Métodos de Clasificación Ordinal Basados en Regresión y su Aplicación

Evangelina Garza Elizondo
Centro de Investigación en Matemáticas. Unidad Monterrey
Email: evangelina.garza@cimat.mx

***Resumen*—En el presente trabajo se exponen modelos de regresión que fueron desarrollados específicamente para trabajar con variables ordinales. A partir de ellos se pueden acoplar tareas de clasificación que respeten el orden estricto de este tipo de variables y que en muchos métodos estándar de *machine learning* no se toman en cuenta. Se muestran algunas aplicaciones y sus resultados en comparación a métodos de clasificación estándar. No se encontraron ventajas significativas al utilizar este enfoque, lo cual puede explicar la poca popularidad de este tipo de métodos.**

I. Introducción

En el mundo de la Estadística, el tipo de variables con que se trabajan las tareas de modelización, clasificación, regresión, entre otras, varían ampliamente del área en que se estén aplicando estas tareas. Las ciencias duras por ejemplo, utilizan en su mayoría variables cuantitativas y continuas. Mientras que en las ciencias sociales es común utilizar variables cualitativas. Un tipo de variable que es de interés y que ha sido utilizado en muchas áreas son las variables ordinales, las cuales suponen un orden determinado y no ambiguo entre sus categorías.

A pesar del amplio uso de estas variables en distintas disciplinas, la mayoría de los algortimos de clasificación tratan a las variables asociadas con la clases como valores no ordenados o cantidades nominales. A continuación se exploran modelos de regresión desarrollados para variables ordinales, a partir de los cuáles es posible desarrollar métodos y algoritmos de clasificación.

II. Modelo de Probabilidades Proporcionales

Este modelo supone que cada una de las *k* categorías ordenadas de la variable respuesta tiene probabilidades: $\pi_1(x), \pi_2(x), ..., \pi_k(x)$ cuando las covariables (variables explicativas) tienen el valor de x. Sea Y la variable respuesta que puede tomar valores desde $1, ..., k$, con las probabilidades anteriores, y sea $\kappa_j(x)$ las probabilidades de que $Y \leq j$ es decir $P(Y \leq j|x)$ entonces el modelo de probabilidades proporcionales está dado por:

$$\kappa_j(x) = \kappa_j exp(-\beta^T x)$$

Donde β es un vector de parámetros desconocido. Tomando el radio de las probabilidades correspondientes entre dos covariables $x_2 - x_1$ y definiendo:

$$\gamma_j(x) = \pi_1(x) + \pi_2(x) + ... + \pi_j(x)$$

el modelo anterior se simplifica y toma la forma de un modelo linear logístico:

$$log[\gamma_j(x)/\{1 - \gamma_j(x)\}] = \theta_j - \beta^T x \ 1 \geq j < k$$

en donde $\theta_j = log\kappa_j$.
La estimación de los parámetros se puede hacer por distintos métodos utilizados para los modelos lineares generalizados.

II-A. Variable latente

Este modelo fue propuesto por P. McCullagh (1980) y está motivado por la existencia de una variable latente que definiremos como $U(x)$ la cual no puede ser observada y permite hacer la siguiente analogía:

$$y_i = j, \ U(x_i)\epsilon[\theta_{j-1}, \theta_j]$$

Es decir que cuando y pertenece a una cierta categoría, la variable latente cae en un cierto rango acotado por las probabilidades θ_{j-1}y θ_j. La naturaleza de esta variable permite hacer distintas suposiciones acerca de su distribución que permiten simplificar el modelo y estimar sus parámetros.

III. Modelo de Riesgo Proporcional

La función de riesgo es muy utilzada en el análisis de supervivencia y define la probabilidad de supervivencia de un individuo/objeto a un tiempo dado t, condicionado a su supervivencia a ese momento. Para un individuo con covariables x, esta función se define de la siguiente manera:

$$\lambda(t; x) = \lambda_0(t) exp(-\beta^T x$$

en donde $\lambda_0(t)$ es la función de riesgo para covariables $x = 0$ y β es un vector de parámetros desconocidos. Si definimos una función $S(t; x)$ que nos de la probabilidad de supervivencia más allá del tiempo t bajo ciertos valores de covarianzas, entonces tendríamos que integrar el lado derecho de la ecuación de riesgo proporcional. Obteniendo, como en el anterior modelo, el radio de la función de supervivencia para las covariables $x_2 - x_1$ el modelo se simplifica de la siguiente manera:

$$log\{S(t; x_1)\}/log\{S(t; x_2)\} = exp\{\beta^T(x_2 - x_1)\}$$

Para datos discretos si definimos como $1 - \gamma_j(x)$ la probabilidad de supervivencia más allá de la categoría j dado x, y aplicamos la función logarítmica de ambos lados de la ecuación, entonces tenemos una estructura lineal que es análoga al modelo de probabilidades proporcionales:

$$log[-log\{1-\gamma_j(x)\}] = \theta_j - \beta^T x$$

IV. Propiedades de los modelos de regresión ordinal

IV-A. Orden estocástico

Dado que ambos modelos presentados previamente, presentan la misma forma general, es decir:

$$link\{\gamma_j(x)\} = \theta_j - \beta^T x,$$

Esta función $link$ puede ser utilizada en lugar del modelo Logit presentado anteriormente, siempre y cuando esta función permita un mapeo del intervalo $(0,1)$ al $(-\infty,\infty)$. Algunos ejemplo de funciones que se pueden utilizar se presentan en la siguiente tabla.

Model	Inverse link function
Logit	$ln(\frac{\Delta}{1-\Delta})$
Probit	N^{-1}
Complementary log-log	$ln(-ln(1-\Delta))$

Sin embargo, todas ellas describen un orden estocástico estricto, esto quiere decir que si existe un cambio en las covariables, y por lo tanto en la variable respuesta, esto se verá observado en el mapeo que realiza la función *link*, y por lo tanto en la variable latente.

IV-B. Reversibilidad e Invarianza

Los modelos disctuidos en el presente son invariantes al orden reverso de las categorías pero no a las permutaciones arbitrarias. Esto no sucede para todas las funciones presentadas en la tabla anterior, sin embargo dependiendo de la aplicación es que se debe observar si el modelo es sensible a estas permutaciones arbitrarias, ya que de ser así podría no ser adecuado para variables ordinales.

V. Boosting para clasificación ordinal

V-A. Fixed Split Boosting

Método que reduce los problemas de clasificaciíon a $k-1$ problemas binarios definiendo una nueva variable, de la siguiente manera:

$$Y^{(r)} \begin{cases} 1 & Y\epsilon\{1,...,r\} \\ 2 & Y\epsilon\{r+1,...,k\} \end{cases} \quad (1)$$

para $r = 1,...,k-1$. Este método permite respetar el orden de las categorías, por lo que sólo es sensible para el tipo de variable ordinal. Sea $C^{(r)}(.,L)$ el clasificador para el problema binario definido por $Y^{(r)}$, para una r fija, a partir del clasificador agregado se obtendrá la predicción de clase de acuerdo a la siguiente ecuación:

$$C^{(r)}_{agg}(x) = argmax_j \sum_{m=1}^{M} c_m^{(r)} I(C^{(r)}(x, L_m^{(r)}))$$

en donde m es la m-ésima iteración del algoritmo, L_m la m-ésima versión del conjunto de entrenamiento, y c_m sus pesos correspondientes. Los resultados de este clasificador agregado se combinan en una segunda etapa en la que se transforman las clases obtenidas de la siguiente manera. Se desea obtener ahora la secuencia de predicción de clases $\hat{y}_1^{(r),...,\hat{y}_k^{(r)}}$, para cada clasificador agregado tendremos que esta secuencia toma los siguientes valores:

Si $C^{(r)}_{agg}(x) = 1$, con salida correspondiente a $\hat{Y}(x)\epsilon 1,...,r$ entonces tenemos que:

$$\hat{y}_1^{(r)}(x) = ... = \hat{y}_r^{(r)}(x) = 1/r,\ \hat{y}_{r+1}^{(r)}(x) = ... = \hat{y}_k^{(r)}(x) = 0,$$

Para $C^{(r)}_{agg}(x) = 2$, con salida correspondiente a $\hat{Y}(x)\epsilon r+1,...,k$:

$$\hat{y}_1^{(r)}(x) = ... = \hat{y}_r^{(r)}(x) = 0,$$
$$\hat{y}_{r+1}^{(r)}(x) = ... = \hat{y}_k^{(r)}(x) = 1/(k-r),$$

Obteniendo así el clasificador final de la forma:

$$C_{agg}(x) = argmax_j \sum_{r=1}^{k-1} \hat{y}_j^{(r)}(x)$$

De esta manera la predicción de clase para la observación x está favorecida porla mayoría ponderada de divisiones.

Algorithm 1 Ordinal Discrete AdaBoost

1: **for** $m = 0,1,...,M...$ **do**
2: Dados los pesos de la m-ésima iteración del ciclo de Boosting $w_1,...,w_n$, se construyen los clasificadores para cada problema binario $C^{(r)}(.,L_m)$
3: **for** $r = 0,1,...,k-1...$ **do**
4: Se corre la versión del conjunto de datos de entrenamiento para cada clasificador $C^{(r)}(.,L_m)$. Los resultados se combinan por *mayoría de voto* para obtener el clasificador $C(.,L_m)$
5: **end for**
6: Se actualizan los pesos muestrales de la siguiente manera:

$$\epsilon_i = \frac{|C(x_i,L_m)-Y_i|}{k-1}$$
$$e_m = \sum_{i=1}^{n} w_i\epsilon_i$$
$$b_m = \frac{1-e_m}{e_m}$$
$$c_m = log(b_m)$$
$$w_{i,new} = \frac{w_i exp(c_m\epsilon_i)}{\sum_{j=1}^{n} w_j exp(c_m\epsilon_j)}$$

7: **end for**
8: Después de las M iteraciones, se obtiene la clasificación de cada observación de acuerdo a: $argmax_j(\sum_{m=1}^{M} c_m I(C(x,L_m)=j))$

De manera alternativa se pueden usar como medidas de error en lugar del $\epsilon_i = \frac{|C(x_i,L_m)-Y_i|}{k-1}$, la distancia no estandarizada:

$$\epsilon_i = |C(x_i,L_m) - Y_i|$$

o la distancia estandarizada cuadrada:

$$\epsilon_i = (\frac{|C(x_i,L_m)-Y_i|}{k-1})^2$$

o el error simple:

$$\epsilon_i = 1 - I(C(x_i,L_m) = Y_i)$$

Además la medida de peso c_m propuesta anteriormente se utiliza principalmente para problemas de dos clases. Para problemas multiclase se sugiere la siguiente adaptación en caso de que se use como medida de error el error simple:

$$c_m = log(\frac{(1-e_m)(k-1)}{e_m})$$

para la distancia simple (el caso propuesto en el algoritmo), se propone la siguiente adaptación:

$$c_m = log(\frac{1-e_m}{e_m})$$

V-B. *Ordinal Real AdaBoost*

La construcción del método AdaBost Real Ordinal, también se basa en los problemas binarios entre las clases $\{1, ..., r\}$ y $\{r+1, ..., k\}$, pero de cada problema se extraen las probabilidaddes:

$$p^{(r)}(x) = \hat{P}(Y \leq r|x)$$

Se obtienen a partir de estas probabilidades las puntuaciones:

$$\hat{y}_1^{(r)}(x) = ... = \hat{y}_r^{(r)}(x) = p^{(r)}(x),$$
$$\hat{y}_{r+1}^{(r)}(x) = ... = \hat{y}_k^{(r)}(x) = 1 - p^{(r)}(x),$$

Y la predicción de la observación x_i está dada por:

$$\hat{y}_j(x) = \sum_{r=1}^{k-1} \hat{y}_j^{(r)}(x)$$

Algorithm 2 Ordinal Real AdaBoost

1: **for** $m = 0, 1, ..., M \ldots$ **do**
2: Dados los pesos de la m-ésima iteración del ciclo de Boosting $w_1, ..., w_n$, se construyen los clasificadores para las clases ordenadas $1, ..., k$: $\hat{y}_j(x) = \sum_{r=1}^{k-1} \hat{y}_j^{(r)}(x)$
3: Dadas las predicciones a partir de este clasificador se obtiene:
$$f_j(x_i, L_m) = 0{,}5 * log \frac{\hat{y}_j(x_i)}{(\prod_{l \neq j}^{k} \hat{y}_j(x_i))^{\frac{1}{k-1}}}$$
4: se actualizan los pesos de la siguiente manera:
$$\epsilon_i = \frac{|C(x_i, L_m) - Y_i|}{k-1}$$
$$e_m = \sum_{i=1}^{n} w_i \epsilon_i$$
$$b_m = \frac{1 - e_m}{e_m}$$
$$c_m = log(b_m)$$
$$w_{i,new} = \frac{w_i exp(-f_{Y_i}(x_i, L_m))}{\sum_{j=1}^{n} w_j exp(-f_{Y_j}(x_j, L_m))}$$
5: **end for**
6: Después de las M iteraciones, se obtiene la clasificación de cada observación de acuerdo a: $argmax_j(\sum_{m=1}^{M} f_j(x, L_m))$

VI. Aplicación

Se utilizaron las siguientes bases de datos para realizar una demostración de la funcionalidad de la Regresión Ordinal.

- **Universidad**. Base de datos simulada por la UCLA que predice la posibilidad de que un alumno aplique a alguna Universidad. La variable respuesta tiene 3 niveles: *Unlikely, Somewhat likely, Very likely*.
- **Australian Open 2013**. Base de datos con estadísticas de las jugadoras del torneo en 2013. Se utilizó como variable respuesta la última ronda a la que llegaron las jugadoras (Supervivencia en el torneo).
- **CTG**. Base de datos con medidas de Cardiotocogramas fetales. La variable respuesta es el estado del feto y tiene 3 niveles: *Normal, Sospechoso, Patológico*

En una primera etapa se utilizaron métodos de regresión para clasificar en todas las bases de datos: El primero consistió en clasificación utilizando un método ya implementado de regresión logística regular en la librería de SKLEARN de Python, el segundo consistió en hacer una regresión logística ordinal y realizar la clasificación a partir de los valores de cortes (θ_j), y el último de igual manera fue a partir de regresión logística ordinal pero con un método optimizado implementado ya en la librería MORD de Python. Se muestran los resultados de accuracy score para cada uno de los métodos en las siguientes imágenes. En esta primera etapa sólo mostramos como medida de comparación esta puntuación debido a la diferencia de la naturaleza de ambos conjuntos de datos.

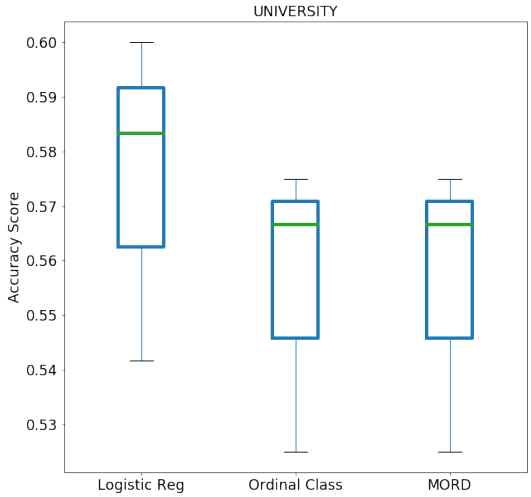

Figura 1. Accuracy Métodos de Clasificación basados en Regresión. Conjunto de datos: Universidad

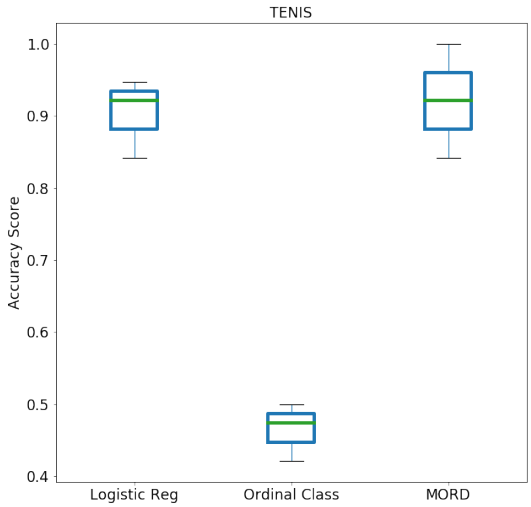

Figura 2. Accuracy Métodos de Clasificación basados en Regresión. Conjunto de datos: Australian Open

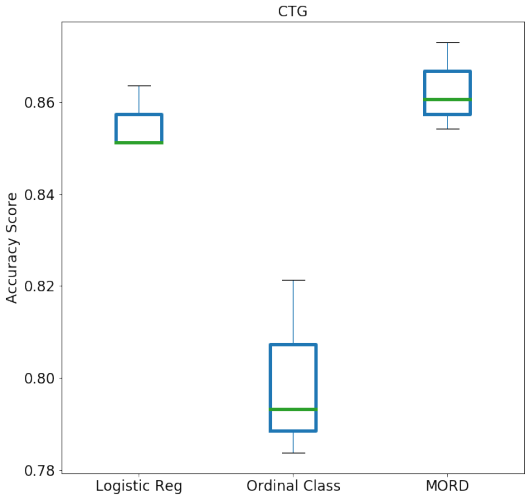

Figura 3. Accuracy Métodos de Clasificación basados en Regresión. Conjunto de datos: CTG

En una segunda etapa sólo se utilizaron los datos correspondientes a los Cardiotocogramas. Se eligió este conjunto de datos debido a la importancia que tienen tanto lograr una muy buena precisión de clasificación, como reducir el número de diagnósticos equivocados para fetos con categoría patológicos, que en realidad son normales (falsos positivos) o viceversa (falsos negativos). Se utilizaron dos métodos de análisis: El primero consiste en utilizar un árbol de clasificación regular y el segundo consiste en realizar una adaptación de este método de clasificación a $k-1$ problemas binarios y a partir de una *mayoría de votos* de estos clasificadores, hacer una asignación final de la clase. Se muestran a continuación para ambos métodos tanto las matrices de confusión como las puntuaciones con respecto a los falsos positivos como se definieron anteriormente. Esta última puntuación se calcula con el radio de precisión:

$$\frac{tp}{tp+fp}$$

En donde tp son los verdaderos positivos (patológicos clasificados correctamente) y fp los falsos positivos (patológicos clasificados como normales).

	Ctree Reg	CTree Ordinal
0	0.918495	0.926332
1	0.916928	0.934169
2	0.920063	0.927900

Figura 4. Scores de precisión en 3 distintos muestreos de los datos

Como se puede ver los scores de precisión para el método ordinal, superaron a los de la clasificación regular en todos los muestreos. En las matrices de confusión presentadas a continuación se puede observar que el método regular tiene un mayor número de diagnósticos de fetos patológicos correctos. Las principales diferencias entre ambos métodos se presentan en la cantidad de diagnósticos que se hicieron de fetos sospechosos pero que realmente son normales (mayor número en el método ordinal) o que eran sospechosos y se diagnosticaron como normales (mayor número en el método regular).

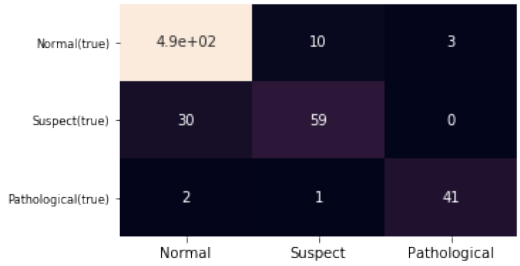

Figura 5. Matriz de Confusión - Árbol Ordinal

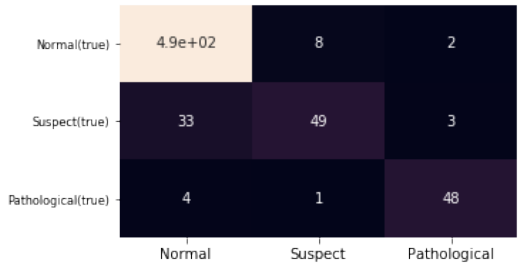

Figura 6. Matriz de Confusión - Árbol Regular

A partir de estas matrices se obtuvieron las puntuaciones de precisión de $0{,}95$ para el método Ordinal y $0{,}92$ para el método Regular. Por lo que en esta aplicación en específico en donde es de mucha relevancia tener una clasificación correcta de los diagnósticos patológicos el método de clasificación de un árbol de decisión adaptado a tratar los problemas binarios de las variables ordenadas obtiene mejores resultados.

VII. Conclusiones

A pesar de la importancia que tienen las variables ordinales en diversas áreas como lo son las ciencias sociales y las ciencias de la salud, los métodos de clasificación para variables ordinales han sido poco implementados en el área de *machine learning*. Una de las razones que se pudo encontrar en el presente trabajo es la falta de atractivo en contraste con los métodos estándar en problemas reales en donde las características de orden de las variables ordinales pueden ser ignoradas sin mayores consecuencias negativas. Sin embargo, se encontró que, por ejemplo, en la aplicación de diagnóstico de gravedad del estado de salud de un feto a partir de las características de su CTG, el *score* de precisión de diagnósticos correctos vs diagnósticos erróneos con respecto a una patología del feto es mayor para el enfoque ordinal que se le dio al árbol de clasificación utilizado. La naturaleza de las variables ordinales requiere la implementación de métodos que permitan explotar y explorar las ventajas que tiene trabajar con ellas. Uno de los enfoques no explorados en el presente es la modelación a partir

de la variable latente, la cual tiene propiedades que podrían ser aprovechadas para mejorar los métodos mencionados.

REFERENCIAS

[1] P. McCullagh. *Regression models for ordinal data*. Journal of the Royal Statistical Society B, 42 (2):109–142, 1980.

[2] P. McCullagh and J. A. Nelder. *Generalized Linear Models*. Chapman and Hall, London, 1983.

[3] Hechenbichler K., Tutz G. *Bagging, Boosting and Ordinal Classification*. In: Weihs C., Gaul W. (eds) Classification — the Ubiquitous Challenge. Studies in Classification, Data Analysis, and Knowledge Organization. Springer, Berlin, Heidelberg. 2005.

[4] Herbrich, R., Graepel, T., and Obermayer, K. *Regression Models for Ordinal Data: A Machine Learning Approach*. 1999.

[5] Frank, E. and Hall, M., *A Simple Approach to Ordinal Classification*. 12th European Conference. 2001.