UNIVERSIDAD DEL PACÍFICO
Departamento Académico de Economía
Machine Learning para Economistas
Segundo semestre de 2022
Profesor: F. Rosales, JP: R. Arauco

PRÁCTICA CALIFICADA 3

Indicaciones Generales:

Esta práctica calificada es sobre penalización. Es para la casa y se debe desarrollar en grupos de 2 ó 3 personas. Para su desarrollo usted debe leer el documento en blackboard `MaxEnt.pdf`.

Preguntas:

El modelamiento de distribuciones geográficas es una aplicación popular de los algoritmos de Machine Learning. En específico, el modelamiento de la distribución de individuos de una especie, o de condiciones propicias para la existencia de la especie, son comunes. En este caso, queremos estimar la probabilidad de encontrar anchoveta en diferentes zonas del mar peruano.

1. Proceso Generador de Datos

   (a) Construya un data frame que contenga la información de una grilla de 100×100 puntos. Considere que la distancia horizontal representa la longitud i, y la vertical, la latitud j.

      i. Genere la variable `temperatura` $z_{i,j}$ de la siguiente manera

$$z_{i,j} = 10 + \cos\left(\frac{4i\pi}{100}\right) + \sin\left(\frac{4j\pi}{100}\right)$$

      ii. Genere la variable `presencia` $y_{i,j}$ de la siguiente manera

$$y_{i,j} = \begin{cases} 1 & \text{si } p_{i,j} \geq 0.8 \\ 0 & \text{si } p_{i,j} < 0.8 \end{cases}$$

      donde

$$p_{i,j} = \frac{e^{10 - z_{i,j} + \epsilon_{i,j}}}{1 + e^{10 - z_{i,j} + \epsilon_{i,j}}}, \quad \epsilon_{i,j} \overset{iid}{\sim} \mathcal{N}(0,1),$$

      que representa la probabilidad de encontrar anchoveta.

      iii. Genere la variable `presencia observable` $\tilde{y}_{i,j}$ de la siguiente manera

$$\tilde{y}_{i,j} = y_{i,j} x_{i,j}, \quad x_{i,j} = \mathbf{1}_{j \geq i}.$$

   Ilustre el proceso generador de datos con gráficos. (2 pts.)

(b) Compare los histogramas de la temperatura para las ubicaciones en las que se observa anchoveta y en las que no. Comente. (2 pts.)

2. Modelamiento con Penalización

(a) Realice un train-test split con un tamaño en test de 30%. Tenga en cuenta que cada punto de la grilla debe tener un valor, sea este cero o uno. (1 pts.)

(b) Seleccione los estimadores $\hat{\lambda}_0$ y $\hat{\lambda}_1$ óptimos en data de entrenamiento y el valor β que maximiza el AUC en data de prueba. Para hacerlo optimice la función de ganancia descrita en el paper de Merow, Smith & Silander (2013) tomando como respuesta la `presencia observable` y como variable predictora, la `temperatura`. Considere que la probabilidad a priori de encontrar peces en cualquier punto de la grilla sigue una distribución uniforme. Hint: puede asumir que los estimadores óptimos están en el vecindario de los parámetros poblacionales. (7 pts.)

(c) Estime el $\text{RoR}_{i,j}$ (relative ocurrence ratio) para cada punto de la grilla. ¿Todas estas probabilidades suman uno? ¿Por qué? (3 pts.)

(d) Calcule la medida de entropía de la distribución posterior r y ajuste las probabilidad de cada punto según una función logística:

$$\tilde{p}_{i,j} = \frac{\tau e^{\hat{\lambda}_0 + \hat{\lambda}_1 z_{i,j} - r}}{1 - \tau + \tau e^{\hat{\lambda}_0 + \hat{\lambda}_1 z_{i,j} - r}},$$

donde $\tau = 0.00001$. (3 pts.)

(e) Calcule la curva ROC y el AUC del modelo a partir de las estimaciones $\text{RoR}_{i,j}$ y $\tilde{p}_{i,j}$. ¿Son diferentes? ¿Por qué? (2 pts.)

Indicaciones Específicas:

- Prepare su código en python como jupyter notebook. Asegúrense de que sea ordenado y fácil de entender. Utilicen comentarios, tablas y gráficos cuando lo consideren necesario.
- Envíe un único archivo comprimido con sus resultados a `francisco@brein.pe`. En el tema del mensaje coloque `MLE-PC3`. En el cuerpo del mensaje indique los integrantes del grupo.
- La fecha límite de entrega es el día domingo 23 de Octubre a las 11:59pm. Cualquier entrega posterior será penalizada.