UNIVERSIDAD DE BUENOS AIRES - 2025

Taller de Programación

TRABAJO PRÁCTICO N° 4

CLASIFICANDO POBRES EN LA EPH: MÉTODOS DE REGULARIZACIÓN Y CART

Fecha de entrega: 28 de noviembre a las 13:00 hs.

Contenido*:* Aplicar los métodos de regularización y árboles de regresión vistos en clase en la EPH para predecir la pobreza en 2025. Discutir las métricas de desempeño de clasificación y evaluar entre modelos la identificación de pobres en una región en Argentina

Modalidad de entrega

- El informe debe subirse a la carpeta correspondiente en repositorio de GitHub del grupo en formato PDF con el nombre **Program_TP4_Grupo#.pdf** (donde **#** es el número de grupo), incluyendo gráficos e imágenes dentro del mismo archivo. El mismo debe tener link al repositorio del grupo en la primer pagina. La extensión máxima es de 8 páginas (sin apéndices) y se espera una redacción clara y precisa.

- Se debe publicar el código con los comandos utilizados en el repositorio, indicando claramente a qué inciso corresponde cada uno. El nombre del archivo deberá ser **Program _TP4_Grupo#**.
  - Al finalizar el trabajo práctico deben hacer un último commit en su repositorio de GitHub llamado "*Entrega final del TP*".
  - El Jupyter Notebook y el correspondiente al TP3 deben estar dentro de esa carpeta.
  - La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
    - No suban el pdf en la sección de "**Actividades/Entregas**" del campus hasta no haber terminado y estar seguros de que han hecho el *commit* y *push* a la versión final que quieren entregar.
    - No hagan nuevos *push* despues de haber entregado su versión final. Esto generaría confusión acerca de que versión es la que quieren que se les corrija.

- **Cualquier detección de copia o plagio será sancionada.**

El objetivo de este trabajo es evaluar la predicción de pobres utilizando técnicas de regularización y de árboles. Van a utilizar las mismas bases `X_train`, `y_train`, `X_test`, `y_test` para cada año que crearon en el TP3.

Usen las mismas variables que usaron en el TP3 para predecir. Si piensan que hay más variables que son relevantes para predecir la pobreza las pueden incluir. **Aclaración**: Por simplicidad, usen únicamente la base de 2025.

A. Modelo de Regresion Logistica con Regularización: Ridge y LASSO

1. *Visualización:* Grafiquen en dos paneles distintos los coeficientes de la penalidad de LASSO y Ridge para la grilla del parámetros de penalidad $\lambda = 10^n$ con $n \in \{-5, -4, -3, \dots, 4, 5\}$ e interpreten la regularización con cada penalidad. (*Hint*: en la función Logistic Regression la opción `C` es la inversa de la fuerza de penalidad, en lugar de alpha. Según el tamaño de su matriz, se recomienda probar con distintas opciones de `solver` para el uso eficiente computacional).

2. *Penalidad óptima por Cross-validation y visualización*: Usando Logistic Regression CV dividiendo la base 5 partes (5-fold), elijan la penalidad óptima λ en la grilla mencionada arriba. ¿Qué λ^{cv} seleccionó en cada caso? Hagan un grafico de linea con puntos conectados mostrando la distribución del error de clasificaición promedio entre folds para cada λ. *Opcional*: para la regularización LASSO, generen otro plot de puntos conectados, para la proporción de variables ignoradas por el modelo en función de λ, es decir la proporción de variables para las cuales el coeficiente asociado es cero.

3. *Estimación con λ^{cv} y comparación de coeficientes*: Estimen una **Regresión Logística** usando *X_train* de respondieron_2025 sin penalidad, con penalidad L1 y L2 siendo la penalidad óptima λ^{cv} elegida en el ítem anterior. Exporten una tabla con las siguientes columnas: (i) los coeficientes estimados para cada variable sin penalidad, (ii) coeficientes de penalidad L1, y (ii) coeficientes de penalidad L2. Interpreten los resultados de la tabla y conteste: ¿cómo son los coeficientes de la regularización con respecto a logit sin penalidad? ¿La penalidad L2 elimina variables de las que incluyeron en su matriz *X_train*?

B. Árboles de Decisión

4. Estimen un árbol de decisión podado (CART) eligiendo el hiperparámetro de costo de complejidad del árbol (`ccp_aplha`) con 10-fold Cross-Validation. En un gráfico de línea, muestre el error de clasificación en función de la grilla de valores que uso para `ccp_aplha`. Comenten los resultados.

5. *Visualización del árbol podado por cross-validation:* En un panel A, muestren el gráfico del árbol que obtuvieron e interpreten. En un panel B, muestren en un gráfico la importancia de cada uno de los predictores.

¿Podría decirse que los coeficientes de las variables menos importantes son los que LASSO "achicó" a cero?

C. Comparación entre métodos

6. Computen la matriz de confusión (p>0,5), la curva ROC y las dos métricas que utilizó en el TP3. Comparen el desempeño de Regresiones Logísticas sin penalidad, KNN con K-CV (ambos en TP3), logit con penalidad LASSO y Ridge, y el árbol de decisión podado. Comenten los resultados. ¿Qué pueden decir del trade-off entre comunicación de los resultados y performance predictiva? En este caso, ¿hay una ventaja por utilizar un método no lineal? Incorporen dentro de la justificación (1 - *accuracy*) para argumentar.
7. Recuerde que el Ministerio de Capital Humano está interesado en identificar a grupos vulnerables para dirigir los recursos de un programa de alimentos. ¿Cambió su respuesta con respecto a cuál es el "mejor" modelo para asignar recursos escasos a los más necesitados?