CC408 - Ciencias de Datos

PRIMAVERA 2024

TRABAJO PRÁCTICO N 4

CLASIFICACIÓN Y REGULARIZACIÓN DE DESOCUPACIÓN USANDO LA EPH

Reglas de Formato y Presentación

Fecha de entrega: domingo 24 de noviembre a las 23.59 hs.

Contenido: Análisis de hogares para los determinantes de la desocupación, problema de clasificación de desempleo entre cohortes usando métodos de regularización y elección de hiperparámetros por cross-validation.

Modalidad de entrega

Al finalizar el trabajo práctico deben hacer un último `commit` en su repositorio de GitHub con el mensaje Entrega final del TP.

- Asegúrense de haber creado una carpeta llamada TP4. Deben entregar un reporte (pdf) y el código (Jupyter notebook). Ambos deben estar dentro de esa carpeta.
- Deberán enviar el link a su repositorio -para que pueda ser clonado y corregido- al siguiente email: ispiousas@udesa.edu.ar
- La última versión en el repositorio es la que será evaluada. Por lo que es importante que:
    - No envíen el mensaje hasta no haber terminado y estar seguros de que han hecho el commit y `push` a la versión final que quieren entregar. Debido a que se pueden tomar hasta 3 días de extensión a lo largo del curso, no se corregirán sus tareas hasta no recibir el correo.
    - No hagan nuevos `push` después de haber entregado su trabajo. Esto generaría confusión acerca de qué versión es la que quieren que se les corrija.

Modalidad de entrega

- El informe debe ser entregado en formato PDF, con los gráficos e imágenes en este mismo archivo. Se espera una buena redacción en la resolución.

- Puede tener una extensión máxima de hasta 10 paginas (no se permite Apéndice).

- Entregar el código con los comandos utilizados, identificando claramente a que inciso corresponde cada comando.

- **Importante**: Todos los miembros del equipo deben haber hecho al menos un commit durante la realización del TP para asegurar que todos hayan aportado a su resolución.

Parte I: Análisis de la base de hogares y tipo de ocupación

Ahora que ya están familiarizados con la Encuesta Permanente de Hogares (EPH) y la desocupación, vamos a complejizar un poco la construcción de las tasas del desempleo. Relacionaremos la información a nivel hogar.

1. Exploren el diseño de registro de la base de **hogar**: *a priori*, ¿qué variables creen pueden ser predictivas de la desocupación y seria útil incluir para perfeccionar el ejercicio del TP3? Mencionen estas variables y justifiquen su elección.

2. Descarguen la base de microdatos de la EPH correspondiente al primer trimestre de **2004** y **2024** en formato .dta y .xls, respectivamente. La base de hogares se llama `Hogar_t104.dta` y `usu_hogar_T124.xls`, respectivamente. Eliminen todas las observaciones que **no** corresponden a los aglomerados de Ciudad Autónoma de Buenos Aires o Gran Buenos Aires y unan ambos trimestres en una sola base. Esto es, a la base de la encuesta individual de cada año (que usaron en el TP3) unan la base de la encuesta de hogar. **Asegúrese de estar usando las variables CODUSU y NRO_Hogar para el merge.**

3. Limpien la base de datos tomando criterios que hagan sentido. Explicar cualquier decisión como el tratamiento de valores faltantes (*missing values*), extremos (*outliers*), o variables categóricas. Justifique sus decisiones.

4. Construya variables (mínimo 3) que no estén en la base pero que sean relevantes para predecir individuos desocupados (por ejemplo, la proporción de personas que trabajan en el hogar).

5. Presenten estadísticas descriptivas de tres variables de la encuesta de hogar que ustedes creen que pueden ser relevantes para predecir la desocupación. Comenten las estadísticas obtenidas.

Parte II: Clasificación y regularización

El objetivo de esta parte del trabajo es nuevamente intentar predecir si una persona está desocupada o no. Esta vez utilizando distintas variables de características individuales y del hogar del encuestado. A su vez, incluiremos ejercicios de regularización y de validación cruzada.

1. Para cada año, partan la base respondieron en una base de prueba y una de entrenamiento (`X_train, y_train, X_test, y_test`) utilizando el comando `train_test_split`. La base de entrenamiento debe comprender el 70% de los datos, y la semilla a utilizar (*random state instance*) debe ser 101. Establezca a desocupado como su variable dependiente en la base de entrenamiento (vector *y*). El resto de las variables serán las variables independientes (matriz *X*). Recuerden agregar la columna de unos (1).
2. Expliquen brevemente cómo elegirían λ por validación cruzada (en Python es alpha). Detallen por qué no usarían el conjunto de prueba (test) para su elección.
3. En validación cruzada, ¿cuáles son las implicancias de usar un *k* muy pequeño o uno muy grande? Cuando $k = n$ (con *n* el número de muestras), ¿cuántas veces se estima el modelo?
4. Para regresión logística, implementen la penalidad, L1 como la de LASSO y L2 como la de Ridge con $\lambda = 1$ (como en la Tutorial 10), usando la opción `penalty` y reporten la matriz de confusión, la curva ROC, los valores de AUC y de Accuracy para cada año.[1] ¿Cómo cambiaron los resultados con respecto al TP3? ¿La performance de regresión logística con regularización es mejor o peor?
5. Realicen un barrido en $\lambda = 10^n$ con $n \in \{-5, -4, -3 \ldots, +4, +5\}$ y utilicen 10-fold CV para elegir el λ óptimo en regresión logística con Ridge y con LASSO. ¿Qué λ seleccionó en cada caso? Usando la librería de seaborn, generen box plot mostrando la distribución del error de predicción para cada λ. Cada box debe corresponder a un valor de λ y contener como observaciones el error medio de validación (MSE) para cada partición. Además, para la regularización LASSO, generen un line plot del promedio de la proporción de variables ignoradas por el modelo en función de λ (como vieron en el tutorial 10), es decir la proporción de variables para las cuales el coeficiente asociado es cero.[2]
6. En el caso del valor óptimo de λ para LASSO encontrado en el inciso anterior, ¿qué variables fueron descartadas? ¿Son las que hubieran esperado? ¿Tiene relación con lo que respondieron en el inciso 1 de la Parte I?

---

[1] En la clase magistral 9, vimos el método de regularización en regresión lineal donde la variable dependiente es numérica. En este caso, nuestra variable dependiente es binaria (ocupado, desocupado), por lo que usamos la regresión logística y aprovechamos la opción de penalidad para aplicar los métodos de regularización vistos en clase.

[2] *Hint*: a mayor penalidad, esperamos que más coeficientes sean 0, por lo tanto, esta figura debe tener una forma de "S".

7. Elijan alguno de los modelos de regresión logística donde hayan probado distintos parámetros de regularización y comenten: Compare los resultados de 2004 versus 2024, ¿qué método de regularización funcionó mejor: Ridge o LASSO? ¿LASSO hizo una selección distinta de predictores en 2004 versus 2024? Comenten mencionando el error cuadrático medio (MSE).