- El modelo de regresión lineal puede usarse sólo para predecir valores dentro del rango utilizado para estimar el modelo.
- El modelo de regresión lineal puede usarse sólo para realizar predicciones sobre la población de la cual se tomó la muestra.

Regresión lineal

Estimación puntual

Es el valor predicho al usar el modelo de regresión para hacer predicciones sobre nuevas observaciones.

A pesar de que la estimación puntual representa nuestra mejor estimación del valor de la nueva observación, es poco probable que coincida exactamente con el valor de la nueva observación. Con el fin de considerar esta incertidumbre, podemos utilizar intervalos.

En lo que sigue, n denota la cantidad de puntos sobre los cuales se ajusta el modelo.

Intervalos de confianza

El intervalo de confianza del $(1-\alpha)100\%$ se suele interpretar como que, en una probabilidad del $(1-\alpha)100\%$, la verdadera recta de regresión lineal de la población se encuentra dentro de este intervalo, que se calcula a partir de los datos de una muestra.

Formalmente, se estima un rango de valores medios de $\mathrm{E}(Y|x)$ que, con alta confianza, contengan la verdadera media de los valores dados por los predictores x.

Intervalos de confianza para regresión lineal simple

$$\hat{y} \pm t_{n-2;1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{\frac{1}{n} + \frac{(x-\bar{x})^2}{(n-1)s_x^2}},$$

donde $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n-2}}$ es el error estándar residual.

Intervalos de predicción

Es un rango de valores donde, con una confianza del $(1-\alpha)100\%$, es probable que se encuentre la nueva observación. Observar que cuanto más angosto sea este rango, más precisa será la predicción.

Intervalos de predicción para regresión lineal simple

$$\hat{y}_0 \pm t_{n-2;1-\frac{\alpha}{2}} \hat{\sigma} \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{(n-1)s_x^2}},$$

donde $\hat{\sigma} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$ es el error estándar residual.

El rango del intervalo de predicción siempre es mayor que el rango del intervalo de confianza debido a la gran incerteza de predecir un valor individual en vez de la media.

Intervalos para regresión lineal múltiple

Sea k la cantidad de variables predictoras.
La varianza de los errores de estimación está dada por:

$$s^2_{\text{error}} = \frac{(\boldsymbol{y} - \hat{\boldsymbol{y}})'(\boldsymbol{y} - \hat{\boldsymbol{y}})}{n - k - 1}.$$

Sean $\boldsymbol{x}_i$, $i = 1, 2, \ldots, n$ los vectores de los valores de las variables predictoras para los cuales se construyen los intervalos. Se define:

$$h_i = \boldsymbol{x}_i'(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{x}_i.$$

Intervalos de confianza para regresión lineal múltiple

$$\hat{y}_i \pm t_{n-k-1;1-\frac{\alpha}{2}}\, s_{\text{error}} \sqrt{h_i},\ i = 1, 2, \ldots, n$$

Intervalos de predicción para regresión lineal múltiple

$$\hat{y}_0 \pm t_{n-k-1;1-\frac{\alpha}{2}}\, s_{\text{error}} \sqrt{1 + h_0}$$

Bondad de las predicciones

Para un modelo de regresión con una buena bondad de ajuste, es necesario evaluar cómo es su desempeño al realizar predicciones.

Para $i = 1, 2, \ldots, n$:

1. Eliminar la observación $(\boldsymbol{x}_i, y_i)$.
2. Hallar la ecuación del modelo de regresión con las $n-1$ observaciones restantes.
3. Calcular la predicción $\hat{\hat{y}}_i$ para $\boldsymbol{x}_i$.

Calcular la suma de cuadrados residual de predicciones:

$$\text{PRESS} = \sum_{i=1}^{n} (y_i - \hat{\hat{y}}_i)^2.$$

Bondad de las predicciones

Recordemos que $\text{SST} = \sum_{i=1}^{n}(y_i - \bar{y})^2$.

R^2 de predicción

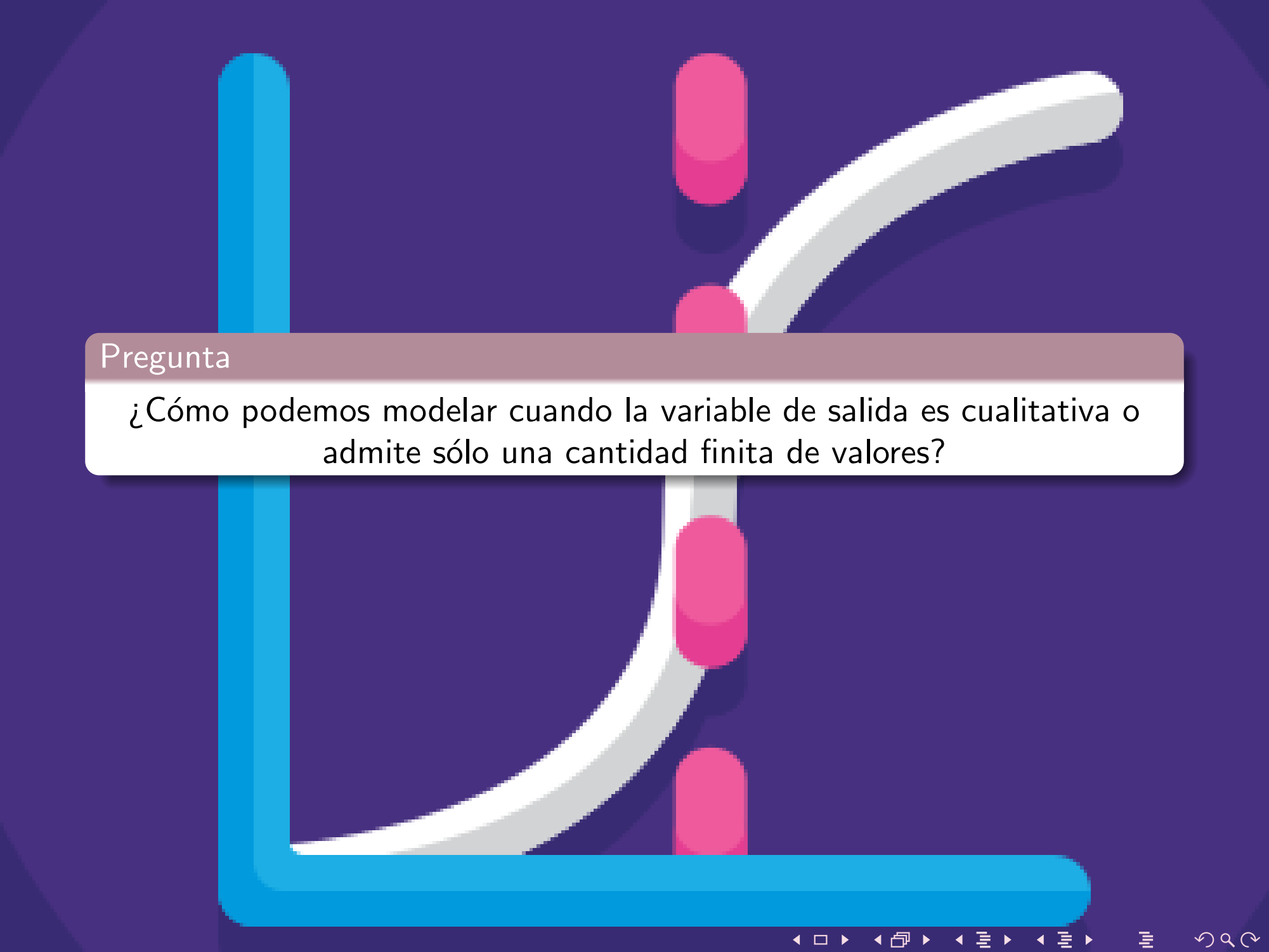
$$R^2_{\text{pred}} = 1 - \frac{\text{PRESS}}{\text{SST}}$$

Observación

Si el valor de R^2_{pred} es considerablemente menor que el de R^2, significa que el modelo no predice tan bien como ajusta, por lo que deberíamos desconfiar de las predicciones realizadas por el modelo.

Ejemplo*

*Base de datos `cars` de `R`



Pregunta

¿Cómo podemos modelar cuando la variable de salida es cualitativa o admite sólo una cantidad finita de valores?

Regresión logística

- Es un procedimiento usado en análisis de datos que encuentra relaciones entre dos factores para luego utilizarlas en la predicción del valor de uno de esos factores en función del otro.
- Es una técnica similar a la regresión lineal múltiple pero con el agregado de que la respuesta predicha es una probabilidad.
- Permite predecir la probabilidad de que ocurra un evento (valor de 1) o no (valor de 0) a partir de la optimización de los coeficientes de regresión.
- Ayuda a comprender las relaciones entre los datos y prever los resultados, posibilitando una mejor toma de decisiones.

Regresión logística

Aplicaciones

- **Fabricación:** La regresión logística se utiliza para estimar la probabilidad de falla de las máquinas de producción con el fin de planificar los programas de mantenimiento.
- **Sanidad:** La regresión logística se utiliza para estudiar el impacto de antecedentes familiares o genéticos con el fin de planificar tratamientos preventivos.
- **Finanzas:** La regresión logística se utiliza para analizar transacciones financieras con el fin de evitar fraudes.
- **Aseguradoras:** La regresión logística se utiliza para evaluar solicitudes de seguros con el fin de buscar posibles riesgos.
- **Marketing:** La regresión logística se utiliza para predecir el impacto de un anuncio en línea con el fin de mejorar el rendimiento publicitario.

Regresión logística

Ventajas ☺

- **Simplicidad:** son modelos menos complejos que otros utilizados en aprendizaje automático (*machine learning*).
- **Velocidad:** son modelos que pueden procesar grandes volúmenes de datos a alta velocidad debido a que requieren poca memoria y potencia de procesamiento.
- **Flexibilidad:** son modelos aplicados para hallar respuestas a preguntas dicotómicas (admiten sólo dos respuestas) o con una cantidad finita de resultados, aunque también pueden utilizarse para preprocesar datos.
- **Visibilidad:** son modelos que ofrecen mayor visibilidad de los procesos de programación internos, permitiendo solucionar problemas o corregir errores de una manera más sencilla.

Tipos de regresión logística

Binaria

Se utiliza en problemas de clasificación binaria donde sólo existen dos resultados posibles. Funciona redondeando los valores de la función logística por cercanía al 0 o al 1.

Multinomial

Se utiliza en problemas de clasificación con un número finito de resultados posibles. Funciona agrupando el resultado a los valores posibles más cercanos.

Ordinal

Se utiliza en problemas en los que los números representan rangos en lugar de valores reales.

Regresión logística

Pasos

1. Formular una pregunta.
2. Identificar los factores.
3. Recopilar datos.
4. Procesar los datos con el modelo de regresión logística.
5. Realizar predicciones para valores desconocidos.

Función logística

Definición

La **función logística** se define para $p \in [0, 1]$ como:

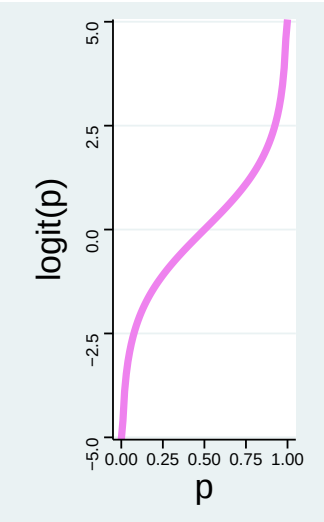
$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right).$$

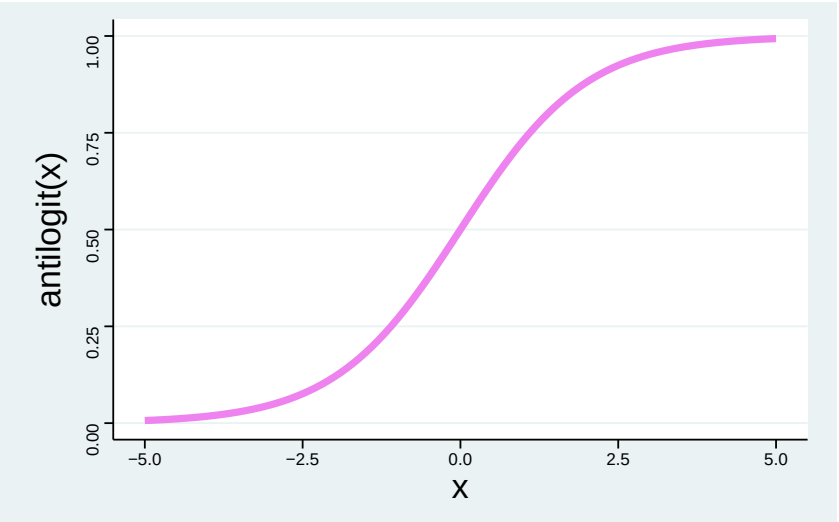
Su función inversa se define para $x \in \mathbb{R}$ como:

$$\text{antilogit}(x) = \frac{e^x}{1+e^x}.$$

Función logística

Gráficas





Observación

La inversa de la función logística devuelve valores sólo entre 0 y 1 para la variable dependiente.

Modelo de regresión logística

Sea $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ un vector de observaciones de una variable respuesta binaria; es decir, $y_i \in \{0, 1\}$ para todo $i = 1, 2, \ldots, n$.

El modelo logístico se basa en el supuesto de que la probabilidad de observar $y_i = 1$ dado un valor particular $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ está dada por:

$$p(x_i) = P(y_i = 1|x_i) = \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}.$$

Entonces:

$$1 - p(x_i) = P(y_i = 0|x_i) = 1 - \frac{\exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}$$

$$= \frac{1}{1 + \exp\left(\beta_0 + \sum_{j=1}^{k} \beta_j x_{ij}\right)}.$$

Modelo de regresión logística

Observemos que:

$$\ln\left[\frac{p(x_i)}{1-p(x_i)}\right] = \ln[p(x_i)] - \ln[1-p(x_i)]$$

$$= \ln\left[\frac{\exp\left(\beta_0 + \sum_{j=1}^{k}\beta_j x_{ij}\right)}{1+\exp\left(\beta_0 + \sum_{j=1}^{k}\beta_j x_{ij}\right)}\right] - \ln\left[\frac{1}{1+\exp\left(\beta_0 + \sum_{j=1}^{k}\beta_j x_{ij}\right)}\right]$$

$$= \ln\left[\exp\left(\beta_0 + \sum_{j=1}^{k}\beta_j x_{ij}\right)\right] = \beta_0 + \sum_{j=1}^{k}\beta_j x_{ij}.$$

Esto es equivalente a un modelo log-lineal para el cociente $\frac{p(\boldsymbol{x})}{1-p(\boldsymbol{x})}$.

Modelo de regresión logística

Pregunta

¿Cómo estimamos los coeficientes de regresión del modelo?

Estimador de máxima verosimilitud

Se base en hallar los valores de los parámetros que maximicen la probabilidad de que el proceso descripto por el modelo produzca los datos que realmente se observaron.

Sea $X_1, X_2, \ldots, X_k$ una muestra aleatoria de una distribución con parámetro θ.
Supongamos que observamos $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$.
Evaluando la función de densidad de probabilidad conjunta en los datos observados de la muestra, obtenemos una función del parámetro.

Estimador de máxima verosimilitud

Función de verosmilitud

$$L(x_1, x_2, \ldots, x_k; \theta) = f_{X_1, X_2, \ldots, X_k}(x_1, x_2, \ldots, x_k; \theta).$$

Bajo el supuesto de independencia,

$$L(x_1, x_2, \ldots, x_k; \theta) = \prod_{i=1}^{k} f_{X_i}(x_i; \theta).$$

Estimador de máxima verosimilitud

Conocido por su equivalente en inglés *maximum likelihood estimator* (MLE), se define como:

$$\hat{\theta}_{MV} = \arg\max_{\theta} L(x_1, x_2, \ldots, x_k; \theta).$$

Estimador de máxima verosimilitud

En muchos contextos es más sencillo trabajar con su logaritmo.

Función de log-verosmilitud

$$\ell(x_1, x_2, \ldots, x_k; \theta) = \ln L(x_1, x_2, \ldots, x_k; \theta).$$

Bajo el supuesto de independencia:

$$\ell(x_1, x_2, \ldots, x_k; \theta) = \sum_{i=1}^{k} \ln[f_{X_i}(x_i; \theta)].$$

Estimador de máxima verosimilitud: Regresión lineal

El modelo lineal sigue una distribución Normal:

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \sim \mathcal{N}(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I}).$$

La función de verosmilitud es:

$$L(\boldsymbol{\beta}, \sigma^2) = (2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right].$$

La función de log-verosmilitud es:

$$\ell(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}).$$

Derivando, igualando a cero y resolviendo el sistema se encuentan los estimadores de máxima verosimilitud.

Estimador de máxima verosimilitud: Regresión lineal

Estimación de los coeficientes

El estimador de máxima verosimilitud para β coincide con el estimador de mínimos cuadrados.

Estimación de la varianza

El estimador de máxima verosimilitud es:

$$\hat{\sigma}^2_{\text{MV}} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2.$$

$\hat{\sigma}^2_{\text{MV}}$ es un estimador sesgado de σ^2, aunque es asintóticamente insesgado.

Estimador de máxima verosimilitud: Regresión logística

Para observaciones independientes idénticamente distribuidas (iid), la función de verosmilitud es:

$$L(\beta_0, \beta) = \prod_{i=1}^{n} [p(x_i)]^{y_i} [1 - p(x_i)]^{1-y_i}.$$

La función de log-verosmilitud es:

$$\ell(\beta_0, \beta) = \sum_{i=1}^{n} \{y_i \ln[p(x_i)] + (1 - y_i) \ln[1 - p(x_i)]\}.$$

Derivando, igualando a cero y resolviendo el sistema se encuentan los estimadores de máxima verosimilitud.

Regresión logística

Limitaciones

- Las variables independientes deben ser válidas, puesto que variables incorrectas, faltantes o incompletas degradan el valor predictivo del modelo.
- Se deben evitar los resultados continuos, los que hacen que el modelo sea mucho menos preciso.
- Si algunas observaciones están relacionadas entre sí, el modelo tenderá a sobrestimar su importancia.

Ejemplo*

*Base de datos `PimaIndiansDiabetes2` del paquete `mlbench` de `R`

Clasificación

Realizar una predicción usando el modelo de regresión logística es equivalente a asignar una clase o nivel a nuevas instancias observadas.

Un modelo de regresión logística puede verse como un **clasificador** puesto que el mismo posibilita realizar una tarea de **clasificación**.

Pregunta

¿Cómo podemos evaluar el rendimiento de un clasificador?

Evaluación de un modelo de clasificación

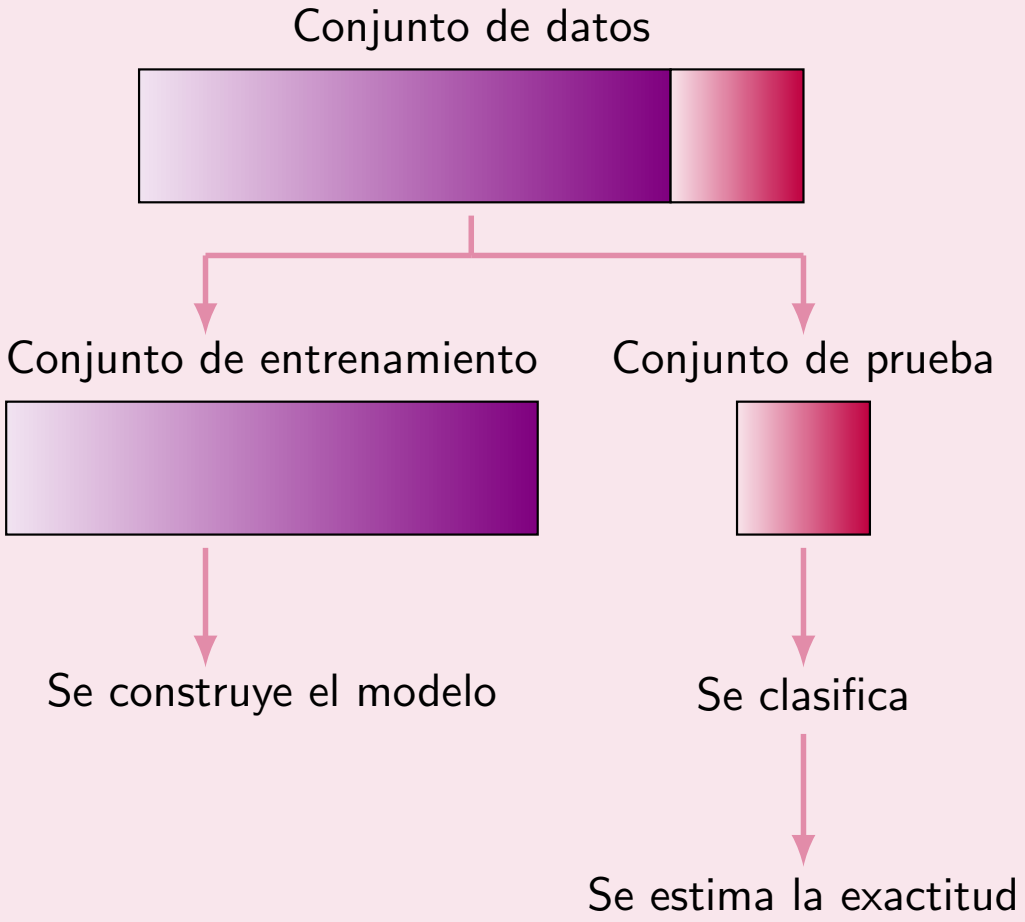
Exactitud predictiva

Es la proporción de instancias que han sido correctamente clasificadas.

La exactitud predictiva suele estimarse a partir de un conjunto de datos que no fue usado para construir el modelo, para lo cual las estrategias más utilizadas son:

- división de los datos en un conjunto de entrenamiento y un conjunto de prueba,
- validación cruzada k-fold,
- validación cruzada N-fold,
- método de *bootstrap*.

Conjuntos de entrenamiento y prueba



Conjuntos de entrenamiento y prueba

Una partición del $a-(100-a)\%$ implica que el conjunto de entrenamiento contiene el $a\%$ del conjunto de datos, mientras que el conjunto de prueba contiene al $(100-a)\%$ restante.

No hay una regla establecida para el tamaño de cada conjunto, aunque las particiones más usadas son $60-40\%$, $70-30\%$, $75-25\%$, y $80-20\%$.

Si el conjunto de prueba contiene N datos de los cuales C fueron correctamente clasificados, la exactitud se estima como:

$$\hat{p} = \frac{C}{N}.$$

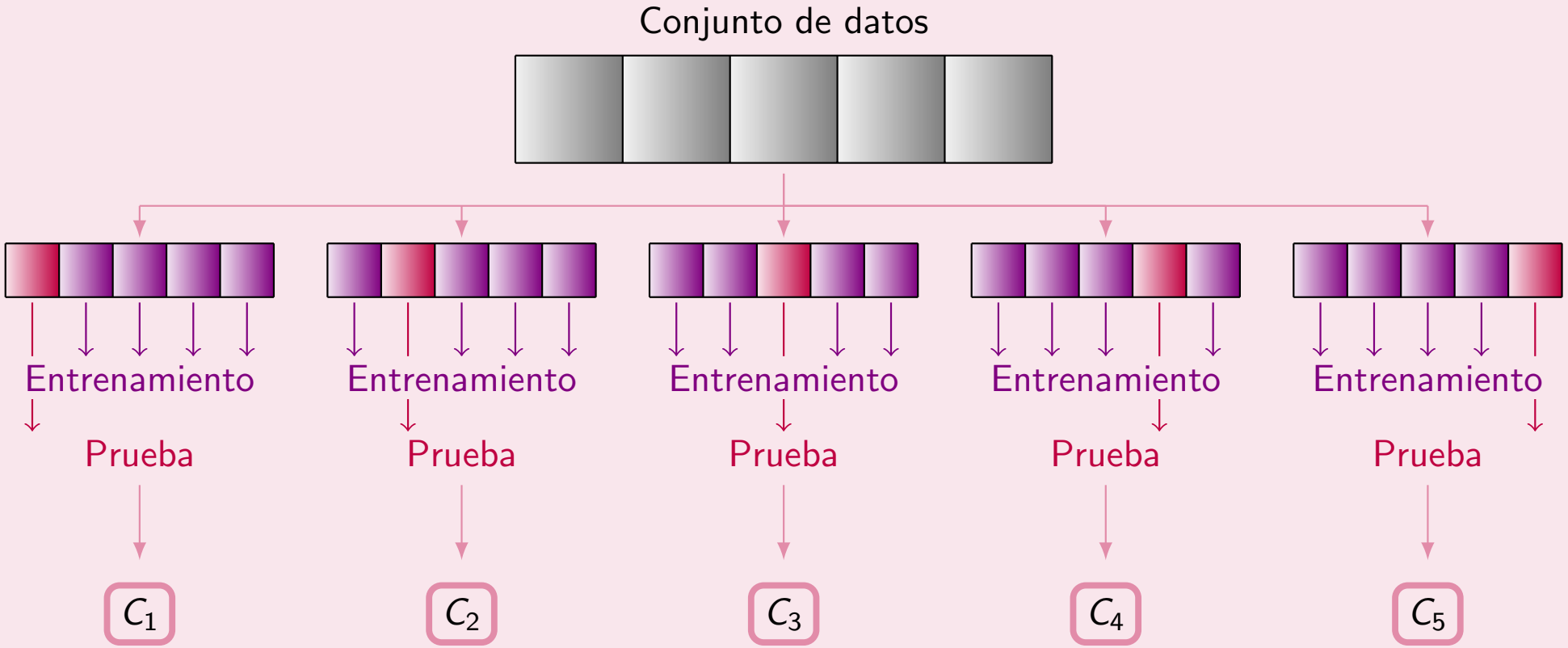
El intervalo de confianza del $(1-\alpha)\%$ para esta estimación es:

$$\hat{p} \pm z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{N}},$$

donde $z_{\frac{0.1}{2}} = 1.64$, $z_{\frac{0.05}{2}} = 1.96$ y $z_{\frac{0.01}{2}} = 2.58$.

Validación cruzada k-fold

Esquema para $k = 5$



Validación cruzada k-fold

El conjunto de datos se divide en k grupos con la misma cantidad de elementos cada uno.

Si la cantidad total de datos N no es un múltiplo de k, digamos $N = kq+r$, las primeras $k-1$ partes tendrán q elementos y la parte final r.

Si el conjunto de prueba en el paso i contiene N_i datos de los cuales C_i fueron correctamente clasificados, la exactitud se estima como:

$$\hat{p} = \frac{\sum_{i=1}^{k} C_i}{\sum_{i=1}^{k} N_i}.$$

Validación cruzada N-fold

Es un caso extremo de la validación cruzada k-fold donde k coincide con el número total de datos.

También se conoce como la validación cruzada dejando uno afuera, o su equivalente en inglés *leave-one-out cross-validation*.

Observar que en este caso los conjuntos de prueba tienen un único elemento y el proceso se repite N veces.

Método de bootstrap

Es una técnica de remuestreo en la cual el conjunto de entrenamiento se elige haciendo un muestreo aleatorio, con reemplazo, del conjunto de datos.
El conjunto de prueba está formado por el resto de los elementos.

Se realiza mediante los siguientes pasos.

1. Elegir B, la cantidad de bootstraps a llevar a cabo.
2. Elegir m, el tamaño de la muestra.
3. Para $i = 1, 2, \ldots, B$:
   i. Elegir aleatoriamente m elementos del conjunto de datos, permitiendo elementos repetidos.
   ii. Ajustar el modelo con los datos de la muestra.
   iii. Estimar la exactitud de predicción en el conjunto de datos restante.
4. Calcular el promedio de las exactitudes de predicción de cada bootstrap.

La exactitud de predicción es una medida global del rendimiento de un clasificador.

En muchas ocasiones, es de interés desglosar el rendimiento del clasificador en cada una de las clases bajo estudio. Es decir, con qué frecuencia los elementos de una clase se clasificaron correctamente o, en caso contrario, con cuál de las restantes clases se confundió el clasificador.

Pregunta

¿Cómo podemos evaluar en rendimiento de un clasificador en cada clase?

Matriz de confusión

Comencemos por el caso de un conjunto de datos con sólo dos clases posibles.
En este caso una de las clases, en general la de principal interés, se trata como la de los casos **positivos**, y la otra clase como la de los casos **negativos**.

Matriz de confusión

		Predicción	
		Positivo	Negativo
Verdad	Positivo	Verdadero positivo (TP)	Falso negativo (FN)
	Negativo	Falso positivo (FP)	Verdadero negativo (TN)

Matriz de confusión

En el caso en que existan k clases $\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_k$, la matriz de confusión es de la forma:

Verdad \ Predicción	$\mathcal{C}_1$	$\mathcal{C}_2$	$\cdots$	$\mathcal{C}_k$
$\mathcal{C}_1$	C_{11}	C_{12}	$\cdots$	C_{1k}
$\mathcal{C}_2$	C_{21}	C_{22}	$\cdots$	C_{2k}
$\vdots$	$\vdots$	$\vdots$	$\ddots$	$\vdots$
$\mathcal{C}_k$	C_{k1}	C_{k2}	$\cdots$	C_{kk}

C_{ij} indica la cantidad de elementos de la clase $\mathcal{C}_i$ que fueron clasificados como pertenecientes a la clase $\mathcal{C}_j$.

En la práctica, suele existir una clase lo suficientemente "importante" para ser considerada como la de los casos positivos, combinando el conjunto de las demás clases como la clase de los casos negativos.

Medidas basadas en la matriz de confusión

Sensibilidad o Exhaustividad

Mide la probabilidad de que el clasificador detecte un caso positivo cuando en verdad lo es. Se calcula como:

$$\frac{TP}{TP + FN}.$$

Especificidad

Mide la probabilidad de que el clasificador detecte un caso negativo cuando en verdad lo es. Se calcula como:

$$\frac{TN}{TN + FP}.$$

Medidas basadas en la matriz de confusión

Exactitud

Mide la probabilidad de que el clasificador acierte. Se calcula como:

$$\frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}.$$

Exactitud balanceada

Suele utilizarse cuando las clases no están balanceadas; es decir, sus tamaños muestrales difieren significativamente. Se calcula como:

$$\frac{\text{Sensibilidad} + \text{Especificidad}}{2}.$$

Medidas basadas en la matriz de confusión

Precisión

Mide la probabilidad de que el clasificador detecte correctamente un caso positivo. Se calcula como:

$$\frac{TP}{TP + FP}.$$

F1 score

Combina las medidas de precisión y exhaustividad para devolver una medida de calidad más general del modelo. Se calcula como:

$$\frac{2TP}{2TP + FP + FN}.$$

Ejemplo*

*Base de datos `Cleveland Heart Disease Data` del paquete `MixAll` de R

Pregunta

Cuando la cantidad de atributos es muy grande, ¿cómo podemos seleccionar las variables importantes para que el modelo sea adecuado?


COMING
SOON