

Clasificación _



Clasificación desde la Econometría

Casos de Uso

- Los problemas de clasificación corresponden a un ejemplo de aprendizaje supervisado donde el vector objetivo responde a un atributo discreto.
- Existen muchos fenómenos cuya primera aproximación es mediante la binarización: ¿**existe o no existe una condición?**
- Esta aproximación toma forma de un ensayo de Bernoulli.

Modelo de Probabilidad Lineal

- **Primera aproximación:** utilizar una regresión lineal asumiendo que nuestra variable dependiente mide la probabilidad de suceso.

$$y = \beta_0 + \beta_1 \times \text{dist100} + \varepsilon_i$$

- La interpretación de los coeficientes se hace en consideración a la probabilidad de ocurrencia del suceso.

Limitantes del Modelo de Probabilidad Lineal

- El modelo LPM presenta fallas en la estimación:
 - Los parámetros estimados pueden tomar valores mayor a uno y menos que cero.
 - Los errores no siguen una distribución normal.
 - La forma funcional lineal restringe las no linealidades en los extremos de la muestra.

Regresión Logística

- La estimación de los coeficientes en la regresión logística se realiza mediante el **método de máxima verosimilitud**.

$$\log\left(\frac{\text{Pr}(y)}{1 - \text{Pr}(y)}\right) = \beta_0 + \beta_1 \times \text{dist}100 + \varepsilon_i$$

Bondad de Ajuste


- La bondad de ajuste en los modelos estimados se evalúa con las métricas de Log-Likelihood.
- Buscamos encontrar un máximo de verosimilitud en una función: **Esto implica un problema de optimización argmin.**
- Existen dos métricas de interés:
 - Log-Likelihood: La verosimilitud del modelo ajustado.
 - LL-Null: La verosimilitud del modelo sin regresores.

Interpretación de Coeficientes

- Importante: **No debemos interpretar los coeficientes como lineales.**
- En la regresión logística los coeficientes estimados corresponden a los logaritmos de las chances de ocurrencia en el cambio en una unidad de x .
- El problema con la interpretación de los coeficientes como log-odds es que no tiene sentido para nosotros.

De log-odds a probabilidad

- El objetivo es traducir de log-odds a una declaración probabilística. Así generamos una explicación intuitiva sobre el efecto de una variable en la probabilidad de ocurrencia.
- Esto lo podemos lograr con la función logística inversa:

$$\text{logit}^{-1}(x) = \frac{\exp(x)}{1 + \exp(-x)}$$


$$\Pr(\text{Cambio de Pozo} = 1|X) = \log\left(\frac{\exp(\beta_0 + \beta_1)}{1 - \exp(\beta_0 + \beta_1)}\right)$$

Efecto Diferencial

- Al convertir una combinación lineal de log-odds estamos obteniendo la probabilidad de un punto específico.
- Para evaluar la contribución de X en la probabilidad de ocurrencia, debemos hacer lo siguiente:
 - Obtener la probabilidad de ocurrencia en escenario 1: $\Pr(y = 1 | dist100 = 100) = \text{logit}^{-1}(\mathbf{x}_i\beta)$
 - Obtener la probabilidad de ocurrencia en escenario 2: $\Pr(y = 1 | dist100 = 200) = \text{logit}^{-1}(\mathbf{x}_i\beta)$
 - Restar ambas probabilidades: $\Pr(y = 1 | dist100 = 200) - \Pr(y = 1 | dist100 = 100)$

Punto equidistante

- Podemos inferir en qué puntaje de X nos encontraremos con el caso equiprobable.
- Esto se conoce como dosis letal media en la literatura bioestadística.

$$x_1 = \frac{-\hat{\beta}_0}{\hat{\beta}_1}$$

Relación entre LPM y Logit

- Podemos tomar los log odds de un modelo logístico y dividirlos por cuatro para obtener un intervalo superior de la contribución de X en y cuando cambia en 1 unidad.

Clasificación desde Machine Learning

Métricas de Desempeño

- No podemos implementar métricas como el Promedio del Error Cuadrático, dado que el método de optimización es distinto.
- Los modelos predictivos de clasificación generan dos tipos de predicciones:
 - **Predicción de probabilidad continua** entre los límites de 0 y 1.
 - **Predicción de clase**, que establece cuál es la más adecuada para una observación.
- Por lo general nos centraremos en la probabilidad de clase para evaluar el desempeño de un modelo de clasificación.

Matriz de Confusión

- ¿Qué es? Cruce de información predicha y etiquetas reales en la muestra de validación.
- Permite observar la cantidad

	Categoría Verdadera	
Predicción	Verdadero	Falso
Positivo	VP: Verdadero positivo	FP: Falso positivo
Negativo	FN: Falso negativo	VN: Verdadero negativo

Accuracy, Precision, Recall

- Exactitud: Casos correctamente predichos del total de observaciones.

$$\text{Exactitud} = \frac{VP + VN}{VP + VN + FP + FN}$$

- Precision: Etiquetas correctas en las positivas.

$$\text{Precision} = \frac{VP}{VP + FP}$$

- Recall: Verdaderos positivos entre los predichos del modelo

$$\text{Recall} = \frac{VP}{VP + FN}$$

Curva ROC

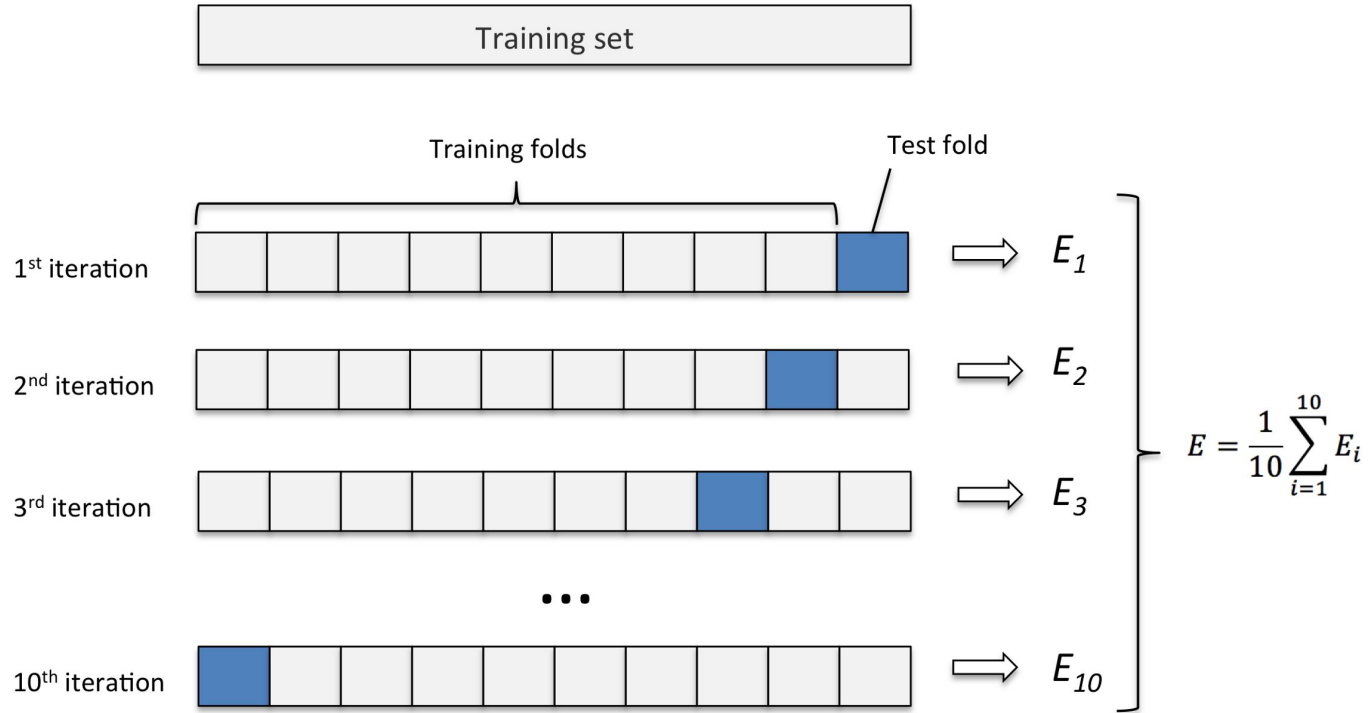
- Permite evaluar el rango de errores del modelo.
- Evalúa la relación entre falsos positivos y verdaderos positivos.
 - En el eje X va la tasa de Falsos Positivos (falsas alarmas).
 - En el eje Y va la tasa de Verdaderos Positivos.

Validación Cruzada

Motivación

- Situación común: no existen suficientes observaciones como para generar un estadístico de prueba robusto.
- Solución: **Iterar de forma sucesiva simulando el entrenamiento del modelo en múltiples muestras.**
- Por cada muestra se estima una métrica de desempeño

K-Fold Cross Validation



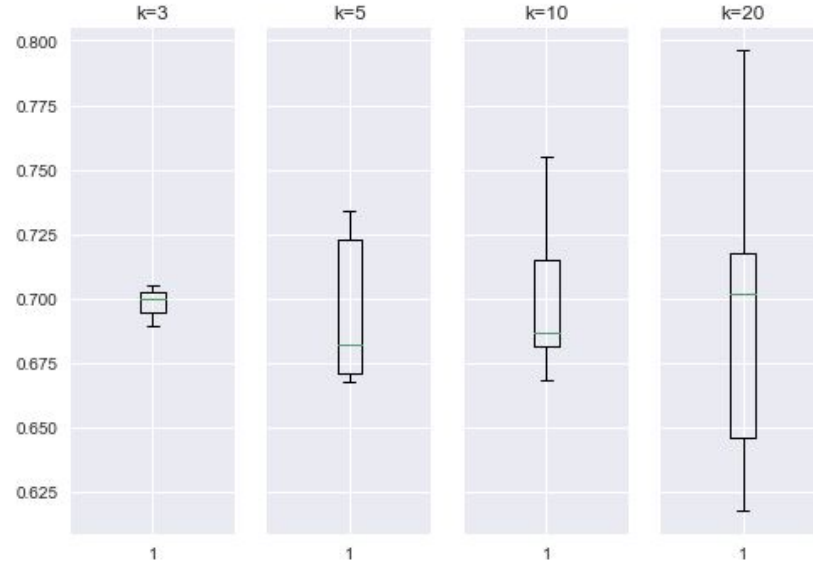
Leave One Out Cross Validation

- Versión extrema de K-Fold Cross Validation.
- Generamos tantos modelos con $(n-1)$ observaciones como observaciones existan en una muestra.

¿Y qué es mejor?

- Ambos métodos representan posiciones extremas. La elección repercute en el trueque Sesgo-Varianza del modelo.

Desempeño del modelo logístico (Puntaje F1) condicional a la cantidad de validaciones cruzadas



{desafío}
latam_

*Academia de
talentos digitales*

www.desafiolatam.com