# **Guia 1: Regresión logística**

## **Regularización y colinealidades**

En esta primera parte vamos a tratar de convencernos de dos conceptos importantes a la hora de interpretar los resultados de un modelo de regresión logística.

El primero, es que sus coeficientes NO son independientes.

El segundo concepto que vamos a tratar de fijar es que, las regularizaciones L1 y L2 tienen formas cualitativamente distintas de reducir la complejidad de la frontera de decisión. En particular, que tratan de manera muy distinta las colinealidades presentes en nuestro sistema (i.e. los features muy correlacionados). Quiero que ganemos intuición de qué hace cada método de regularización con las variables altamente correlacionadas.

### **Ejercicio. 1.**

Supongamos que queremos publicar nuestro modelo. Queremos antes entender cómo funciona, es decir, queremos entender la importancia que les da a las distintas variables. Ahora queremos en particular, cuantificar la importancia que tiene la ‘conservación genómica’ en nuestro modelo. Se nos ocurre pues, medir la caída del desempeño al eliminar las variables de conservación. Miramos la figura de Feature Importance que hicimos en el notebook y nos damos cuenta que las únicas variables de conservación que sobrevivieron (coeficiente distinto de cero) son **phylop\_100\_vert** y  **phastcons100\_vert.** De hecho, una de ellas está como la variable de mayor importancia del modelo.

1. Elimine los feature **phylop\_100\_vert** y  **phastcons100\_vert** del dataset y vuelva a correr todo el modelito.

Ayuda para remover las variables reemplace la línea del notebook donde define X e y por: X,y=data.drop([targetname,'phylop\_phylop100\_vert','phastcons\_phastcons100\_vert'],axis = 1), data[[targetname]]

Respondamos:

* 1. ¿Cuánto cayó la performance del modelo? Discuta.
  2. Grafique el nuevo feature importance del modelo. Cuál es ahora la variable más importante? De qué tipo es? (ayúdese buscándola en la matriz de correlaciones del wb0\_exploratory.ipynb) Qué coeficiente tenía en el modelo anterior nuestra variable ‘más importante’ en el nuevo modelo?
  3. La variable que estaba en segundo lugar en el modelo original, ‘subió’ al puesto número uno? Por qué?
  4. ¿Qué pasa si sacamos la nueva variable de mayor importancia y volvemos a correr? ¿Qué variable (y de qué tipo) aparece en primer lugar? Mire su nivel de importancia en su modelo previo.

1. En base a estos resultados discuta: en un modelo de regresión logística con regularización, es correcto interpretar la importancia de una variable en forma independiente?
2. Discuta en grupo: ¿Cómo le parece que ‘opera’ L1 con las colinealidades del sistema? Dicho de otro modo, cómo ‘distribuye’ los pesos de los coeficientes entre variables muy correlacionadas?
3. Evidentemente, el método de eliminar las dos variables de conservación con coeficiente distinto de cero no parece servir para cuantificar la importancia de la conservación genómica en nuestro modelo. Elimine todas las variables de conservación simultáneamente y mida la caída en desempeño del modelo. (Hint: revise el wb0\_exploring.ipynb para no olvidarse ninguna variable de conservación).
   1. Le parece que puede ser importante hacer un análisis preliminar de clustering sobre las variables de nuestro modelo?.
4. Ejercicio teórico **opcional para el hogar:** considere dos variables perfectamente correlacionadas, vea cuál es la solución óptima de coeficientes que obtiene con regularización L1 (a suma por simplicidad el hyperparámetro de regularización C = 1)

### **Ejercicio. 2.**

Veamos ahora cómo trabaja L2. Cómo penaliza los coeficientes en general, y en particular ante la presencia de colinealidades.

1. Vuelva a considerar todas las variables en la matriz de datos X. Corra un modelo con regularización L2 ('clasificador\_\_penalty': ['l2']) en la grilla de hyperparámetros)
   1. Cuántos coeficientes tira a cero? Puede verlo en la curva de ‘survival\_features’ vs ‘C’.
   2. Tome el valor de C óptimo y haga el gráfico de Feature Imporance. Refuerce lo concluido en el ítem anterior: cuántas variables ‘sobrevivieron’? Dónde están las variables de conservación genómica.
   3. Discuta en su grupo el efecto que tiene la regularización de tipo L2 sobre colinealidades, es decir, cómo distribuye sus pesos.
   4. Ejercicio teórico **opcional para el hogar:** considere dos variables perfectamente correlacionadas, vea cuál es la solución óptima de coeficientes que obtiene con regularización L2. Asuma C =1 por simplicidad
2. Comparación directa L1 vs L2:
   1. Compare las curvas de optimización del hiperparámetros C para las regularizaciones L1 y L2. ¿Qué efecto nota en valores de penalización muy elevados (osea, C pequeño)?
   2. Ejercicio práctico **opcional para el hogar.** Compare los histogramas de coeficientes en una regularización de tipo L2 para penalizaciones fuertes (C<<1, C=óptimo, y C>>1)

### **Ejercicio. 3**

1. Vuelva a la primera línea del notebook, donde cargamos los datos. Reemplace

data=pd.read\_csv(datapath,sep = ',',index\_col='ChrPosRefAlt',nrows=500) por

data=pd.read\_csv(datapath,sep = ',',index\_col='ChrPosRefAlt')

Puede interpretarlo como un proceso usual de la vida real, pasó el tiempo y su campo de research evolucionó y la cantidad de datos disponibles ahora aumentó.

Más aún, procure que el set de testing sólo tenga un 20% :

X\_train, X\_test, y\_train, y\_test = train\_test\_split(X, y, test\_size=**0.20**, random\_state=SEED,stratify = y)

* 1. Vea la nueva dimensión del dataset (agregue una línea de código así:

display(data.shape)

* 1. Corra el modelo pero con penalización L1. Sólo grafique la curva de performance en fucnón de C. Sería ideal que tenga a mano la curva de performance en función de C que hicimos inicialmente en clase. Compárelas. Vea qué cambia en valores de C ‘grandes’ (baja regularización).
  2. Se aplana? (si corre con L2 no le van a quedar dudas). Por qué?

Lo que pasó: se aplanó la curva? Si seguimos aumentando el número de datos (muestras)? Créame, se aplanará más. Entonces valores de C extremadamente altos pueden ser solución? Si se aplana, parecería que sí. Pero entonces, eso tiende a la solución sin regularización? Sí.

Y el motivo es que con suficientes datos, un modelo de regresión logística (recuerde que es un modelo lineal al final del día) no puede aportar una superficie de separación lo suficientemente compleja como para llegar a sobreajustar el training set. Eso pasa en casos como los que vimos en clase, donde la relación entre cantidad de variables y de ítems en su dataset está más comprometida. Ni hablar si tenemos más columnas (variables) que muestras en el dataset.

e) Quédese con L1. vuelva a una proporción de testing del 50% y corra todo el modelo hasta el final. Me interesa que nos guardemos el valor de performance del modelo.

Luego de haber corrido el notebook agregue la siguientes líneas:

res = y\_test

res['y\_pred'] = ypred\_test

res.to\_csv('./models/LR\_whole\_data\_testing\_set.tsv',sep = '\t')