Actividad 1. Evaluación del método de aprendizaje

**Objetivos**

El desarrollo de esta actividad tiene el objetivo de evaluar correctamente un modelo de aprendizaje automático en función de si es un problema de clasificación o regresión.

**Pautas de elaboración**

La actividad consiste en la ejecución de dos modelos de aprendizaje proporcionados por el profesor de la materia bajo el *software* informático Matlab (opcional otro tipo de *software*) para que el alumno pueda ejecutar dicho *script* y analizar los resultados atendiendo a los siguientes criterios:

* Selección de las métricas correctas.
* Evaluación de si existe sobreajuste o infraajuste en alguno de los modelos.
* Conclusiones sobre la precisión de los modelos.
* Planteamiento de posibles líneas futuras de trabajo.

**Extensión y formato**

El estudiante debe entregar una memoria en PDF de no más de diez páginas detallando los puntos anteriores.

**Ejemplo**

* Considerando los datos acerca de la liga de futbol americano y los resultados proporcionados en la base fotbal.sav, en la cual cada observación corresponde a un equipo distinto y el significado de cada variable para *y*, *x*1,...,*x*9 es Games won, Rushing yards (season), Passing yards (season), Punting average (Yds/Punt), Field goal percentage (Fgs made/Fgs attempted-season), Turnover differential, Penalty yards (season), Percent rushing (rushing plays/total plays), Opponents’ rushing yard (season), Opponents’ passing yards (season):
  + Ajuste un modelo adecuado (regresión o de clasificación) relacionando el número de juegos ganados con las variables *x*2, *x*7 y *x*8. Interprete el ajuste.
  + Calcule una métrica adecuada para evaluar el ajuste e interprétela.
  + Construya las gráficas de los residuos y discuta lo bueno y lo malo del modelo.
  + Prediga el número medio de juegos cuando el equipo tiene *x*2 = 2300, *x*7 = 56 y *x*8 = 2100.
  + Seleccione las variables que mejor describan a la respuesta de acuerdo con algún criterio de su elección (el p-valor en pruebas t para ver si una variable es significativa, un método iterativo, el ajuste de algunos modelos con menos variables, su conocimiento sobre futbol americano, etc.). Obtenga los resultados básicos de este modelo y compárelos con la métrica del modelo original.
* Los datos en la base ej5tarea5.sav corresponden a 700 clientes a los que se les negó o no se les negó un crédito (en la base corresponde a la variable *default* con valor 1 si se negó el crédito y 0 si no se negó). Simultáneamente, a cada cliente se le midieron distintas variables socioeconómicas, algunas de las cuales influyeron en esa decisión. La finalidad de un nuevo análisis es determinar una regla de decisión que permita establecer para nuevos clientes si se les otorga o no el crédito. Si no puede recuperar las etiquetas de las variables, estas son Age in years, Level of education (1=Did not complete highschool, 2 = High school, 3 = Some college, 4 = College, 5=Post undergraduate degree), Years with current employer, Years at current address, Household income in thousands, Debt to income ratio (x100), Credit card debt in thousands, Other debt in thousands, Previously defaulted.
  + Elegir y ajustar un modelo adecuado (regresión o clasificación) para predecir el *output.*
  + Partiendo de todas las variables, elegir aquellas que mejor sirven para determinar el modelo, por ejemplo, usar un método de selección, el p-valor de pruebas asociadas a una variable, su intuición sobre el tema, etc.
  + Para el modelo seleccionado determine qué tan bien se clasifican los clientes obteniendo métricas adecuadas. Primero considerando que el enfoque del análisis es solo explicativo (i.e. sobre la muestra completa), por lo que estas solo se usan como una medida más del buen o mal ajuste, y luego desde un enfoque predictivo en el cual se quiere evaluar el funcionamiento del modelo en nuevos datos.
  + Suponga se tiene un cliente nuevo al que se le va a determinar si se le aprueba o no el crédito. Este cliente tiene 37 años, no terminó la preparatoria, tiene 20 años en el mismo empleo, 13 en la misma dirección, un ingreso (en miles) de 41, una razón de deuda e ingreso de 12.9, una deuda en su tarjeta de crédito (en miles) de 0.9 y en otras deudas de 4.49 ¿Le asignaría o no el crédito?
* **Observaciones:**
  + Para leer un archivo .sav en Python necesita instalar pyreadstat y usar la instrucción pd.read spss.
  + En el caso de clasificación puede usar cualquiera de los modelos discutidos hasta el momento o incluso algún otro que se ve en el curso, como regresión logística, k vecinos más cercanos, máquinas de soporte vectorial, etc. Lla técnica es por el momento lo de menos.
  + Si hay variables cualitativas incluidas como *inputs (factors),* recuerde generar variables *dummies* (variables 1 y 0 para todas las categorías, excepto una que se toma de referencia).
  + Puede usar cualquier software de su elección. Los normalmente usados en el área son R y Python, que solo presentan tanto salidas como códigos.

**Rúbrica**

|  |  |  |  |
| --- | --- | --- | --- |
| Evaluación del método de aprendizaje | Descripción | Puntuación máxima  (puntos) | Peso  % |
| Criterio 1 | Selección de métricas | 2 | 20% |
| Criterio 2 | Evaluación del sobreajuste/infraajuste | 2.5 | 25% |
| Criterio 3 | Conclusiones | 2.5 | 25% |
| Criterio 4 | Líneas futuras | 2 | 20% |
| Criterio 5 | Presentación | 1 | 10% |
|  |  | **10** | **100 %** |