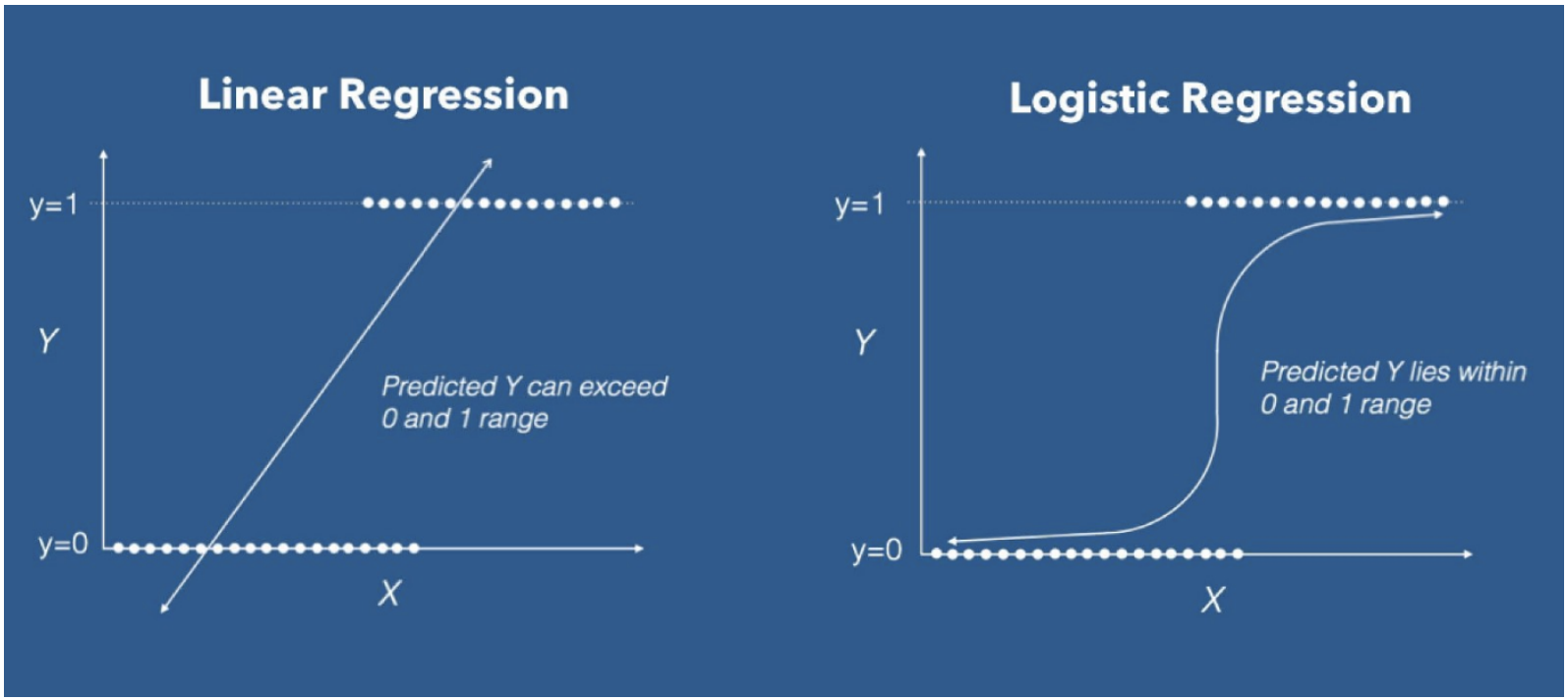
Clase 6: Clasificadores Basados en Regresión

Simón Leiva Meza

Pontificia Universidad Católica de Chile
Facultad de Matemáticas
Programa de Magister en Estadística

Intuición

Uno de los modelos estadísticos más clásicos y utilizados es la regresión lineal, que normalmente lo utilizamos para explicar una variable en función de otras covariables, sin embargo puede ser utilizado para clasificar, estableciéndola como linea frontera entre las dos clases



Modelo de Regresión Logística

Definición

Se suele utilizar como mecanismo para estimar la probabilidad de ocurrencia (o no ocurrencia) de algún evento.

Regresión Logística

Un modelo de regresión logística, es un modelo lineal generalizado con respuesta binomial, y función de enlace logística. El caso binario, sea $Y = 1$ el evento exitoso, un modelo de regresión logistica es

$$\log\left(\frac{P(Y=1|X=x_i)}{1-P(Y=1|X=x_i)}\right) = \beta^T x_i$$

Regresión Logística
Estimación

La forma habitual de estimar un modelo de regresión logística es por máxima verosimilitud. Para el cual no existe una expresión cerrada como el caso de la regresión lineal, sin embargo si se puede obtener la estimación mediante métodos numéricos.

Regresión Logística
Estimación

El problema de estimación finalmente se traduce en un problema de optimización. Para optimizar, se puede utilizar el método de Newton-Raphson (el cual no es el único legal).

Método de Newton-Raphson

Sea f una función a optimizar, y sea β un vector de parámetros. El metodo de Newton-Raphson, actualiza los parámetros β de acuerdo a la siguiente iteración

$$\beta_{n+1} = \beta_n + H^{-1}(\beta_n)\nabla f(\beta_n)$$

Donde H^{-1} corresponde a la inversa del Hessiano de f y ∇ al gradiente de f

Regresión Logística
Estimación

Matemáticamente, el algoritmo es de rápida convergencia ($O(x^2)$), sin embargo suele ser lento computacionalmente. Si bien el cálculo del Hessiano y el gradiente son rápidos, es el cálculo de de la inversa del Hessiano la que genera los cuellos de botella en la computación.

Alternativas a estos métodos son los llamados Quasi - Newton, básicamente es el cambio en la forma de calcular la matriz H. El ejemplo más común de estos métodos en este contexto es el de Fisher - Scoring

Regresión Logística
Enfoque de interpretación

Hasta el momento hemos sido muy puros teóricamente hablando, más adelante veremos que los parámetros puedes ser obtenidos sin usar máxima verosimilitud. Sin embargo hay que tener en cuenta que usar otro estimador, se pierden las buenas propiedades de los EMV.

Al usar un enfoque interpretativo, se debe velar por conservar la pureza de:

- Variables.
- Función objetivo.
- Método de estimación.

Regresión Logística
Enfoque de interpretación

Lo fundamental es que el modelo nos permita aprender de un fenómeno, una buena interpretación de parámetros y un análisis de la significancia estadística juega un rol protagónico.

En Python la librería **statsmodels** contiene una gran variedad de modelos estadísticos, entre ellos una familia de modelos lineales generalizados. Vamos a utilizar esta librería para ajustar un modelo simple de predicción que sea entendible e interpretable.

Regresión Logística
Enfoque de interpretación

Los parámetros tiene la siguiente interpretación

$$\log\left(\frac{P(Y=1|X=x_i)}{1-P(Y=1|X=x_i)}\right) = \beta_0 + \beta_1 x_i$$

$$\frac{P(Y=1|X=x_i)}{1-P(Y=1|X=x_i)} = e^{\beta_0} \cdot e^{\beta_1 x_i}$$

Notar que bajo esta especificación, el intercepto siempre será nuestra categoría de referencia, es decir si la variable x_i está ausente, el efecto siempre es e^{β_0}, es decir

$$e^{\beta_0} = \frac{P(Y=1|X=x_0)}{1-P(Y=1|X=x_0)}$$

Regresión Logística
Enfoque de interpretación

Finalmente

$$\frac{\frac{P(Y=1|X=x_i)}{1-P(Y=1|X=x_i)}}{\frac{P(Y=1|X=x_0)}{1-P(Y=1|X=x_0)}} = e^{\beta_1 x_i}$$

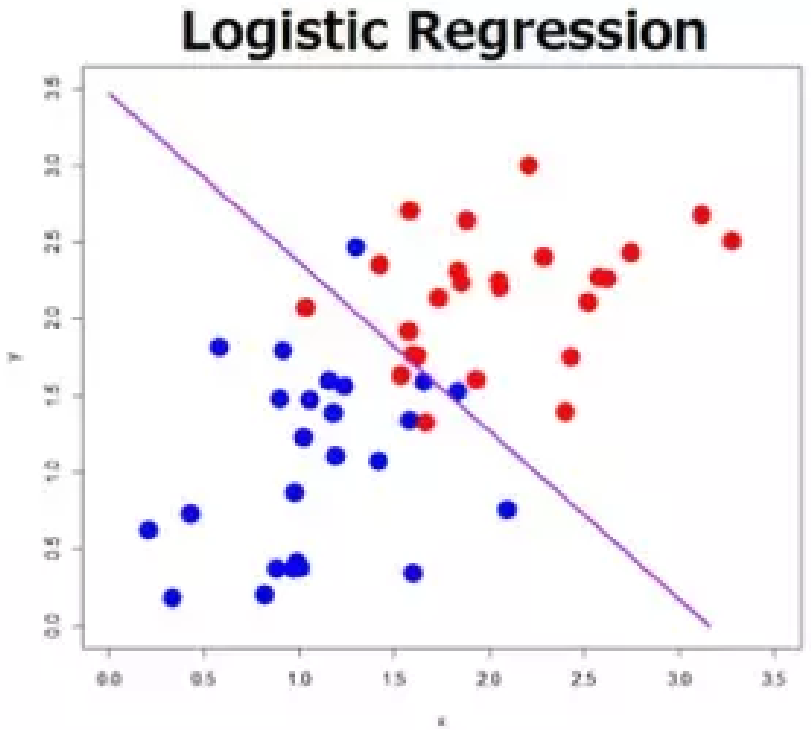
En el caso de regresión múltiple, la interpretación es la misma, pero dejando las otras variables constantes.

Regresión Logística
Enfoque de interpretación

Ejemplo: Usando la tabla glass.csv, vamos a ajustar un modelo de regresión logística para predecir los vidrios tipo 1.

Regresión Logística
Enfoque de predicción

A diferencia del enfoque anterior, nos interesa buscar el hiper plano que mejor separe los datos pertenecientes a una categoría con respecto a otra. El gráfico muestra una pequeña comparativa de un clasificador basado en árboles y otro en regresión.



Regresión Logística
Enfoque de predicción

En general las principales diferencias son:

1. Foco en el resultado, la significancia de las variables es importante pero principalmente por un tema de dimensionalidad.
2. Las variables tiene total libertad de sufrir transformaciones, aunque estas no sean interpretables.
3. Método de estimación libre, por ejemplo se puede minimizar la entropía usando gradiente decendente.
4. Mucha mayor propensión a sobreajuste, es recomendable restringir el espacio de los parámetros.

Una regresión logística de esta naturaleza se vuelve un modelo caja gris, si bien no es del todo interpretable, hay total control de lo que hace el modelo.

Selección de Variables

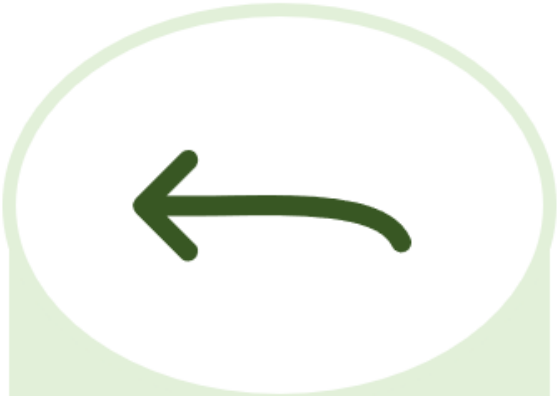
Cuando necesitamos seleccionar variables o bien reducir la dimensionalidad del problema algunos de los métodos más populares son:

- **Selección secuencial:** Ajustar modelos e ir incrementando el número de variables o reduciendo de acuerdo al aporte o pérdida marginal de las mismas.
- **Indicadores de importancia:** Basarse en IV, (y por consiguiente los WOE) o bien IG para decidir que variables utilizar.
- **Shrinkage:** Restringir el espacio de parámetros de tal modo de minimizar las variables poco relevantes.

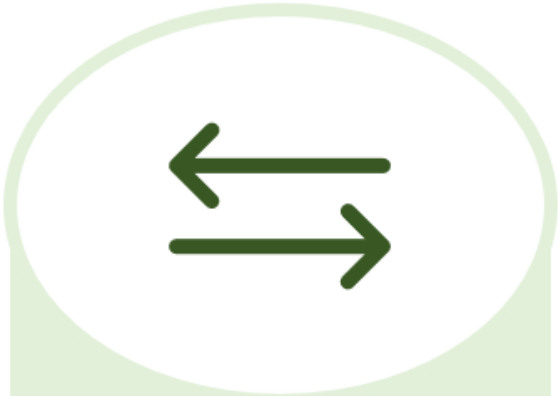
Selección de Variables

Selección secuencial

Sequential Forward Selection: Se inicia el modelo con la variable con mayor incremento en AIC (u otra métrica) y se van incorporando variables secuencialmente hasta que no incremente el AIC.



Sequential Backward Selection: Se inicia el modelo con todas las variables incorporadas, se van eliminando variables de forma secuencial hasta que ya no siga aumentando el AIC (u otra métrica).



Sequential Stepwise Selection: Se inicia el modelo sin variables y estas se van incorporando secuencialmente, se diferencia con el forward, en que cada paso también puede eliminar variables.

Selección de Variables

Indicadores de importancia

IV

Information Value: Indica que tan bien una variable es capaz de discriminar entre las clases de la variable respuesta, se calcula como

$$IV = \sum(\%ceros_i - \%unos_i) \cdot WOE_i$$

En general un IV > 0.2 es una buena variable

IG

Information Gain: Al igual como se utiliza en los árboles de decisión, indica que tanto desorden se reduce en la variable respuesta al usar la variable como discriminadora.

WOE

Weight of Evidence: El poder predictivo de una separación de variables. Principalmente es utilizado para categorizar variables continuas y agrupar grupos parecidos.

$$WOE_i = \log(\frac{\%\ ceros_i}{\%\ unos_i})$$

Selección de Variables

Indicadores de importancia

Llueve	Temperatura	Cielo
1	12	Nublado
1	9	Nublado
1	16	Nublado
1	5	Parcial
0	14	Soleado
0	13	Nublado
0	24	Parcial
0	31	Soleado
0	13	Parcial

Usando el mismo ejemplo calculamos los IV para las mismas reglas

Regla 1: llueve si la temperatura es menor a 15
Regla 2: de acuerdo a como se ve el cielo decidimos.

Para la regla 1 se obtiene que

Regla	% Llueve	% No llueve	WOE
T > 15	0,33	0,67	0,31
T < 15	0,5	0,5	0

Luego su IV es

$$IV_T = (0.67 - 0.33) \cdot 0.31 + 0 = 0.10$$

Para la otra regla

Regla	% Llueve	% No llueve	WOE
Nublado	0,75	0,25	-0,48
Parcial	0,33	0,67	0,31
Soleado	0,75	0,25	-0,48

$$IV_C = (0.25 - 0.75) \cdot -0.48 + (0.67 - 0.33) \cdot 0.31 + (0.75 - 0.25) \cdot -0.48 = 0.82$$

La conclusión es la misma que con el IG, la variable Cielo es mejor

Selección de Variables
Shrinkage

Consiste en agregarle una restricción a la función a optimizar, esta restricción en particular es:

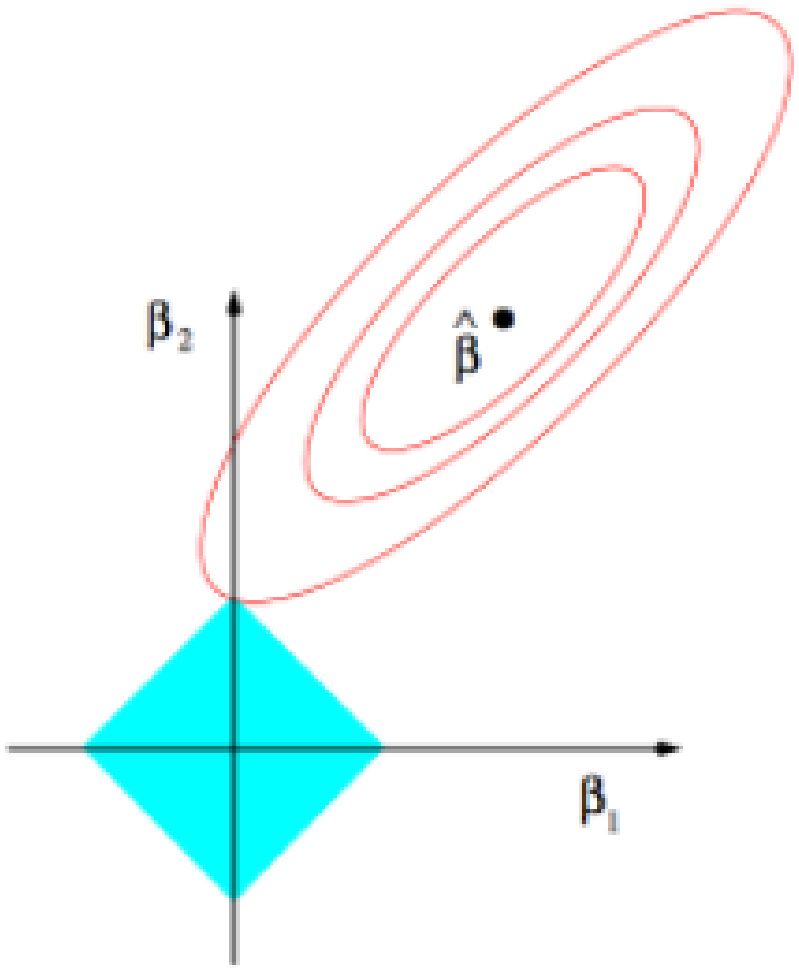
$$\sum_{j=1}^{p} |\beta_j| \leq t$$

Aplicando lagrangiano, el problema se traduce en optimizar

$$\text{argmax}_\beta \left(\sum_{i=1}^{n} \left[y_i(\beta^T x_i) - \log(1 + e^{\beta^T x_i}) \right] - \lambda \sum_{j=1}^{p} |\beta_j| \right)$$

El parámetro λ es un parámetro de penalización de complejidad. Este tipo de penalización se llama del tipo L_1

Selección de Variables
Shrinkage

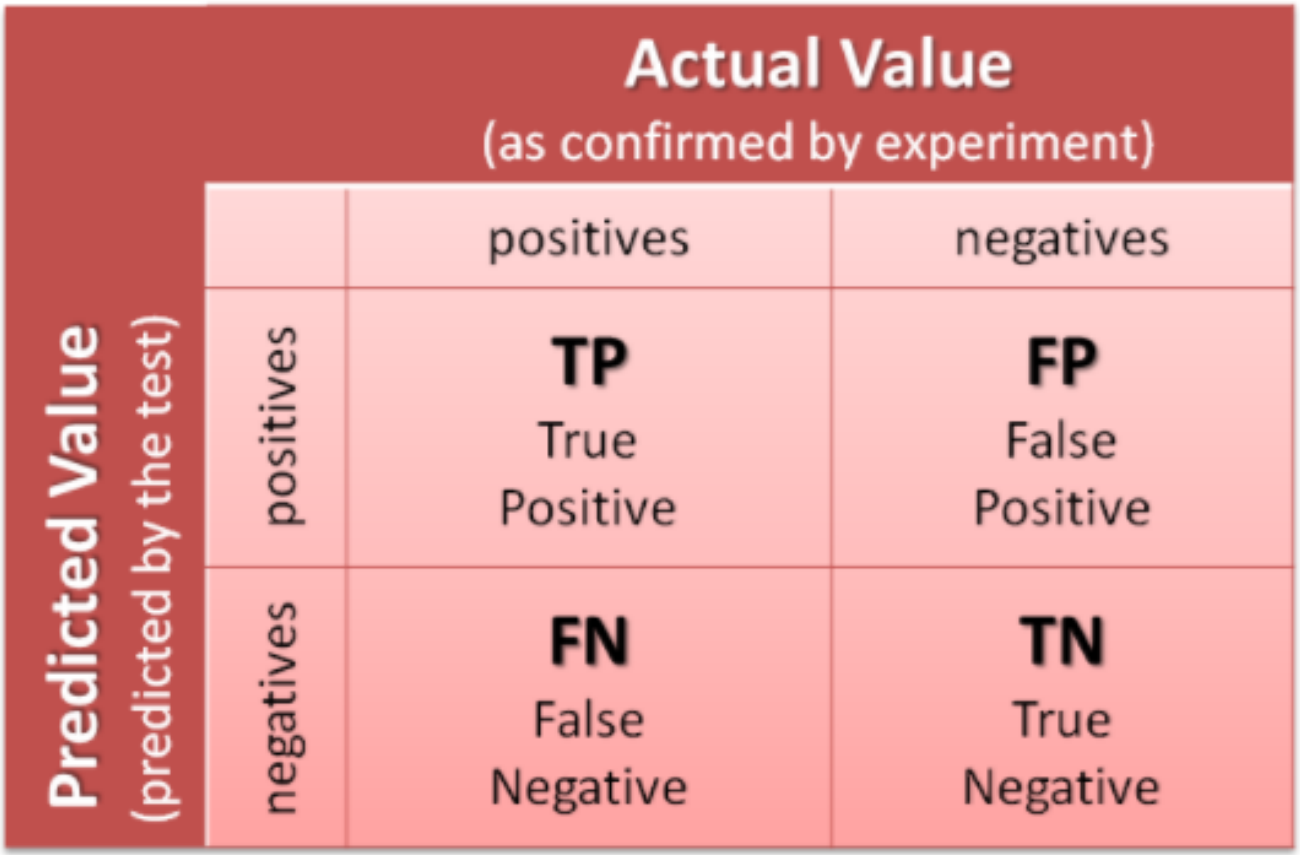


Selección de Variables

	Ventajas	Desventajas
Selección secuencial	Simple, intuitiva, ayuda a reducir dimensionalidad, automático.	Lento para algunos casos, requiere muchos ajustes de modelos. Puede variar según el indicador de ajuste.
Métricas de evaluación	Muy simples, de rápido cálculo, interpretables, ayudan a la creación de variables.	Sólo se ve la importancia de a pares de las variables, no en conjunto. Se asume que las variables son categorías.
Shrinkage	Evaluación conjunta de las variables, rápido.	No indica el peso de las variables, no es interpretable.

Evaluación
Matriz de confusión

La clase pasada vimos



Evaluación
Matriz de confusión

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + FN + TN}$$

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1} = \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

- El accuracy nos dice que proporción el modelo predijo correctamente.
- Precision indica que proporción de los casos predichos son están correctamente clasificados.
- Recall indica que proporción de los casos de interés fueron correctamente clasificados.

Evaluación
Matriz de confusión

Ejercicio: Un modelo obtuvo los siguientes resultados:

Precision	0.34
Recall	0.55
Accuracy	0.93

¿Qué puede concluir del modelo?

Evaluación
Curva de ROC

Se construye a partir de la matriz de confusión, la curva contiene en el eje X la razón de falsos positivos ($FP/(FP + TN)$) y en el y la razón de verdaderos positivos ($TP/(TP + FN)$). Como las predicciones son probabilidades, la curva se construye fijando umbrales de clasificación y calculando las métricas.

El área bajo la curva (AUC) nos da un indicador de la capacidad de discriminación del modelo, en general un AUC entre 0.5 y 0.6 indica un clasificador pobre, 0.6 y 0.7 regular, 0.7 y 0.8 uno bueno, entre 0.8 y 0.9 uno muy bueno y entre 0.9 y 1 es sospechosamente bueno.

Evaluación
Curva de ROC