Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Máquinas de aprendizaje

Reporte: Ejemplo de regresión logística

Docente: Abraham Sánchez López

Alumno	Matrícula
Taisen Romero Bañuelos	202055209

Regresión logística contra regresión lineal

Para empezar, recordemos que la regresión lineal busca la relación entre una variable dependiente y múltiples variables independientes, y su salida es continua, lo que implica que no es ideal para problemas de clasificación. Entonces, podemos inferir que un modelo lineal no es del todo apto para predecir variables binarias (churn). Podría dar una idea de cómo es la correlación entre variables pero no por ello es un modelo apropiado para la tarea. De todas formas, veamos qué sucede si lo hacemos con un modelo lineal.

Algo que hay que aclarar es que para poder hacer una buena comparación entre ambos modelos trabajé ligeramente sobre la salida del modelo lineal, básicamente sólo convertí las predicciones a probabilidades. Para las predicciones negativas las trunqué convirtiéndolas a cero, y para las que son mayores a uno las convertí a uno.

```
> # ==== Comparacción lineal vs logistica ====
> head(churn_test$churn_prob)       #logistica
[1] 0.1650425 0.1046294 0.1710825 0.2438470 0.1003698
[6] 0.2852181
> head(churn_test$churn_prob_lin)  #lineal
[1] 0.1836339 0.1090151 0.1876202 0.2347459 0.1056793
[6] 0.2579945
>
```

En la captura podemos observar que los primeros seis valores son similares, con la diferencia de que el modelo lineal suaviza un poco más las predicciones. Es decir, la regresión logística tiende a producir valores más extremos (lo que tiene sentido porque quiere categorizar los resultados).

```
> summary(churn_test$churn_prob)      #logistica
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.02922 0.09349 0.13489 0.14767 0.18452 0.41604
> summary(churn_test$churn_prob_lin)  #lineal
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
0.00000 0.09781 0.14566 0.14784 0.19567 0.33944
>
```

Aquí podemos corroborar que la regresión logística produce valores más extremos, pues, tiene un máximo mayor al del modelo lineal (0.4160) vs (0.3394). Sobre el mínimo, el modelo lineal pareciera ser más extremo, pero esto se debe a que hice un truncamiento de las predicciones (o sea, que lo forcé a que sea cero). La mediana parece no ser muy diferente entre ambos modelos, lo que significa que tienen una distribución similar (aunque obvio, el modelo logístico iba a estar más sesgado. Esto se ve porque produce valores más alejados de la media).

El diablo se esconde en los detalles, y eso es algo que no podemos omitir en este caso, pues, pese a que los modelos se comportaron de manera similar, el modelo logístico producía valores más realistas para una categorización (más extremos, es esto o no lo es), y en consecuencia genera una mayor separación entre los clientes con alta y baja probabilidad de abandonar.