Clasificación ordinal: oportunidades, retos y aplicaciones

Escuela de Verano de Inteligencia Artificial (EVIA)
Universidad Pablo de Olavide

Pedro Antonio Gutiérrez (pagutierrez@uco.es)

14 de junio de 2023

Grupo de investigación AYRNA, http://www.uco.es/ayrna, Universidad de Córdoba.

Introducción

Aprendizaje supervisado

- Conjunto de descriptores para cada tumor:

Tamaño tumor	Textura	Perímetro	...	Resultado
18.02	27.60	117.50	...	N
17.99	10.38	122.80	...	N
20.29	14.34	135.10	...	R

- Variable de salida:
  - *Resultado*:
    - R: recurrencia (reaparición) del tumor después de la quimioterapia.
    - N: el tumor no vuelve a aparecer después de la quimioterapia.
  - Dado un nuevo paciente (que no está en nuestra tabla), queremos predecir si el tumor va a reincidir (N o R).

Aprendizaje supervisado

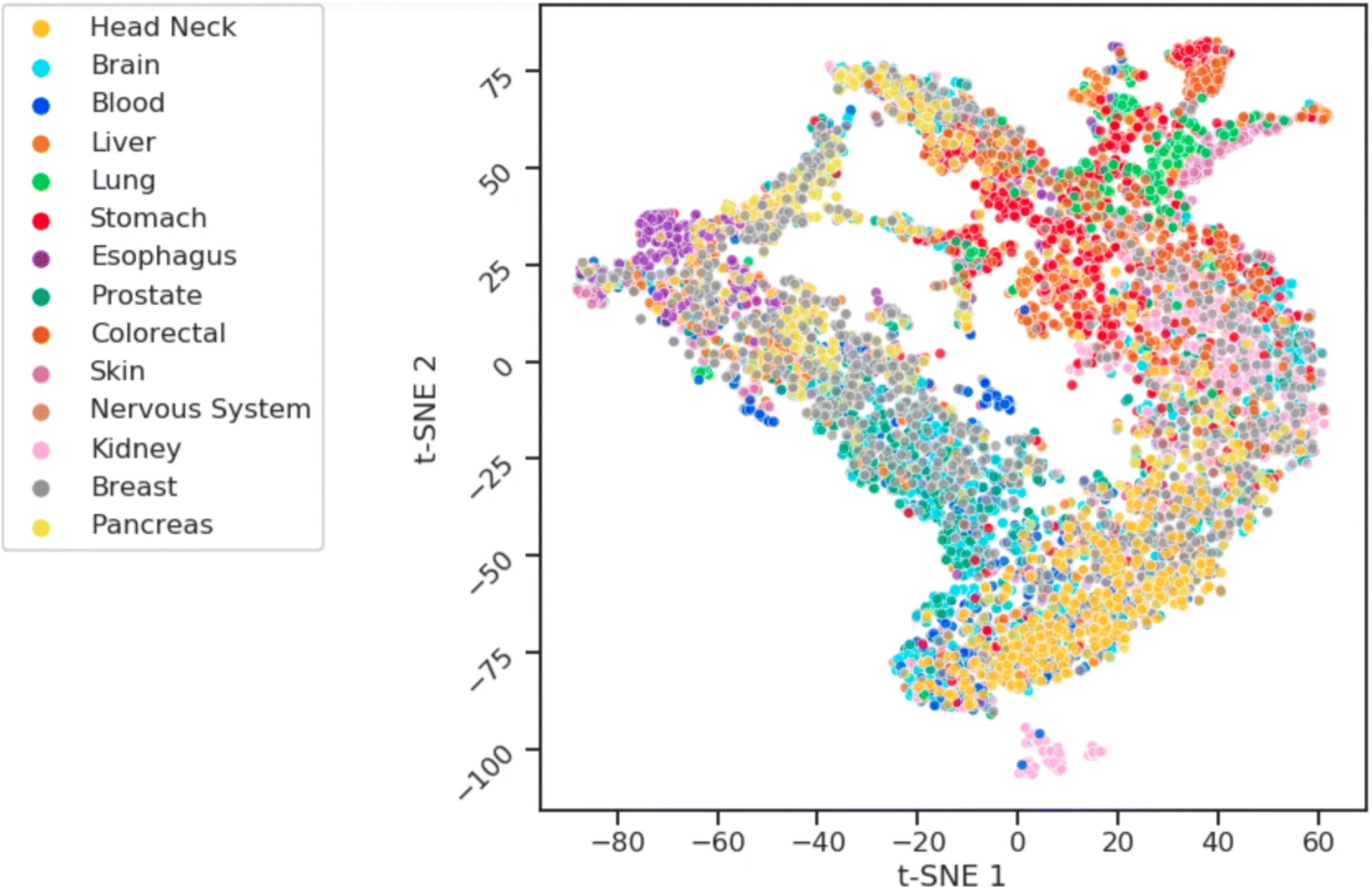
Tamaño tumor	Textura	Perímetro	...	Resultado
18.02	27.60	117.50	...	?
17.99	10.38	122.80	...	?
20.29	14.34	135.10	...	?

Naturaleza de la variable dependiente

- Clasificación binaria: si la variable a predecir es de tipo categórico con dos posibles resultados (recurrente o no recurrente).
- Regresión: si la variable a predecir es de tipo continuo, cuantitativo (p.ej., esperanza de vida del paciente).
- Clasificación ordinal: categorías ordenadas (no recurrente, recurrencia en menos de un año, recurrencia en un año o más).

Clasificación nominal

Tipos de tumores:

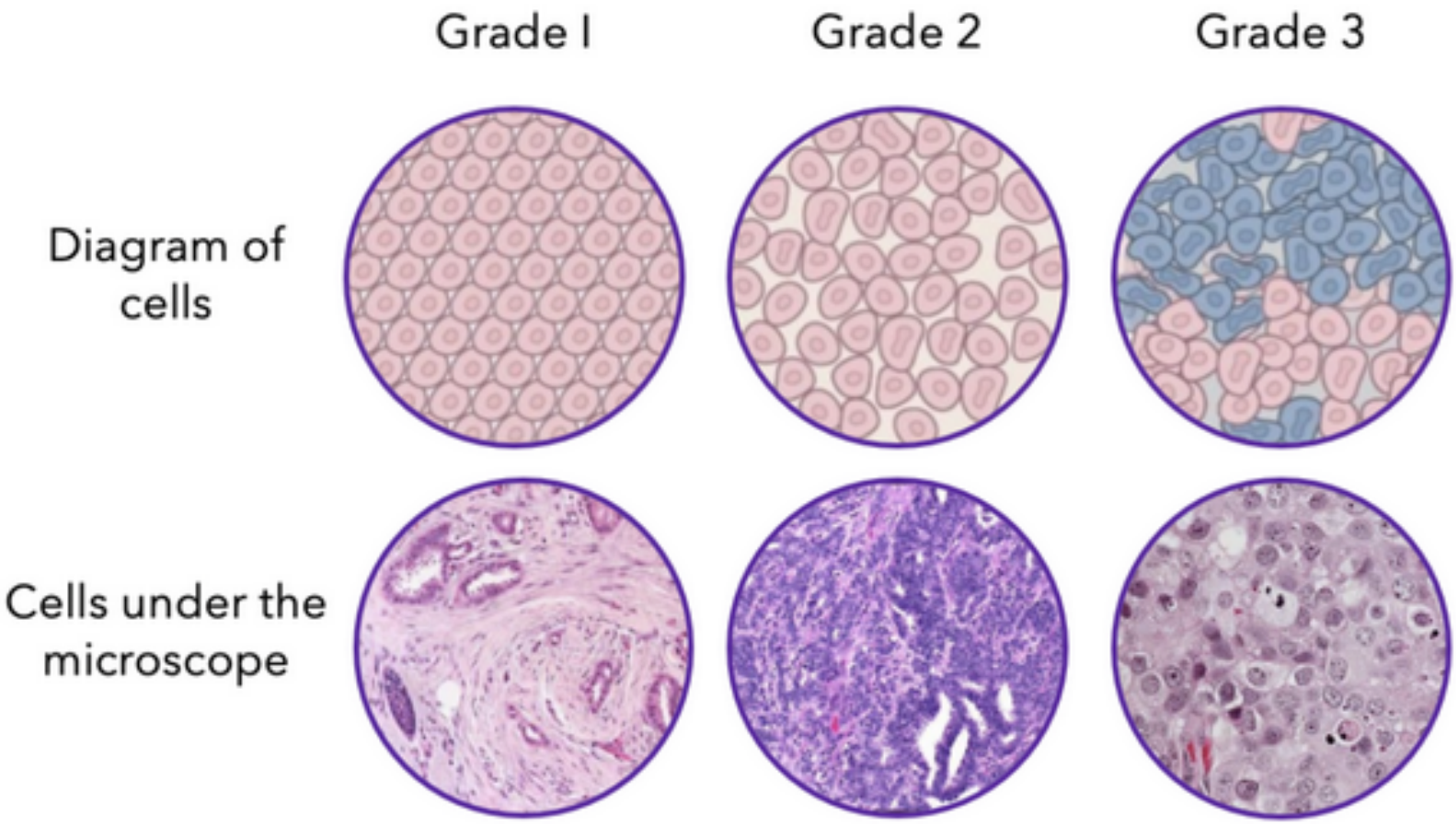


Clasificación ordinal

- Clasificación ordinal $\rightarrow$ (Relativamente) nuevo problema de aprendizaje.
  - Aprender una regla para predecir categorías o etiquetas en una escala ordinal.
  - Las etiquetas son discretas pero hay un orden natural entre ellas.
  - También llamado regresión ordinal.
- Por ejemplo:
  - Un profesor evalúa a sus estudiantes con las notas A, B, C, D, y E, y sabemos que $A \succ B \succ C \succ D \succ E$

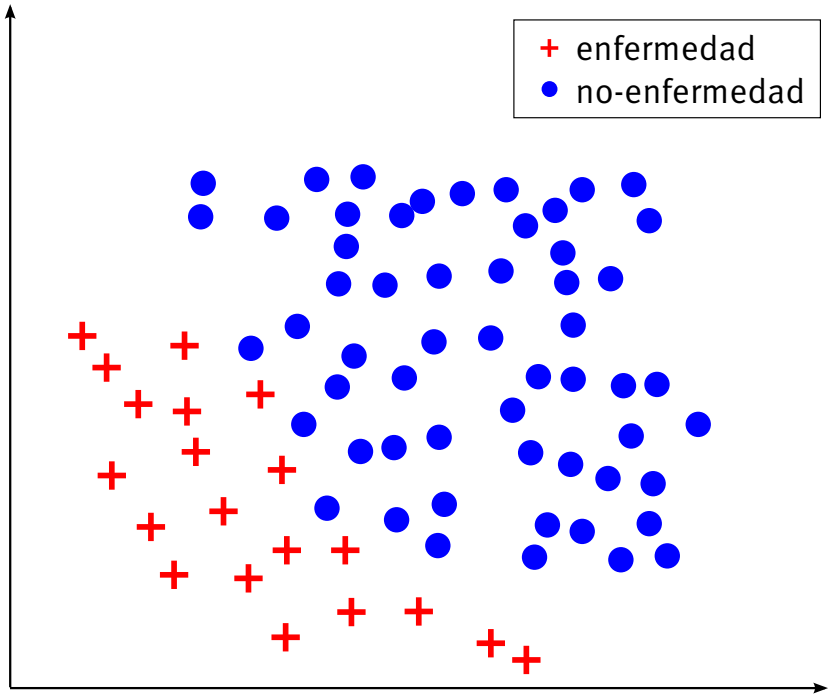
Clasificación ordinal: algunos ejemplos

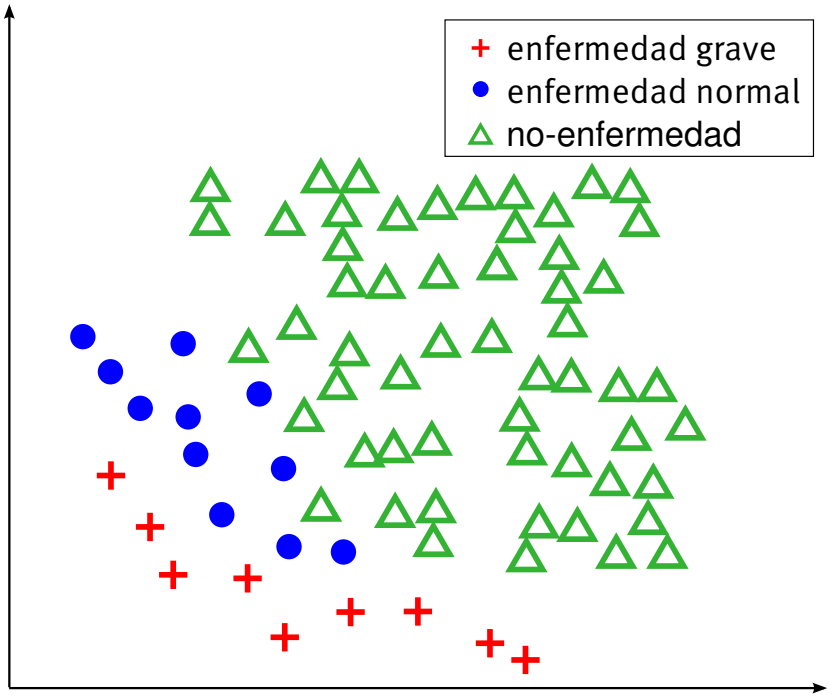
- Ejemplo: *breast cancer*.



¿Regresión? ¿clasificación?

Clasificación binaria Vs. clasificación ordinal:

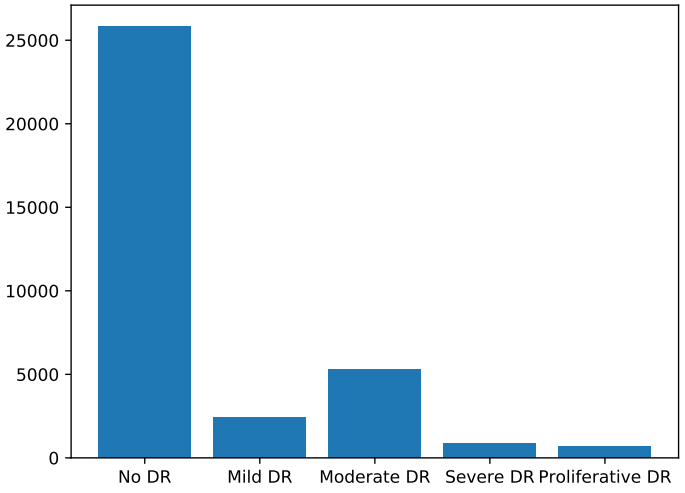


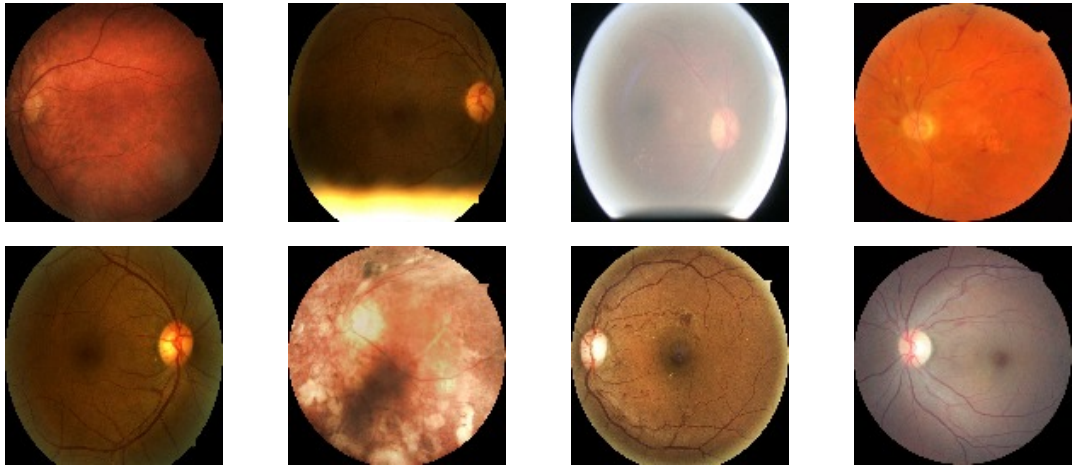


Clasificación ordinal: algunos ejemplos

Diabetic Retinopathy (DR) (Competición Kaggle)

- Imágenes de alta resolución de fondo del ojo para detectar DR.
- Train/test: 35126/53576.
- **5 clases** (no DR $\prec$ DR leve $\prec$ DR moderada $\prec$ DR severa $\prec$ DR proliferativa).





Clasificación ordinal: algunos ejemplos

- Ejemplo: Edad de personas en imágenes.

1:*bebé* 2:*niño* 3:*adolescente* 4:*adulto*

Particularidades

- Las categorías llevan una información de orden: "*niño*" es más joven que "*adulto*".
- Las categorías no llevan información numérica: "*niño*" no es necesariamente la mitad de joven que "*adulto*".

¿Regresión? ¿clasificación?

Clasificación ordinal $\rightarrow$ entre clasificación y regresión

- Distinto de regresión $\rightarrow$ la variable de respuesta es discreta y finita y la distancia entre los rangos no está definida.
- Distinto de clasificación $\rightarrow$ hay un orden entre las categorías.

- Muchos problemas reales requieren la clasificación de objetos en categorías ordenadas:
  - Medicina.
  - Valoración de crédito.
  - Reconocimiento de edad en imágenes.
  - Análisis de riesgo.
  - *Ranking* de universidades.

Clasificación ordinal: subjetividad

- Los problemas de clasificación ordinal generalmente incluyen a seres humanos en la elaboración de la base de datos:
  - La evaluación subjetiva es común cuando intervienen seres humanos.
    - Por ejemplo, el proyecto *European social survey* proporciona datos de encuesta acerca del bienestar de los individuos, que pueden servir para entender mejor los factores que llevan a la felicidad.
  - Las herramientas de *rating* son útiles cuando hay evaluación subjetiva.
    - Las escalas ordinales (escalas de Likert) son una forma de proporcionar información de *rating* imprecisa.
    - También es muy útil cuando varios expertos tienen que proporcionar opiniones sobre los mismos objetos.
    - Por ejemplo, muchos problemas de diagnóstico médico implican una evaluación subjetiva del caso.

Clasificación ordinal: definición formal

Cada ejemplo de entrenamiento $(\mathbf{x}_i, y_i)$ está compuesto de:

- Un vector de entrada $\mathbf{x}_i \in \mathbb{R}^n$.
- Una etiqueta ordinal $y_i \in Y = \{\mathcal{C}_1, \mathcal{C}_2, \ldots, \mathcal{C}_Q\}$,
- pero las etiquetas cumplen la restricción $\mathcal{C}_1 \prec \mathcal{C}_2 \prec \cdots \prec \mathcal{C}_Q$.
- El rango de una etiqueta ordinal se puede definir como $\mathcal{O}(\mathcal{C}_i) = i \Rightarrow$ posición de la etiqueta en la escala ordinal.

Objetivo

Encontrar una función $f : X \to Y$

Métricas de evaluación

Matrices de confusión

- Dado un problema de clasificación con Q clases y N patrones de entrenamiento o test con un clasificador g, la matriz de confusión o de contingencia es:

$$M(g) = \left\{ n_{ij};\ \sum_{i,j=1}^{Q} n_{ij} = N \right\} \tag{1}$$

donde n_{ij} representa el número de patrones de la clase i que el clasificador clasifica como j.

Matrices de confusión

Cuadro 1: Matriz de confusión de un clasificador

Clase	1	2	...	Q	Total
1	n_{11}	n_{12}	...	n_{1Q}	$n_{1\bullet} = \sum_{i=1}^{Q} n_{1i}$
2	n_{21}	n_{22}	...	n_{2Q}	$n_{2\bullet} = \sum_{i=1}^{Q} n_{2i}$
...	...	...	...	...	...
Q	n_{Q1}	n_{Q2}	...	n_{QQ}	$n_{Q\bullet} = \sum_{i=1}^{Q} n_{Qi}$

Matrices de confusión

Definiciones

- Porcentaje de patrones bien clasificados, *accuracy*, $Acc = (1/N)\sum_{j=1}^{Q} n_{jj}$.
- También llamado *Correct Classification Rate* (CCR).
- Número de patrones de la clase i, $n_{i\bullet} = \sum_{j=1}^{Q} n_{ij}$, $i = 1, \ldots, Q$.
- Probabilidad a priori de la clase i, $p_i = (n_{i\bullet}/N) \;\; i = 1, \ldots, Q$.

Problemas del accuracy

- Ejemplo de evaluación de estudiantes:

Clase	A	B	C	D	E	Total
A	15	0	0	0	15	30
B	0	15	0	0	15	30
C	8	0	15	0	7	30
D	15	0	0	15	0	30
E	10	0	0	0	20	30

$$Acc = 80/135 = 0{,}5926$$

¡¡15 estudiantes de A evaluados como E!!

¡¡15 estudiantes de B evaluados como E!!

...

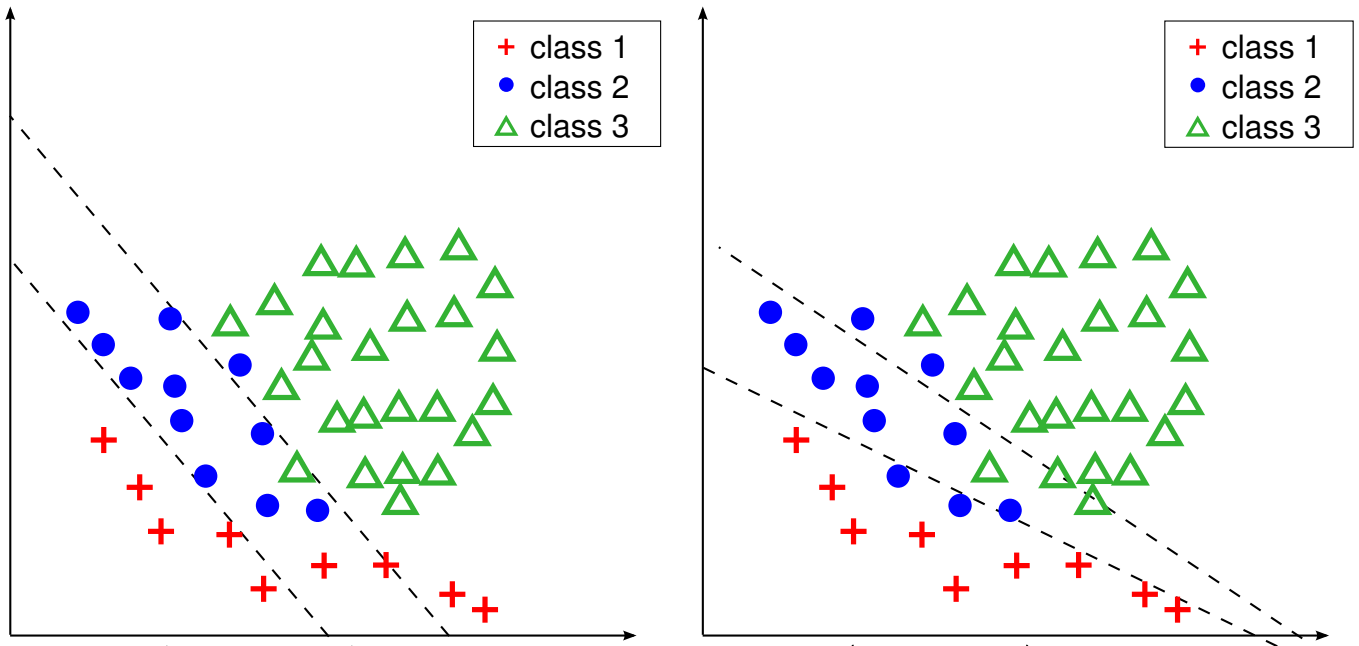
Problemas del accuracy

- Mismo *accuracy* pero mejor orden:

Clase	A	B	C	D	E	Total
A	15	15	0	0	0	30
B	0	15	15	0	0	30
C	0	0	15	15	0	30
D	0	0	0	15	15	30
E	0	0	0	10	20	30

$$Acc = 80/135 = 0{,}5926$$

Necesidad de medidas específicas



$$\mathbf{M}_{\mathrm{A}} = \begin{pmatrix} 5 & 2 & 2 \\ 0 & 10 & 1 \\ 0 & 1 & 18 \end{pmatrix}, \quad \begin{array}{r} CCR = 0.846 \\ MAE = 0.205 \\ AMAE = 0.270 \end{array}$$

a)

$$\mathbf{M}_{\mathrm{B}} = \begin{pmatrix} 9 & 0 & 0 \\ 2 & 7 & 2 \\ 0 & 3 & 17 \end{pmatrix}, \quad \begin{array}{r} CCR = 0.825 \\ MAE = 0.175 \\ AMAE = 0.171 \end{array}$$

b)

El clasificador a) es más preciso, pero el b) respeta mejor el orden

Coste de mala clasificación

- Coste de una mala clasificación:
  - No podemos comparar $\mathcal{C}_4$ y $\mathcal{C}_2$ numéricamente, pero podemos asignar un coste artificial cuando $\mathcal{C}_2$ se confunde con $\mathcal{C}_4$ (una foto "*niño*" etiquetada como "*adulto*").
  - Los costes se organizan en una matriz: C_{ij} es el coste de clasificar una instancia de la clase $\mathcal{C}_i$ como $\mathcal{C}_j$.
  - Elecciones razonables:

$$C_{01} = \begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix} \qquad C_{\mathrm{abs}} = \begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix} \tag{2}$$

Coste cero-uno (nominal) Coste absoluto (ordinal)

Información de orden

Mean absolute error (MAE)

- Cuantificar la precisión de N predicciones ordinales $\{\hat{y}_1, \ldots, \hat{y}_N\}$ con respecto a las etiquetas reales $\{y_1, \ldots, y_N\}$.
- MAE: desviación media de la predicción con respecto a la etiqueta real, cuando se tratan como enteros consecutivos.

$$MAE = \frac{1}{N} \sum_{i=1}^{N} |\mathcal{O}(y_i) - \mathcal{O}(y_i^*)| \,, \tag{3}$$

donde $\mathcal{O}(\mathcal{C}_i) = i$ y el rango es $MAE \in [0, Q-1]$.

  - Supone que las categorías se distribuyen de forma uniforme.
  - No tiene en cuenta que las clases pueden ser de distinto tamaño (probabilidades a priori).

Información de orden

- De la misma forma que la medida *accuracy* sería equivalente a utilizar un matriz de costes 01 sobre la matriz de confusión:

$$A = 1 - \frac{\sum_{i,j} (C_{01})_{ij} n_{ij}}{N}, \tag{4}$$

- la medida *MAE* es equivalente a considerar costes absolutos:

$$MAE = \frac{\sum_{i,j} (C_{\mathrm{abs}})_{ij} n_{ij}}{N}. \tag{5}$$

Información de orden y desequilibrio de clases

Average mean absolute error ($AMAE$)

- Mitiga el efecto de una distribución desigual de patrones por clase. Definimos el MAE para la clase q:

$$MAE_q = \frac{1}{n_{q\bullet}} \sum_{k=1}^{Q} |q-k| n_{qk}, 1 \leq q \leq Q.$$

- El $AMAE$ se define como:

$$AMAE = \frac{1}{Q} \sum_{q=1}^{Q} MAE_q, \text{ donde } AMAE \in [0, Q-1].$$

Maximum mean absolute error ($MMAE$)

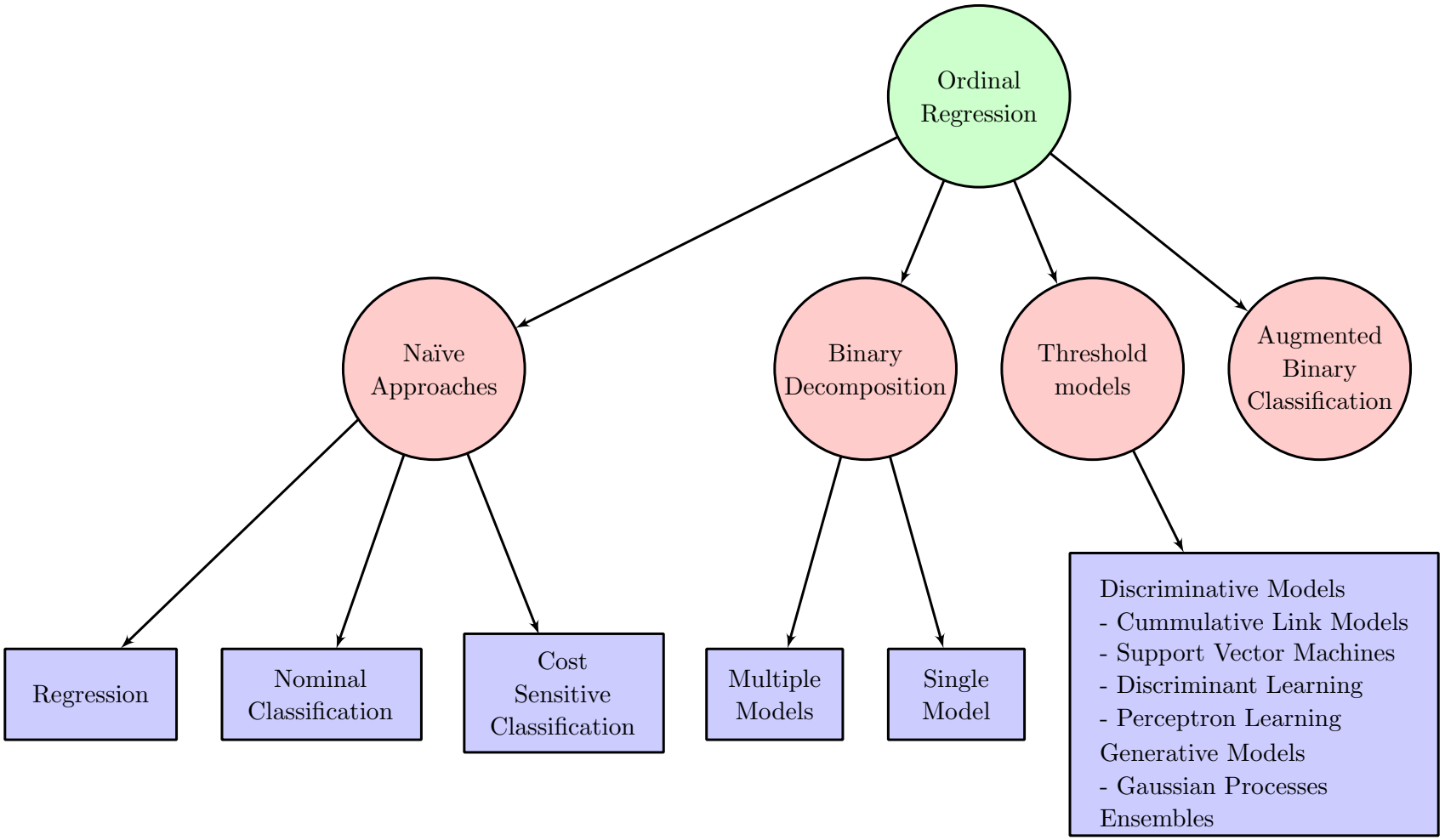
$$MMAE = \text{máx}\{MAE_q; q = 1, ..., Q\},$$

Otras métricas de evaluación

- Hay otras opciones basadas en medir la asociación estadística entre las predicciones y los valores objetivo:
  - Coeficiente de correlación de Spearman (r_{S}): supone que las etiquetas son enteros consecutivos.
  - τ de Kendall (τ_{b}): evalúa todos los pares de patrones observando si el orden asignado es consistente con el esperado.
  - Kappa ponderado (*WKappa*): tiene en cuenta la magnitud del acuerdo que se podría atribuir al azar e incluye pesos para considerar la ordinalidad.
- Análisis ROC para clasificación ordinal [Waegeman et al., 2008].

M. Cruz-Ramírez, C. Hervás-Martínez, J. Sánchez-Monedero y P. A. Gutiérrez. "Metrics to guide a multi-objective evolutionary algorithm for ordinal classification", Neurocomputing, Vol. 135, July, 2014, pp. 21-31.

Taxonomía



P.A. Gutiérrez, M. Pérez-Ortiz, J. Sánchez-Monedero, F. Fernandez-Navarro y C. Hervás-Martínez. "Ordinal regression methods: survey and experimental study", IEEE Transactions on Knowledge and Data Engineering, Vol. 28(1), January, 2016, pp. 127-146.

Framework ORCA

- ORCA es un *framework* en MATLAB/Octave que incluye los distintos métodos de clasificación ordinal comparados.
- Además permite automatizar la selección de parámetros.
- Está disponible bajo licencia GPLv3 en `https://github.com/ayrna/orca/`.
- `https://github.com/ayrna/orca-python/`: en preparación.

J. Sánchez-Monedero, P. A. Gutiérrez and M. Pérez-Ortiz, "ORCA: A Matlab/Octave Toolbox for Ordinal Regression", Journal of Machine Learning Research. Vol. 20. Issue 125. 2019.

Aproximaciones ingenuas

Simplificar a clasificación nominal

Simplificar a clasificación nominal

- La mayoría de trabajos.
- Ignoramos la información de orden lo que nos puede llevar a requerir más datos para aprender lo que realmente necesitamos [Harrington, 2003].
- En ORCA, se incluyen dos versiones de *Support Vector Classifier*, utilizando dos estrategias, "One-vs-One" y "Ove-Vs-All" (posteriormente, estos conceptos serán aclarados): SVC1V1 y SVC1VA.

Simplificar a regresión

Simplificar a regresión

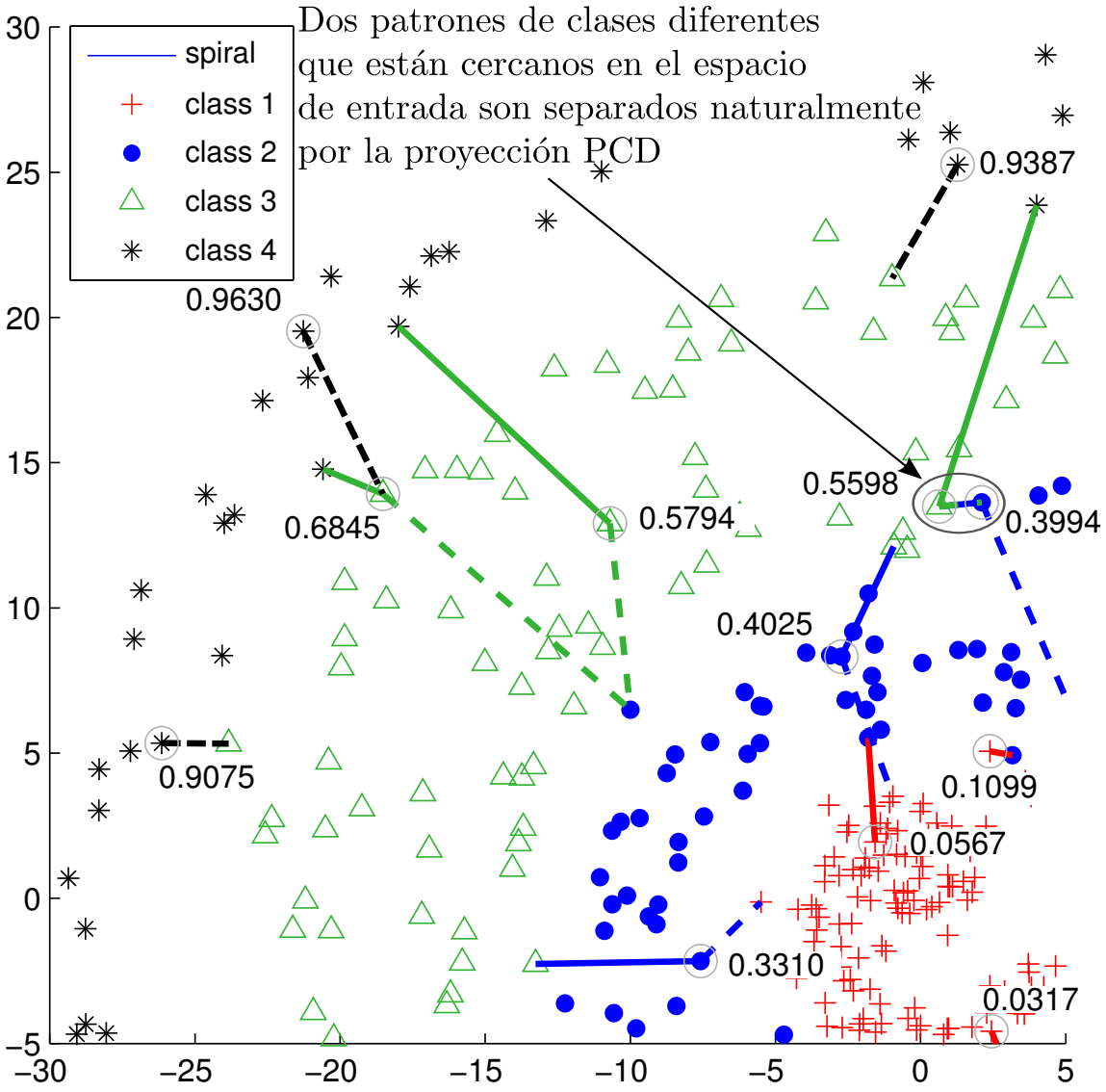
- Kramer, S., Pfahringer, B., Widmer, G., Groeve, M.D.: Prediction of ordinal classes using regression trees. Fundamenta Informaticae 47, 1001-1013 (2001)
- Transformar cada categoría en un valor numérico, aprender una función de regresión. En test, redondear predicciones al entero más cercano.
- Puede que no haya una forma fundamentada de diseñar una correspondencia apropiada ya que la distancia real entre la escalas ordinal suele ser no conocida.
- En ORCA, Support Vector Regression (SVR).

Simplificar a regresión

PCDOC

- J. Sánchez-Monedero, P. A. Gutiérrez, P. Tino, and C. Hervás-Martínez, "Exploitation of pairwise class distances for ordinal classification." *Neural Computation*, vol. 25, no. 9, pp. 2450–2485, 2013.
- Refinar la transformación a un problema de regresión, haciendo uso de las distancias entre los patrones de clases consecutivas.
- Técnica de proyección que ayuda a tener un regresor mejor adaptado al problema de clasificación ordinal.

PCDOC: proyección



Utilizar matrices de coste

Utilizar matrices de coste

- S.B. Kotsiantis and P.E. Pintelas: A Cost Sensitive Technique for Ordinal Classification Problems. LNAI 3025, 220-229 (2004)
- En ORCA, se incluye el método Cost Sensitive SVC (CSSVC), que aplica una estrategia similar a SVC1VA pero incluyendo los pesos correspondientes a la matriz de costes absolutos cuando los patrones son de otra clase.

Costes cero-uno	Costes absoluto	Costes cuadráticos
$\begin{pmatrix} 0 & 1 & 1 & 1 & 1 \\ 1 & 0 & 1 & 1 & 1 \\ 1 & 1 & 0 & 1 & 1 \\ 1 & 1 & 1 & 0 & 1 \\ 1 & 1 & 1 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 2 & 3 & 4 \\ 1 & 0 & 1 & 2 & 3 \\ 2 & 1 & 0 & 1 & 2 \\ 3 & 2 & 1 & 0 & 1 \\ 4 & 3 & 2 & 1 & 0 \end{pmatrix}$	$\begin{pmatrix} 0 & 1 & 4 & 9 & 16 \\ 1 & 0 & 1 & 4 & 9 \\ 4 & 1 & 0 & 1 & 4 \\ 9 & 4 & 1 & 0 & 1 \\ 16 & 9 & 4 & 1 & 0 \end{pmatrix}$

Descomposiciones Binarias

Descomposiciones binarias

Simplificar un problema de clasificación ordinal en múltiples problemas de clasificación binaria.

Descomposiciones nominales	
OneVsAll	*OneVsOne*
$\begin{pmatrix} +,-,-,-,- \\ -,+,-,-,- \\ -,-,+,-,- \\ -,-,-,+,- \\ -,-,-,-,+ \end{pmatrix}$	$\begin{pmatrix} -,-,-,-,\ ,\ ,\ ,\ ,\ , \\ +,\ ,\ ,\ ,-,-,-,\ ,\ , \\ \ ,+,\ ,\ ,+,\ ,\ ,-,-, \\ \ ,\ ,+,\ ,\ ,+,\ ,+,\ ,- \\ \ ,\ ,\ ,+,\ ,\ ,+,\ ,+,+ \end{pmatrix}$

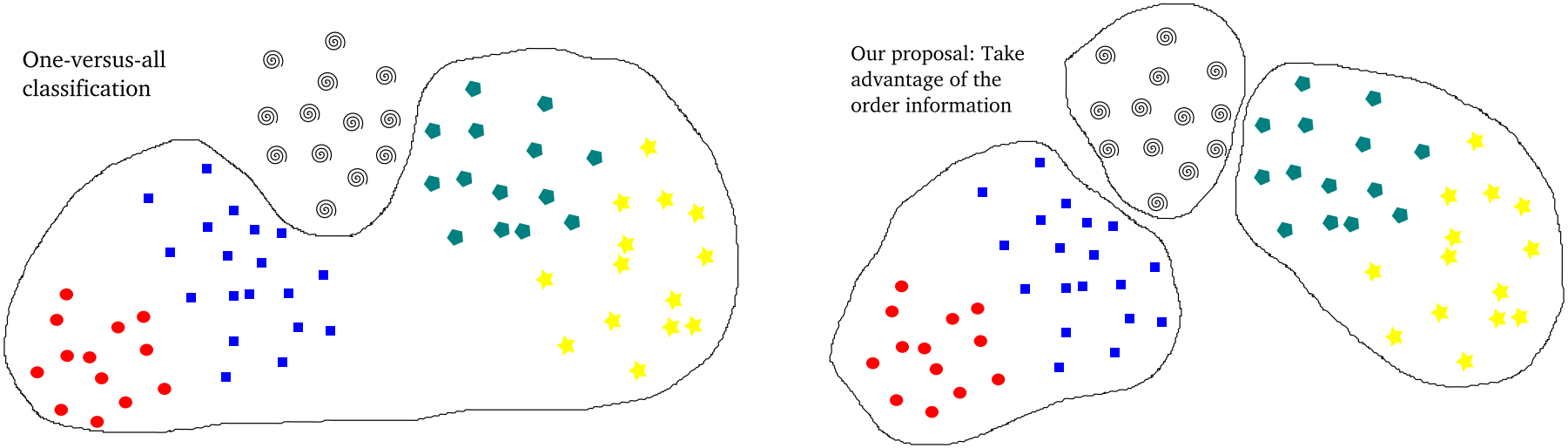
Descomposiciones ordinales			
OrderedPartitions	*OneVsNext*	*OneVsFollowers*	*OneVsPrevious*
$\begin{pmatrix} -,-,-,- \\ +,-,-,- \\ +,+,-,- \\ +,+,+,- \\ +,+,+,+ \end{pmatrix}$	$\begin{pmatrix} -,\ ,\ , \\ +,-,\ , \\ \ ,+,-, \\ \ ,\ ,+,- \\ \ ,\ ,\ ,+ \end{pmatrix}$	$\begin{pmatrix} -,\ ,\ , \\ +,-,\ , \\ +,+,-, \\ +,+,+,- \\ +,+,+,+ \end{pmatrix}$	$\begin{pmatrix} +,+,+,+ \\ +,+,+,- \\ +,+,-, \\ +,-,\ , \\ -,\ ,\ , \end{pmatrix}$

Descomposiciones binarias

- A la hora de aplicar estas descomposiciones, el aprendizaje puede organizarse de distintas formas:
  - Los subproblemas binarios pueden aprenderse por separado, utilizando un clasificador distinto para cada subproblema (columna) $\rightarrow$ aproximaciones con modelos múltiples.
  - Algunos modelos, como las redes neuronales artificiales, nos permiten aprender todos los subproblemas con un solo modelo.
- Por otro lado, hay que decidir como realizar la fase de predicción, una vez se obtiene una respuesta binaria para cada subproblema.

Descomposición en tres clases

Descomposición en tres clases



M. Pérez-Ortiz, P. A. Gutiérrez, and C. Hervás-Martínez, "Projection based ensemble learning for ordinal regression," *IEEE Transactions on Cybernetics*, vol. 44, no. 5, pp. 681–694, 2014.

Descomposición en tres clases

- El conjunto *All* puede dividirse en menores que la categoría analizada y mayores que la categoría analizada.
- En ORCA, este método se denomina *Ordinal Projection Based Ensemble* (OPBE).

Modelos de umbral

Modelos de umbral

- Para modelar un problema de clasificación ordinal desde una perspectiva de regresión, se puede asumir que existe una variable de respuesta con valores reales, pero que esos valores no son observables.
- Estimamos dos cosas:
  - Una función $f(\mathbf{x})$ que predice los valores reales.
  - Un vector de umbrales $\mathbf{b} \in \mathbb{R}^{Q-1}$ que representa rangos en el dominio de $f(\mathbf{x})$, donde $b_1 \leq b_2 \leq \ldots \leq b_{Q-1}$.
  - La función $f(\mathbf{x})$ descubrir la naturaleza de la variable latente.
  - El vector de umbrales $\mathbf{b}$ estima las posibles distintas escalas entre las categorías.

Aproximación general

Métodos de umbral

- Aprender una función $f : X \to \mathbb{R}$,
- junto con un conjunto de umbrales $b^1 \leq \cdots \leq b^{Q-1}$
- La regla para predecir es:

$$g(\mathbf{x}) = \begin{cases} \mathcal{C}_1, & \text{if } f(\mathbf{x}) \leq b^1 \\ \mathcal{C}_2, & \text{if } b^1 < f(\mathbf{x}) \leq b^2 \\ \cdots & \\ \mathcal{C}_Q, & \text{if } f(\mathbf{x}) > b^{Q-1} \end{cases} \tag{6}$$

donde $f(\mathbf{x})$ es una función de *ranking* o de proyección.

Proportional Odds Model (POM)

Proportional Odds Model (POM)

- Proviene de la comunidad estadística.
  McCullagh, P., Nelder, J.A.: Generalized Linear Models, 2nd edn. Chapman and Hall, Boca Raton (1989)
- Modelo lineal probabilista:

$$f(\mathbf{x}) = \sum_{i=1}^{k} \beta_i x_i \tag{7}$$

- Pertenece a un familia de modelos denominados modelos de enlace acumulativo (*Cummulative Link Models*, CLMs) [Agresti, 2010].

Proportional Odds Model (POM)

Proportional Odds Model (POM)

- Utilizando el modelo POM:

$$P(y \preceq \mathcal{C}_q) = P(y = \mathcal{C}_1) + \ldots P(y = \mathcal{C}_q), P(y \preceq \mathcal{C}_Q) = 1 \quad (8)$$

$$odds(y \preceq \mathcal{C}_q) = \frac{P(y \preceq \mathcal{C}_q)}{1 - P(y \preceq \mathcal{C}_q)} \quad (9)$$

$$\mathbf{w}^{\mathrm{T}}\mathbf{x} + b_q = logit(y \preceq \mathcal{C}_q) = \ln\left(\frac{P(y \preceq \mathcal{C}_q)}{1 - P(y \preceq \mathcal{C}_q)}\right) \quad (10)$$

- Esto hace que el ratio de los odds entre dos patrones sea proporcional a su diferencia.

Cummulative Link Models (CLMs)

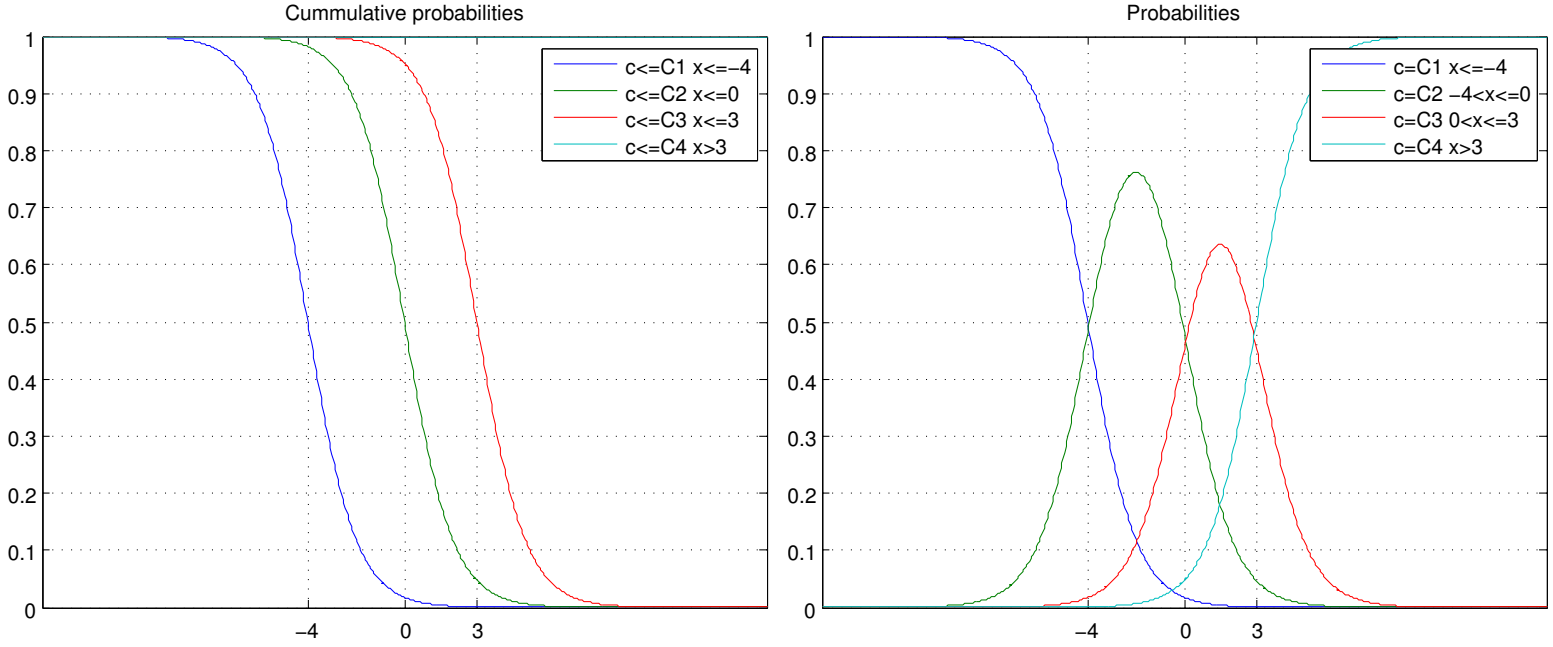
- Hay distintas opciones para la función de enlace (el POM usa `logit`):

model	inverse link function $P_\epsilon^{-1}(\Delta)$	density $dP_\epsilon(\eta)/d\eta$
logit	$\ln \frac{\Delta}{1-\Delta}$	$\frac{\exp(\eta)}{(1+\exp(\eta))^2}$
probit	$N^{-1}(\Delta)$	$\frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{\eta^2}{2}\right\}$
complementary log–log	$\ln(-\ln(1-\Delta))$	$\exp\{\eta - \exp(\eta)\}$

Table 7.1 Inverse link functions for different models for ordinal regression (taken from McCullagh and Nelder [1983]). Here, N^{-1} denotes the inverse normal function.

Proportional Odds Model (POM)

- 4 clases con 3 umbrales $b_1 = -4$, $b_2 = 0$ y $b_3 = 3$.
- Probabilidades acumuladas sobre la línea real (sigmoid$\{f(\mathbf{x})\}$) y transformadas en probabilidades individuales.



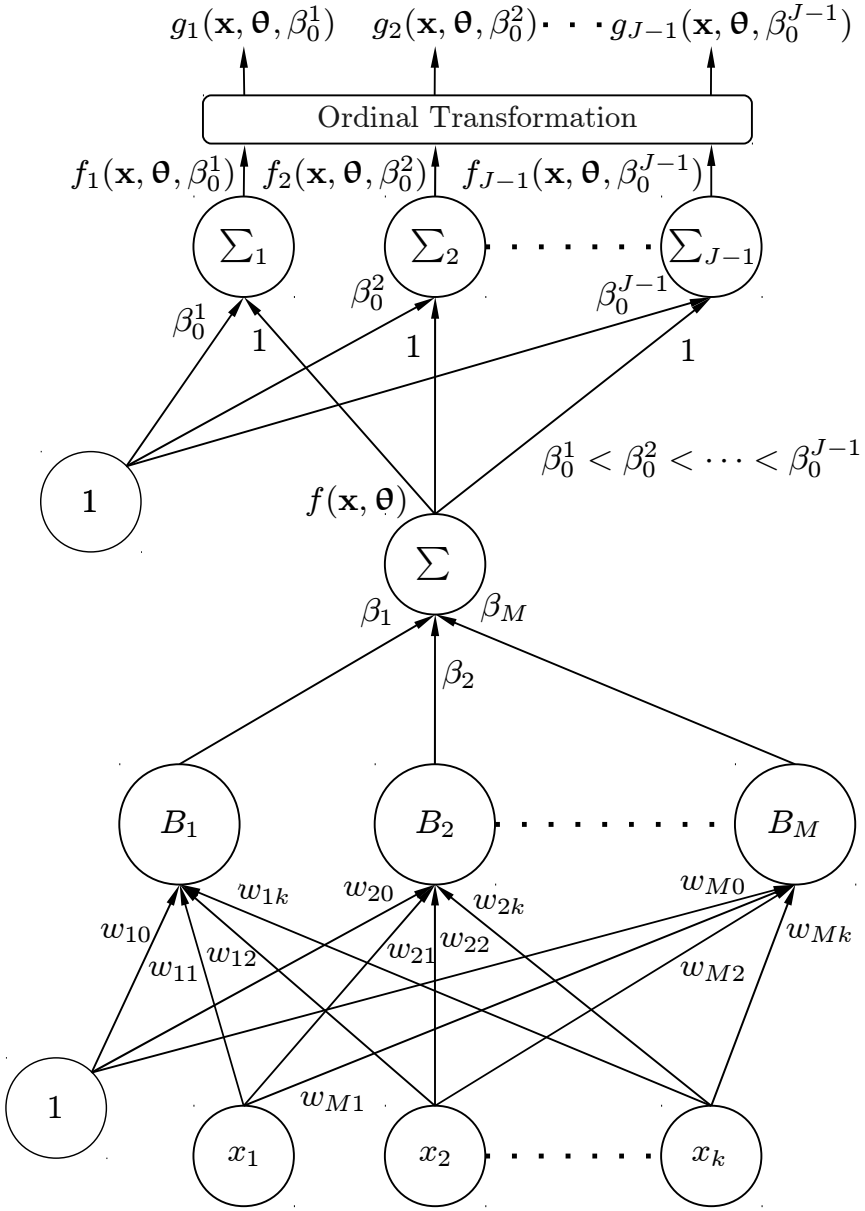
NNPOM

- Una forma directa de abordar estos problemas con clasificación ordinal es aproximar $f(\mathbf{x})$ de los modelos umbral utilizando un modelo de funciones de base, es decir, una suma ponderada de transformaciones no lineales de las variables de entrada:

$$f(\mathbf{x}, \theta) = \sum_{i=1}^{M} \beta_i B_i(\mathbf{x}, \mathbf{w}_i) \tag{11}$$

- Después, se pueden aplicar las ecuaciones del POM para obtener las probabilidades y gradiente descendente para ajustar los parámetros.

NNPOM


$g_1(\mathbf{x}, \boldsymbol{\theta}, \beta_0^1)$
$g_2(\mathbf{x}, \boldsymbol{\theta}, \beta_0^2) \cdots g_{J-1}(\mathbf{x}, \boldsymbol{\theta}, \beta_0^{J-1})$
Ordinal Transformation
$f_1(\mathbf{x}, \boldsymbol{\theta}, \beta_0^1)$
$f_2(\mathbf{x}, \boldsymbol{\theta}, \beta_0^2)$
$f_{J-1}(\mathbf{x}, \boldsymbol{\theta}, \beta_0^{J-1})$
$\sum_1$
$\sum_2$
$\sum_{J-1}$
β_0^1
β_0^2
β_0^{J-1}
1
1
1
1
$\beta_0^1 < \beta_0^2 < \cdots < \beta_0^{J-1}$
1
$f(\mathbf{x}, \boldsymbol{\theta})$
$\sum$
β_1
β_M
β_2
B_1
B_2
B_M
w_{1k}
w_{20}
w_{2k}
w_{M0}
w_{10}
w_{11}
w_{12}
w_{21}
w_{22}
w_{Mk}
w_{M2}
1
w_{M1}
x_1
x_2
x_k

NNPOM

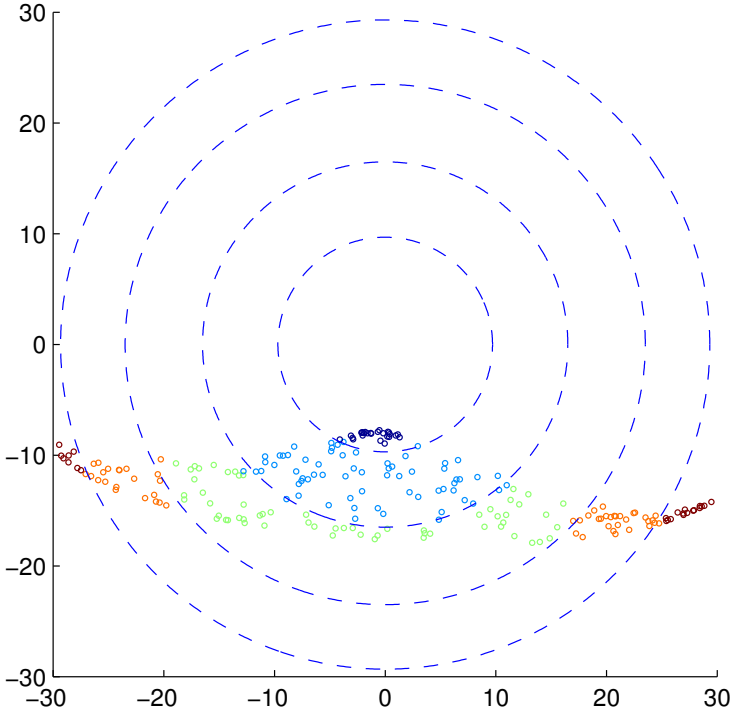
- Tenemos un problema con los umbrales (que deben estar ordenados):

$$b_1 < b_2 < \cdots < b_{J-1} \tag{12}$$

- Solución: definir los umbrales del 2 al $J-1$ como el primero más una cantidad siempre positiva:

$$\{b_1, b_1 + \alpha_1^2, b_1 + \alpha_1^2 + \alpha_2^2 ...\} \tag{13}$$

Red de hiperesferas concéntricas



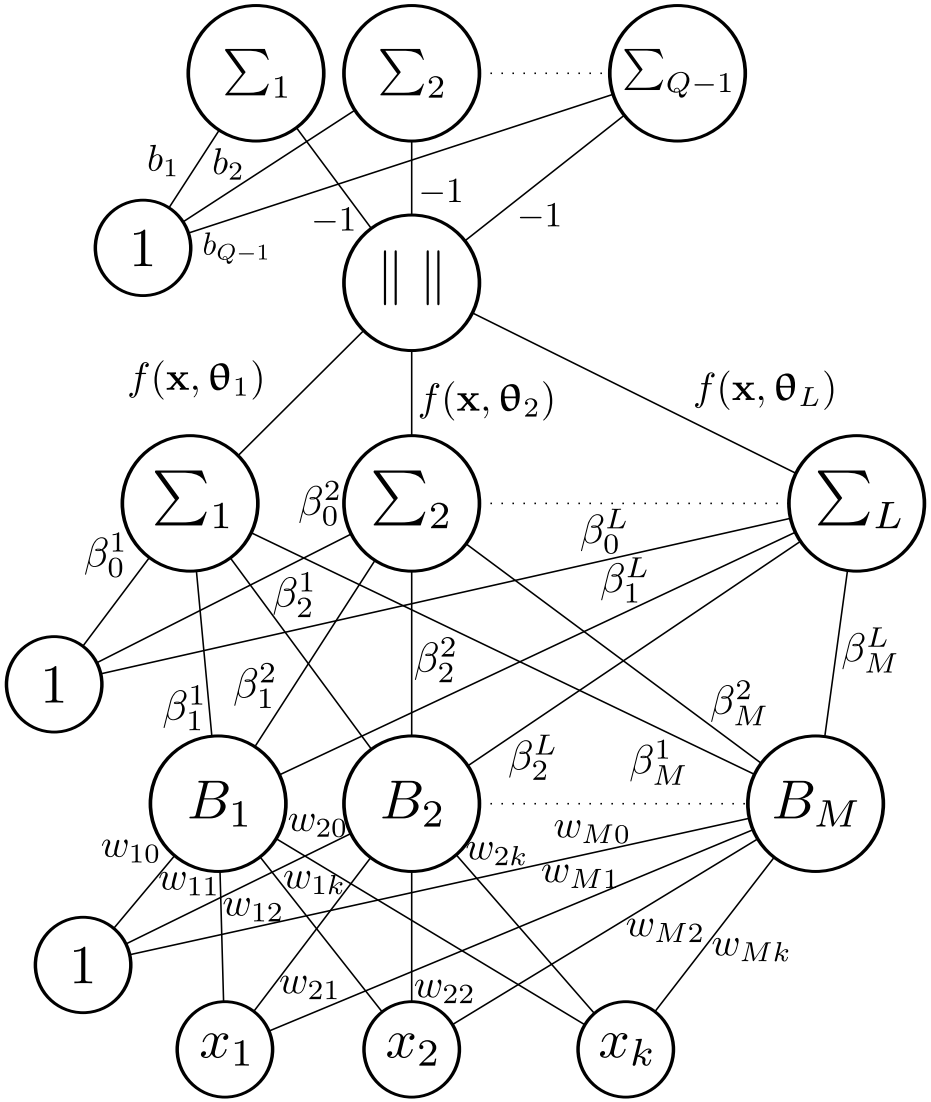
- Parece un poco drástico intentar proyectar todo a una línea recta y que los ejemplos queden correctamente ordenados.
- Una posible relajación es proyectar a un espacio de L dimensiones y luego definir un conjunto de hiperesferas concéntricas que separen las distintas clases.

Gutiérrez, P. A., Tino, P., & Hervás-Martínez, C. (2014). Ordinal regression neural networks based on concentric hyperspheres. Neural Networks, 59, 51

Red de hiperesferas concéntricas

- La idea sería similar a la anterior, pero añadimos varios modelos de funciones de base, uno por cada dimensión (L).
- Cuando llegue un nuevo punto, obtenemos la distancia desde el origen al punto (norma del vector) y aplicamos la transformación ordinal (POM) sobre esa distancia.
- Los umbrales se convierten en el radio de cada una de las esferas.

Red de hiperesferas concéntricas



SVMs para clasificación ordinal

SVM

- Herbrich et al.: algoritmo de máximo margen, similar a las SVM.

  Herbrich, R., Graepel, T., Obermayer, K.: Large margin rank boundaries for ordinal regression. In: Advances in Large Margin Classifiers, pp. 115?132. MIT Press, Cambridge (2000)

- Shashua & Levin: dos algoritmos adicionales.

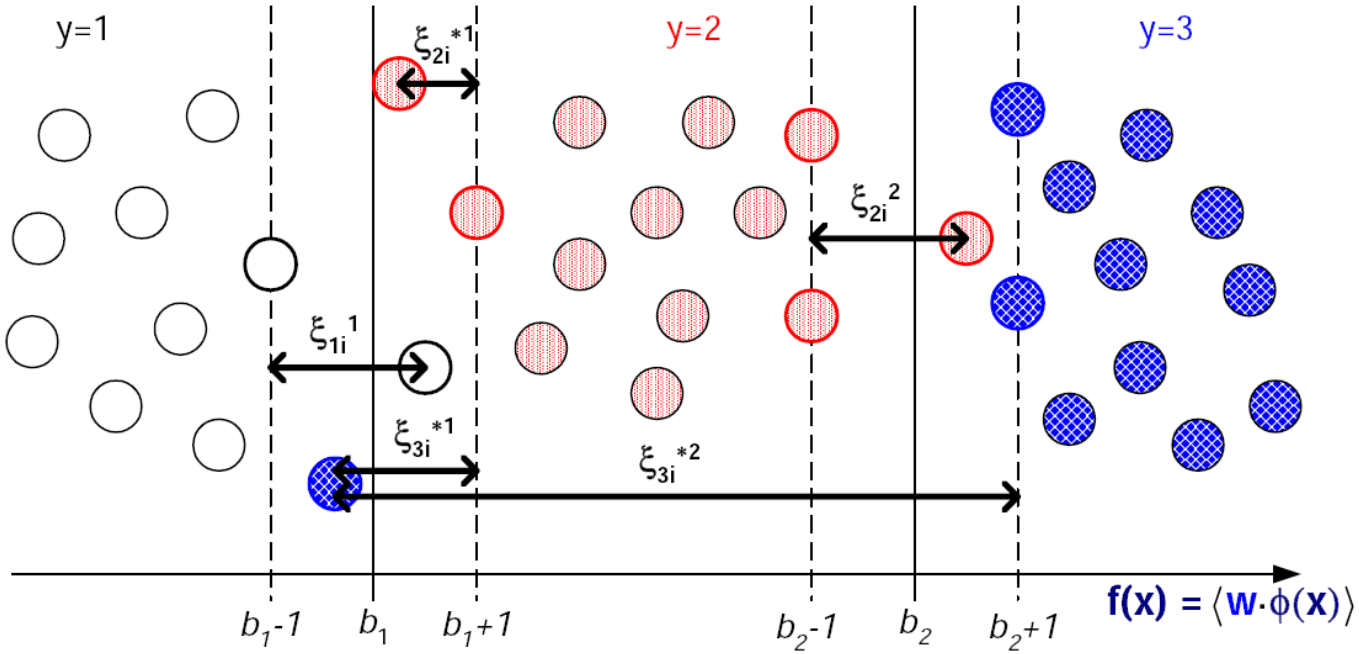
  Shashua, A., Levin, A.: Ranking with large margin principle: Two approaches. In: Advances in Neural Information Processing Systems, vol. 15, pp. 937-944. MIT Press, Cambridge (2003)

- Chu & Keerthi: corrección final más ampliamente extendida.

  Chu, W., Keerthi, S.S.: Support vector ordinal regression. Neural Computation 19(3), 792-815 (2007)

SVMs para clasificación ordinal

- Chu & Keerthi:
  - SVORIM considera que todos los ejemplos en todas las categorías contribuyen a los errores de cada umbral.
  - De esta forma, las desigualdades ordinales de los umbrales se cumplen automáticamente para el valor óptimo.

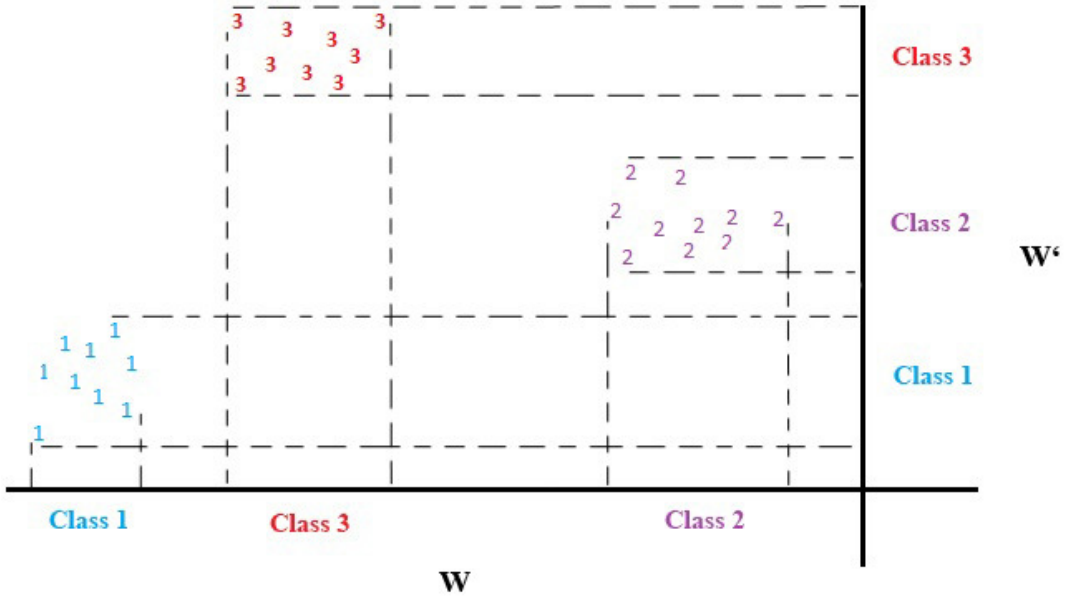


Métodos de análisis discriminante lineal para clasificación ordinal

- Análisis discriminante kernel:
  Bing-Yu Sun, Jiuyong Li, Desheng Dash Wu, Xiao-Ming Zhang, and Wen-Bo Li. "Kernel Discriminant Learning for Ordinal Regression", IEEE Transactions on Knowledge and Data Engineering, Vol. 22, N. 6, June 2010.

Análisis discriminante clas. ordinal

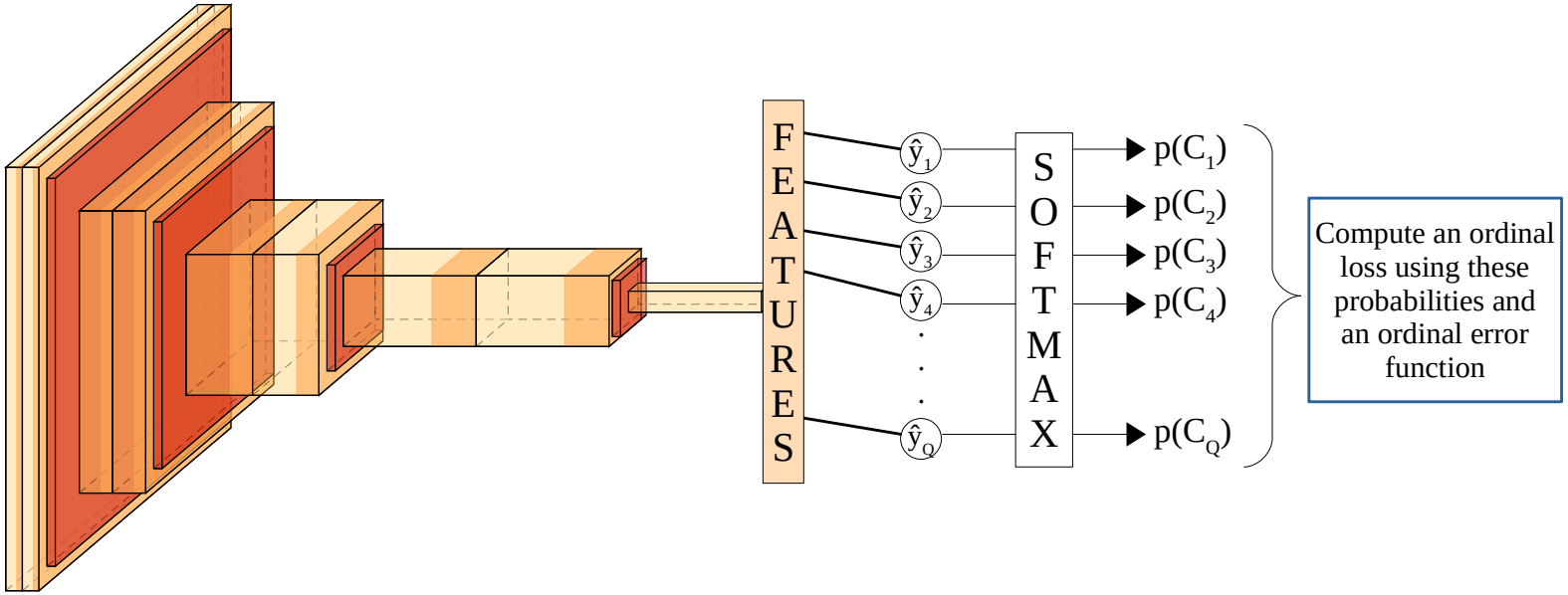
- El LDA se puede aplicar en clasificación ordinal si buscamos proyecciones consistentes.
- Minimizar la mínima distancia entre los patrones proyectados de dos clases consecutivas.
- Ejemplo: Dos proyecciones distintas: W and W'. ¿Cuál será mejor para un problema de clasificación ordinal?



Clasificación ordinal profunda

Formas de utilizar la ordinalidad en una CNN

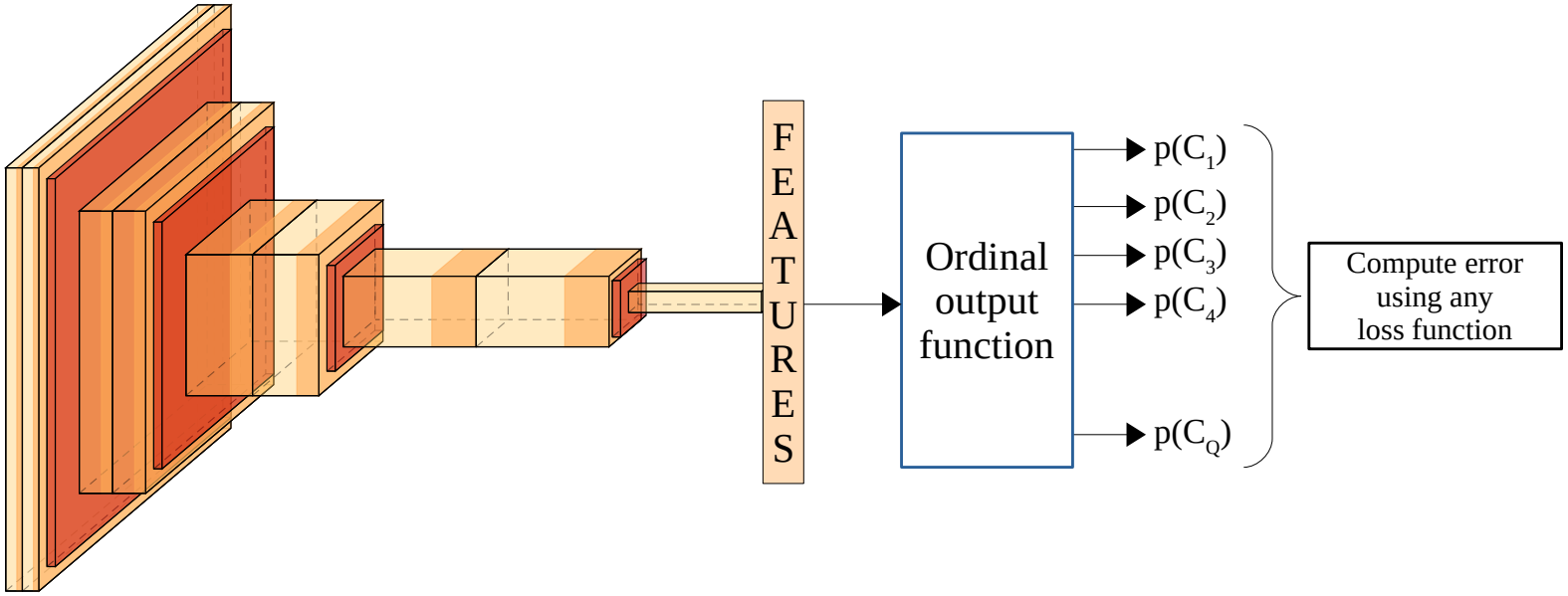
1. Utilizar una función de pérdida ordinal.



2. Utilizar una función de salida ordinal en lugar de la *softmax*.

Formas de utilizar la ordinalidad en una CNN

1. Utilizar una función de pérdida ordinal.
2. Utilizar una función de salida ordinal en lugar de la *softmax*.



Soft labeling unimodal

- Función de entropía cruzada:

$$L = \sum_{i=1}^{Q} q(i)[-\log p(y = C_i|x)]$$

- Para *one-hot*, la distribución es $q(i) = \delta_{i,q}$, donde q es la etiqueta observada y $\delta_{i,q}$ es la delta de Dirac.
- Podemos reemplazar $q(i)$ por una versión suavizada:

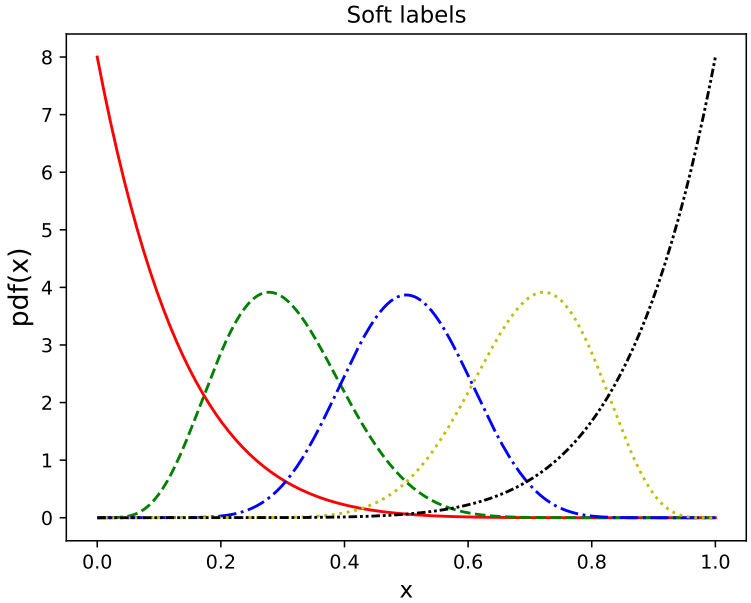
$$L = \sum_{i=1}^{Q} q'(i)[-\log p(y = C_i|x)],$$

donde $q'(i) = (1-\eta)\delta_{i,q} + \eta\frac{1}{Q}$, η es un hiperparámetro de control y hemos usado una distribución uniforme $(1/Q)$.

Soft labeling unimodal

Objetivo: reemplazar las etiquetas 0-1 con una versión "suave" que además sea ordinal.

1 0 0 0 0 0
0 1 0 0 0 0
0 0 1 0 0 0
0 0 0 1 0 0
0 0 0 0 1 0
0 0 0 0 0 1



Codificación estándar *one-hot* (izquierda) y etiquetas suaves unimodales (derecha)

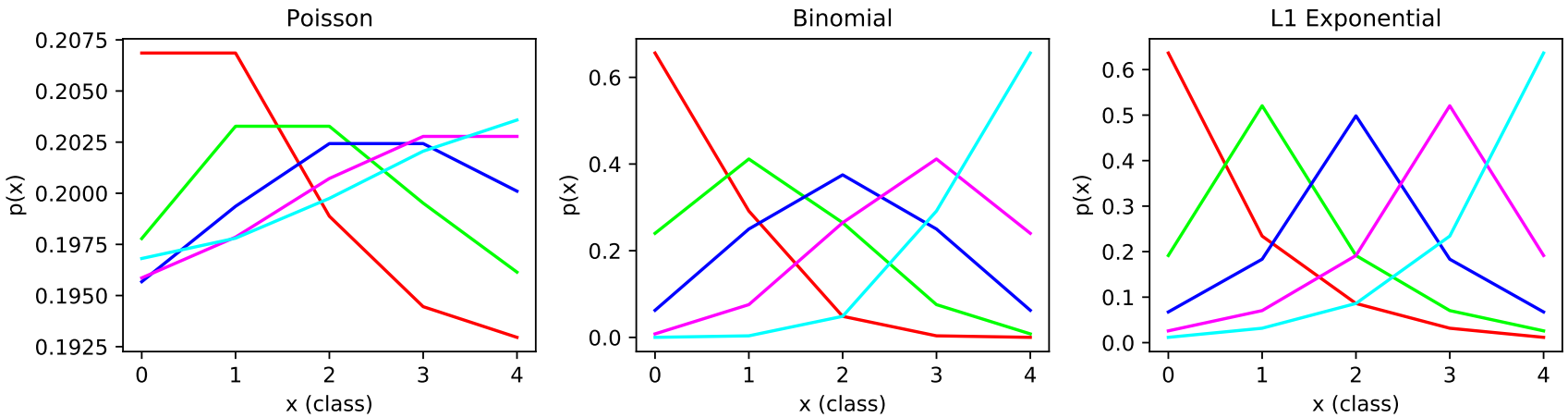
Distintas opciones de *soft labeling* unimodal

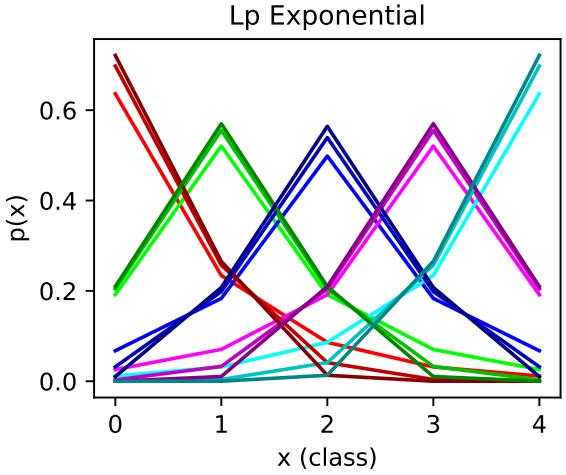
$$q'(i) = (1-\eta)\delta_{i,1} + \eta f(i,q)$$

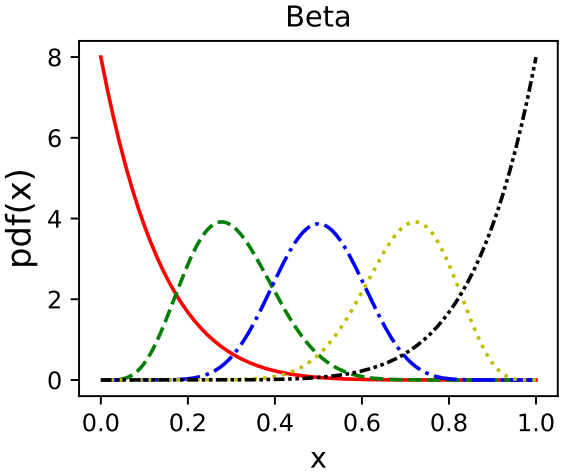
donde $f(i,q)$ define la distribución seleccionada cuando estamos evaluando la clase i y la etiqueta observada es q:

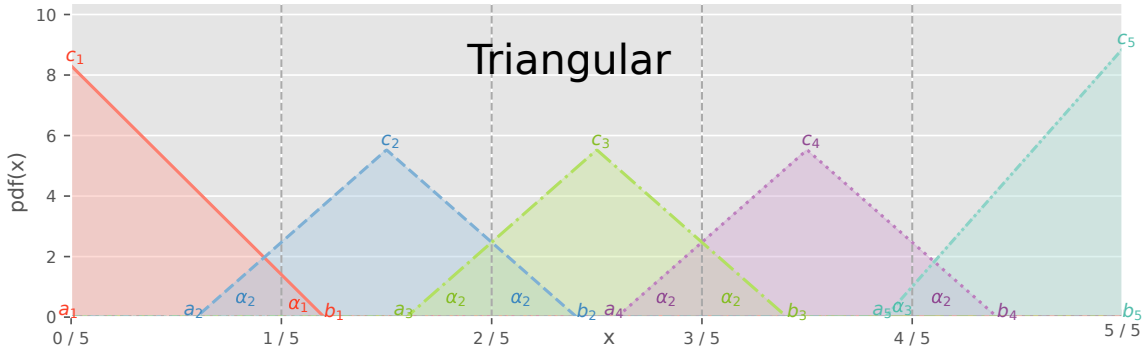
- Distribución de Poisson [Beckham and Pal, 2017].
- Distribución binomial [Pinto da Costa et al., 2008].
- Distribución exponencial [Liu et al., 2020].
- Propuestas:
  - Distrución Beta $\beta(a,b)$ [Vargas-Yun et al., 2022].
  - Exponencial generalizada [Vargas-Yun et al., 2023b].
  - Distribución triangular [Vargas-Yun et al., 2023a].

Comparación de distribuciones





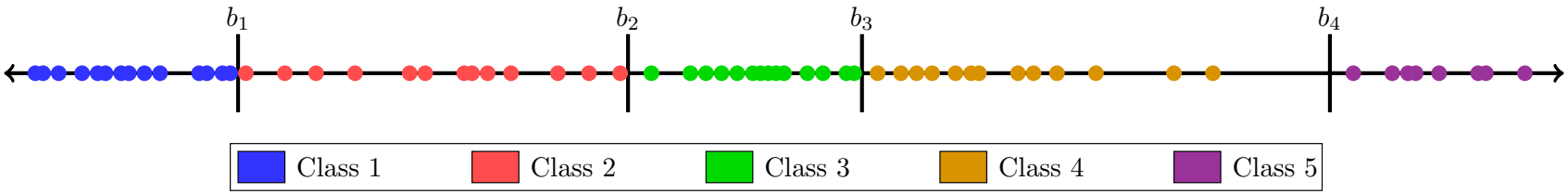




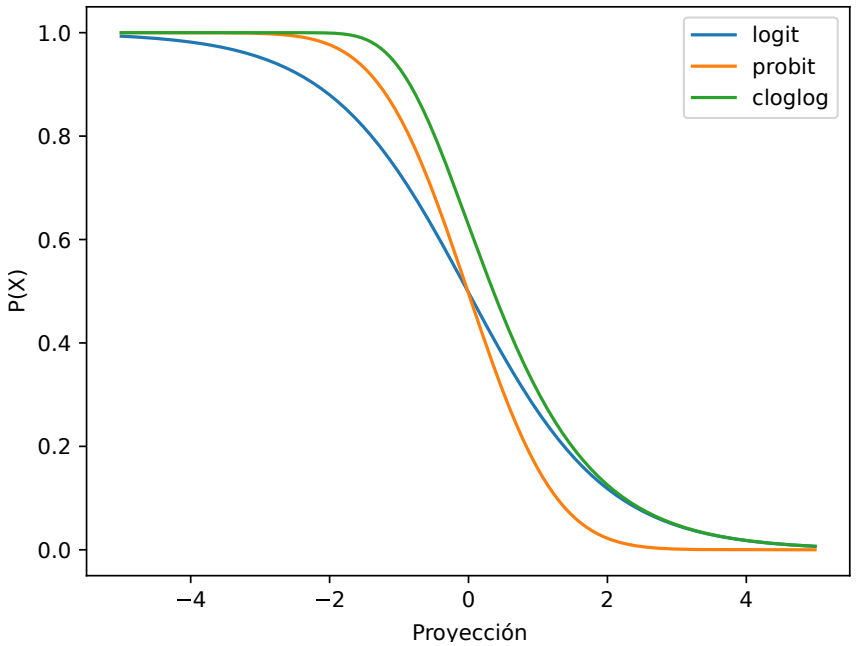
Cumulative Link Models [Vargas-Yun et al., 2020]

- Predicen probabilidades teniendo en cuenta la escala ordinal.
- $P(y \preceq \mathcal{C}_q|\mathbf{x}) = P(y = \mathcal{C}_1|\mathbf{x}) + ... + P(y = \mathcal{C}_q|\mathbf{x})$.
- Basados en una proyección 1D y un conjunto de umbrales **b**.
  - Deben ser no-decrecientes:

$$b_q = b_1 + \sum_{q=1}^{Q-1} \alpha_q^2, \quad n = 2, ..., Q,$$



Funciones de enlace



- Logit.

$$P(y \preceq \mathcal{C}_q|\mathbf{x}) = \frac{1}{1 + e^{-(b_q - f(\mathbf{x}))}}.$$

- Probit.

$$P(y \preceq \mathcal{C}_q|\mathbf{x}) = \int_{-\infty}^{b_q - f(\mathbf{x})} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}\mathbf{x}^2} \mathrm{d}x.$$

- Complementary log-log.

$$P(y \preceq \mathcal{C}_q|\mathbf{x}) = 1 - e^{-e^{b_q - f(\mathbf{x})}}.$$

- ...

Aplicaciones

Enfermedad de Parkinson

Característica neuropatológica de la enfermedad

- Descenso de dopamina en los núcleos basales (caudado y putamen).
- ^{123}I-ioflupano: se une a los transportadores.

Análisis de reducción de transportadores

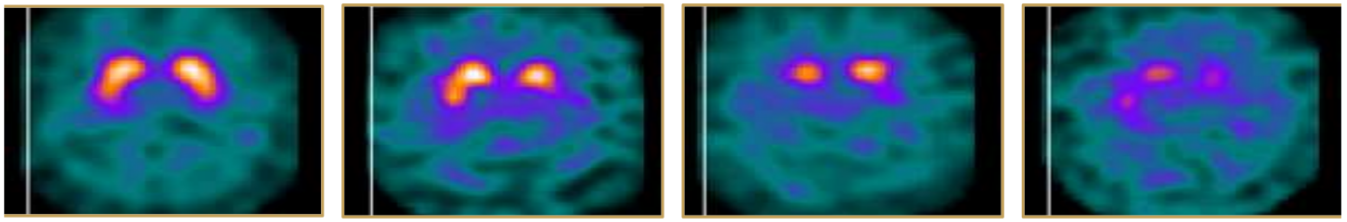
- Técnicas de cuantificación semiautomáticas.
- Diagnósticos visuales dependientes de los especialistas.

Literatura previa

- Clasificación binaria: padece o no padece.
- Uso de regiones de interés (caudado y putamen).

Enfermedad de Parkinson

508 imágenes funcionales 3D ($91 \times 109 \times 91$) (HURS).

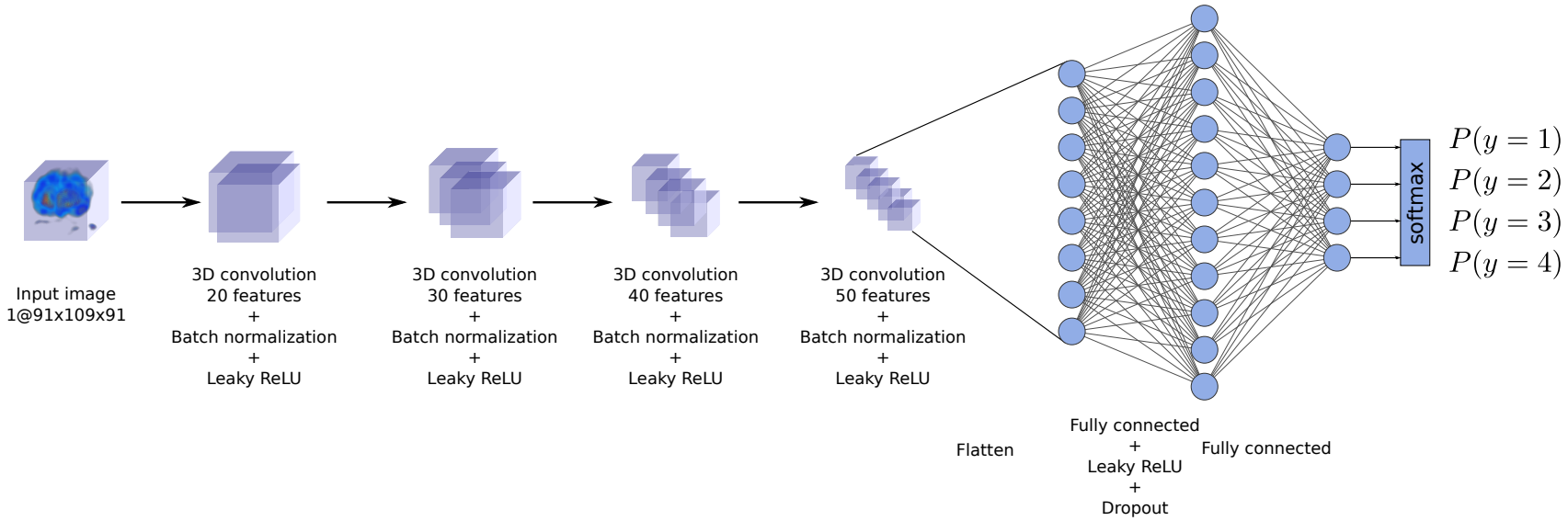


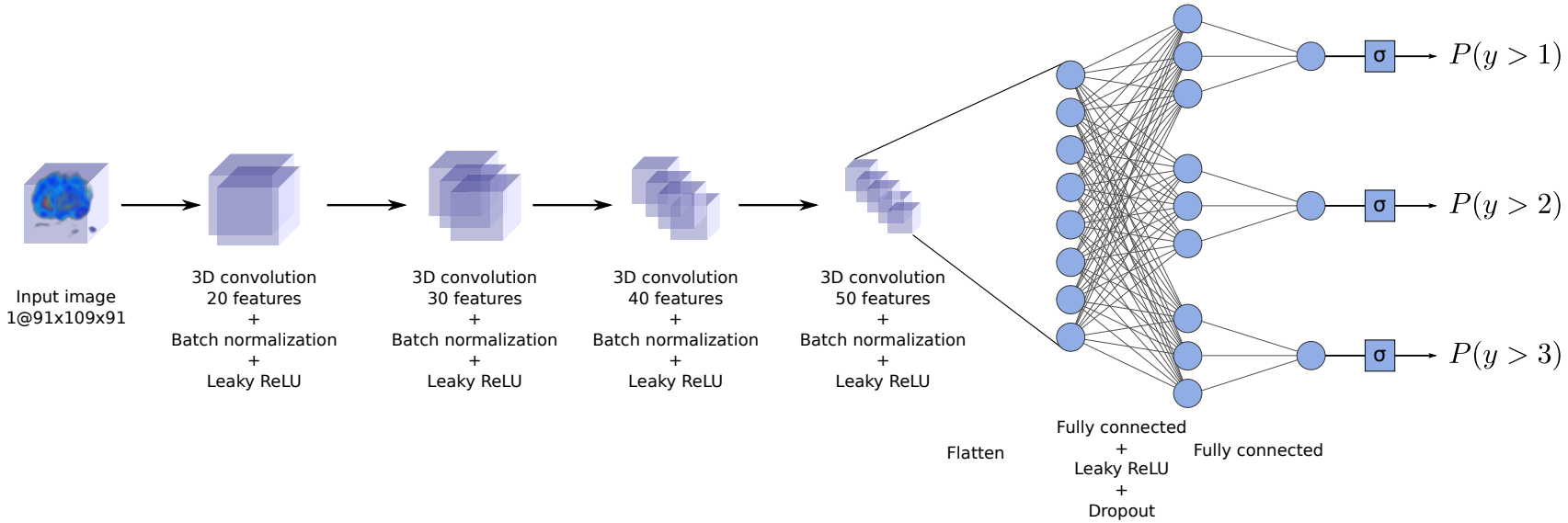
Grado 0 | Grado 1 | Grado 2 | Grado 3

Propuesta

- ¿Podemos distinguir más clases?
  1. Grado 0: sin enfermedad (314, 61.8 %).
  2. Grado 1: afectación leve (42, 8.3 %).
  3. Grado 2: afectación moderada (52, 10.2 %).
  4. Grado 3: afectación grave (100, 19.7 %).
- ¿Puede que otras áreas nos ayuden al diagnóstico?

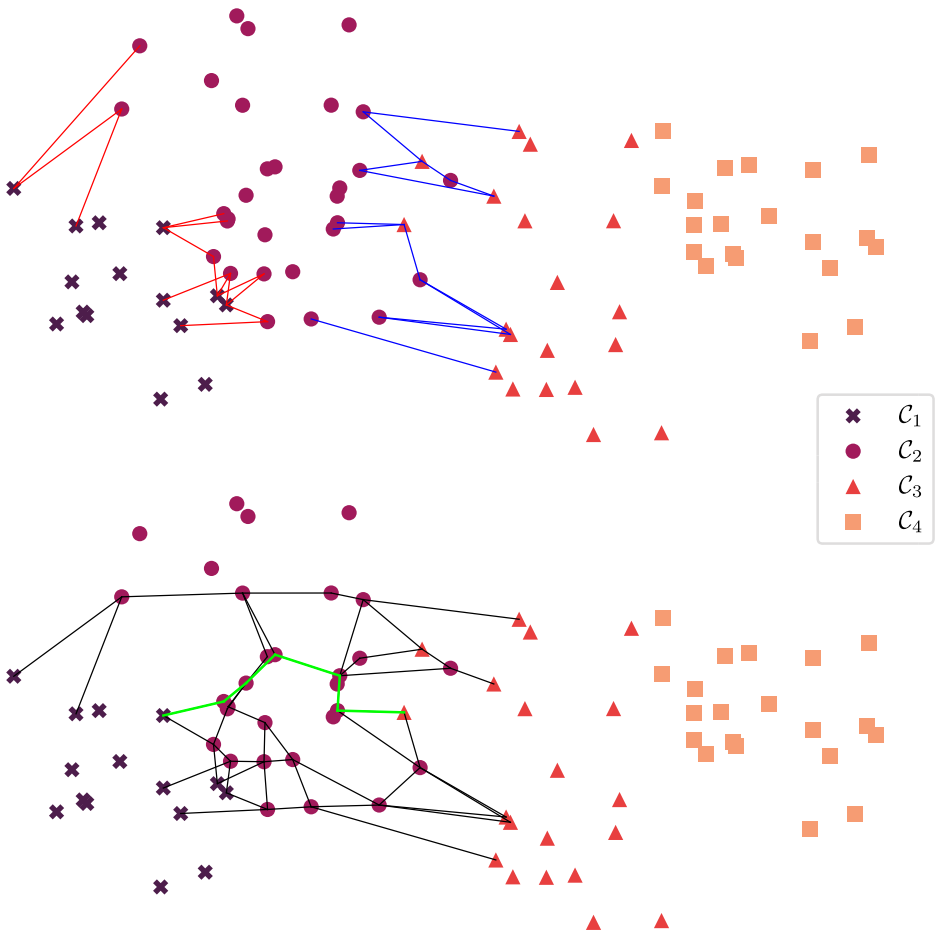
CNN 3D ordinal





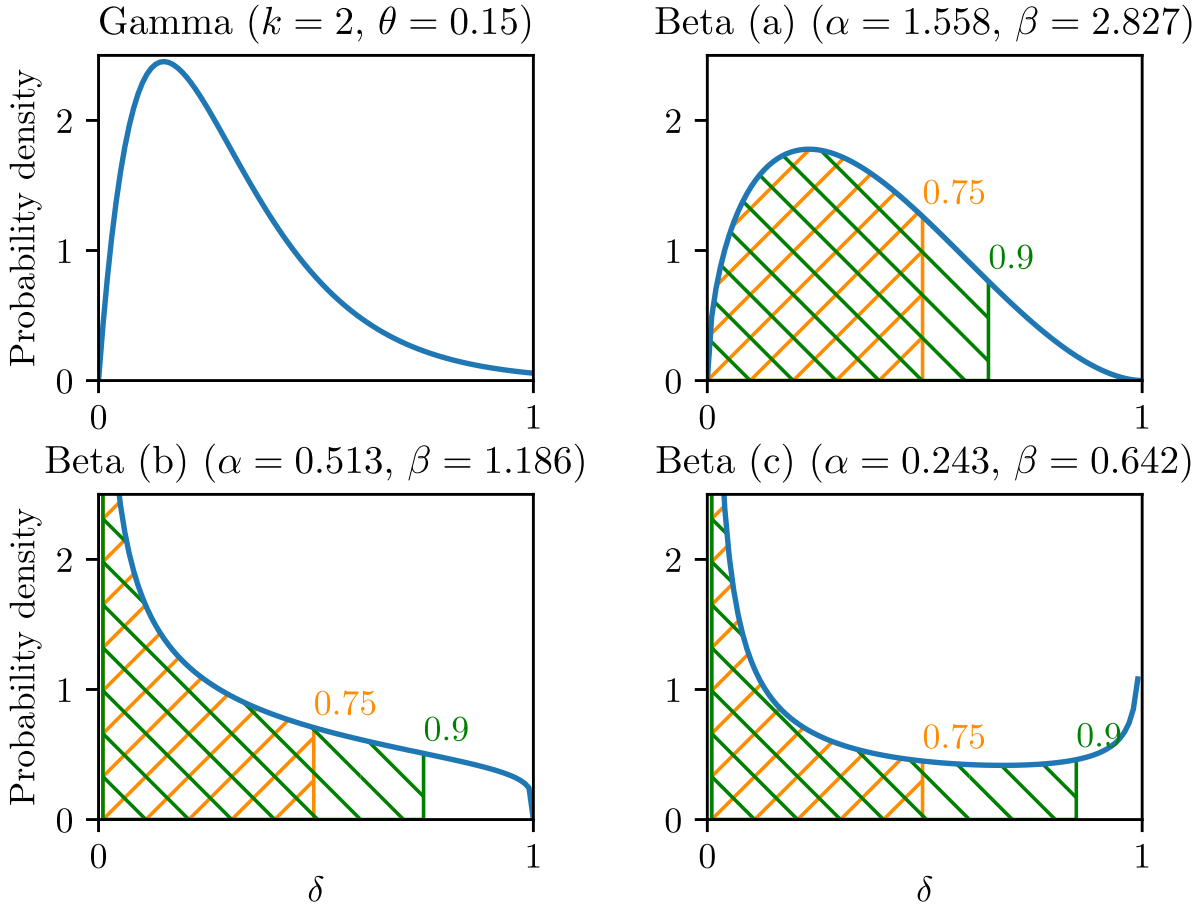
Aumento de datos

Construcción de grafo ordinal + remuestreo en las aristas

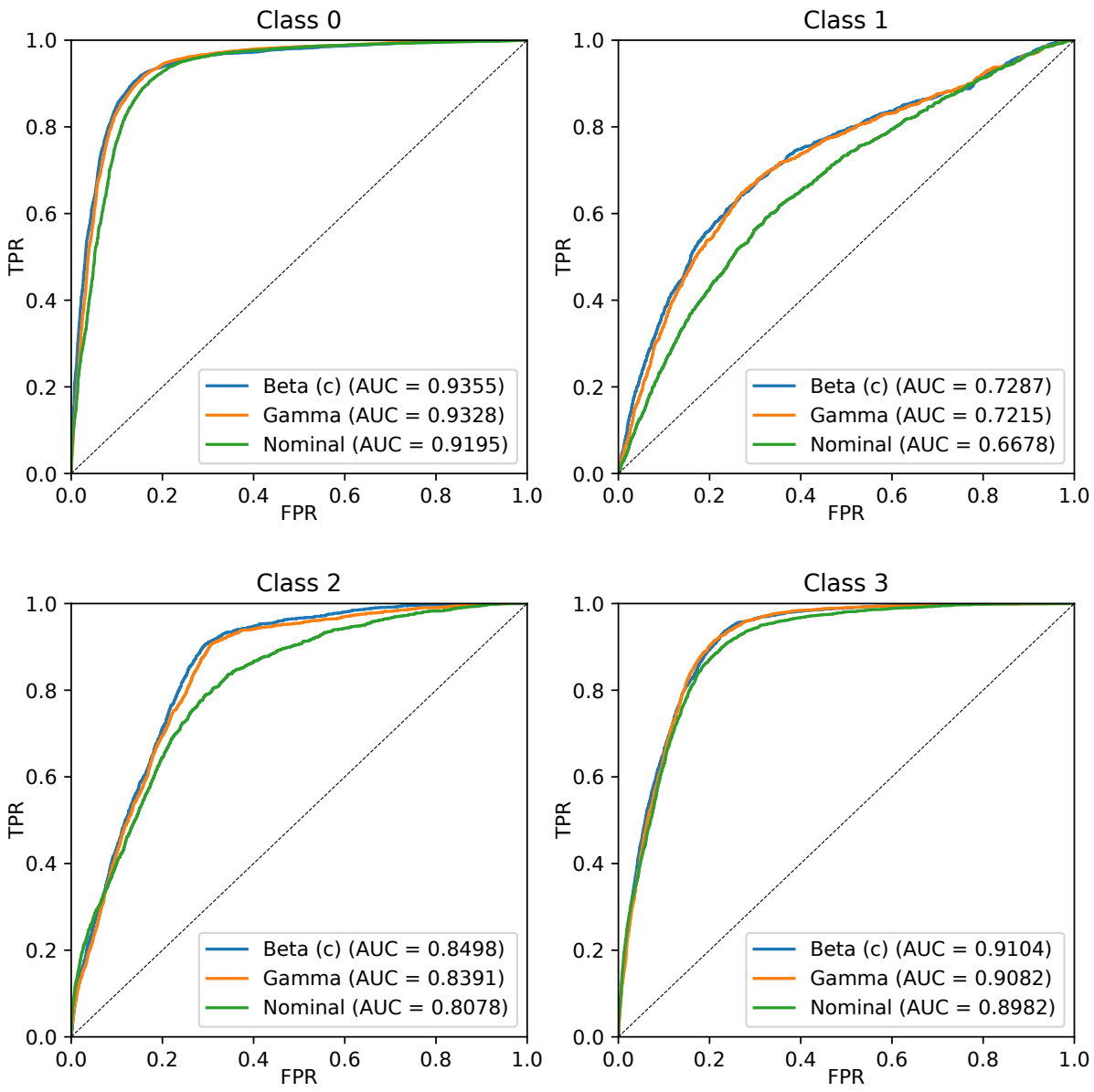


Algoritmo para aumento de datos

Distribuciones para el remuestreo en las aristas

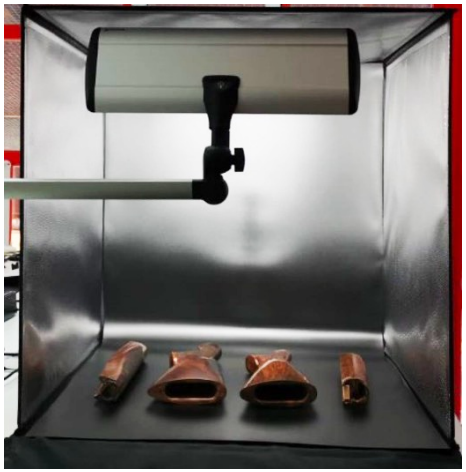


Resultados: curvas ROC por clase



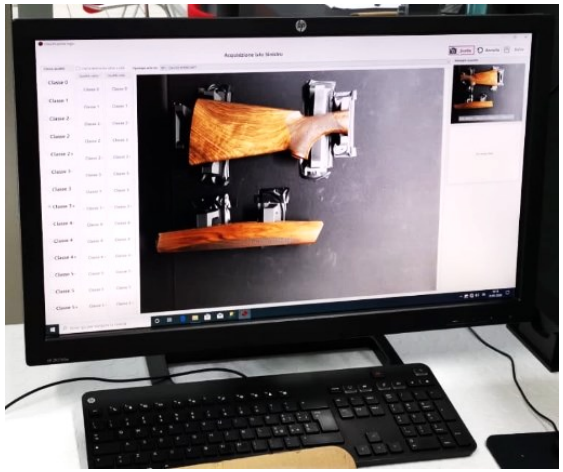
Una aplicación real en Industria 4.0

- Productor de armas conocido: Benelli.
- Evalúa la calidad de las culatas basándose en su estética.
- Necesita un sistema de apoyo a la decisión (DSS) que clasifique las imágenes en cuanto a su calidad estética.



a) Acquisition Box

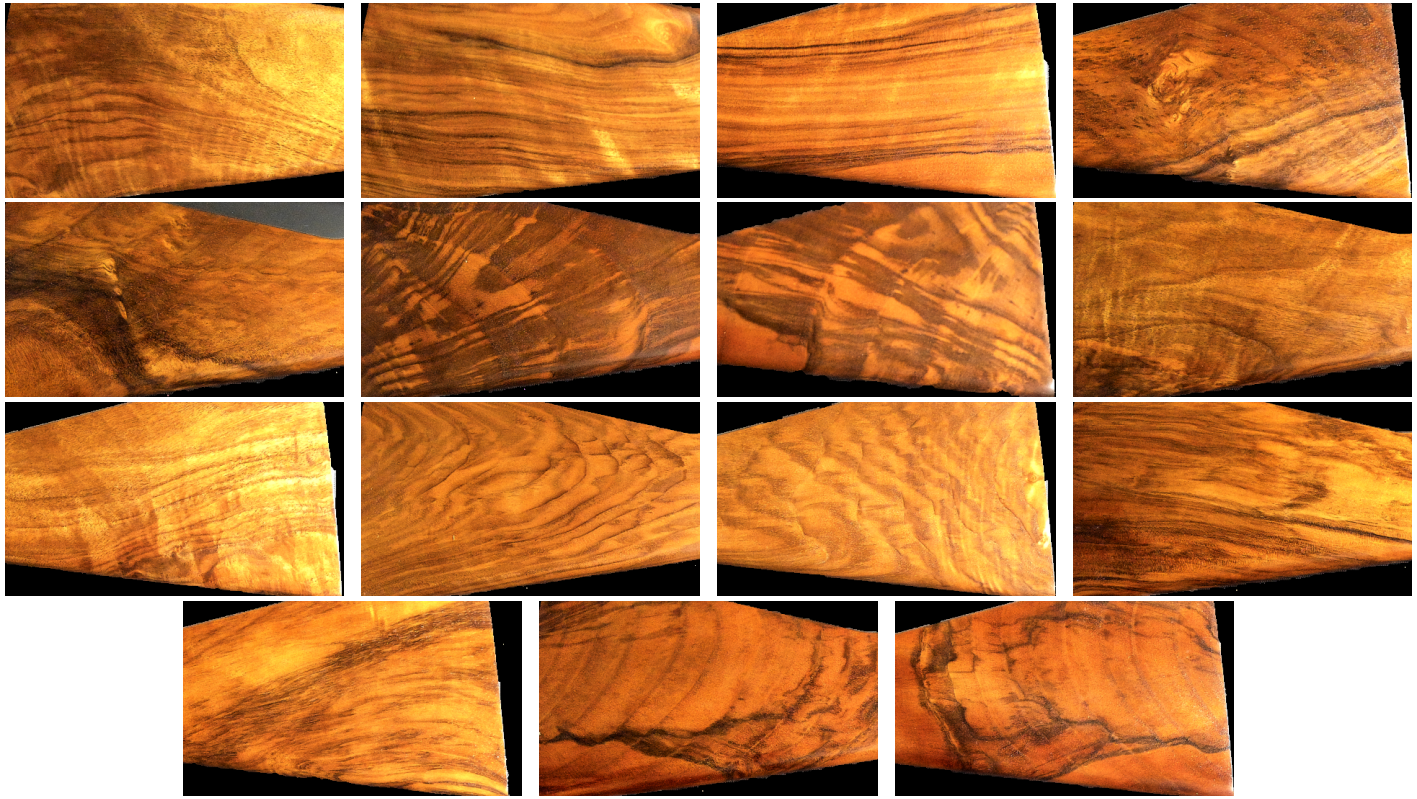
b) Rifle Placement



c) GUI Interface

Una aplicación real en Industria 4.0

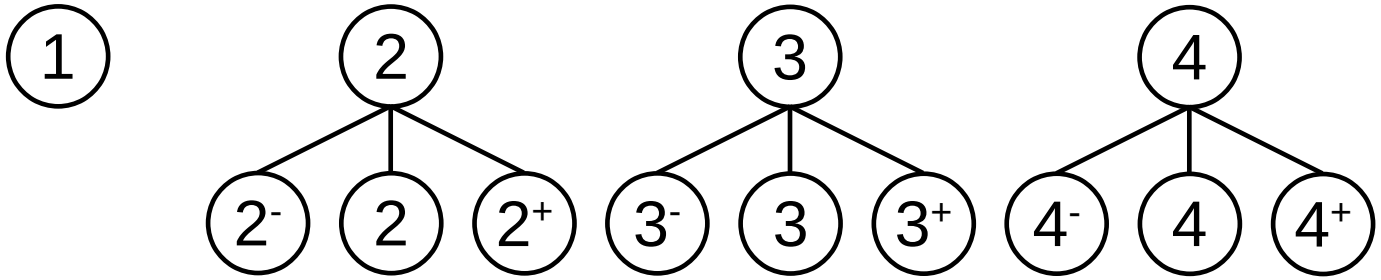
Obtuvieron 2120 imágenes de distintos rifles y las etiquetaron en función de su calidad estética.



Una aplicación real en Industria 4.0

Crearon 4 categorías principales: 1, 2, 3, 4; y 3 subcategorías para las categorías 2, 3 y 4.

Label	1	2^-	2	2^+	3^-	3	3^+	4^-	4	4^+
Images	165	148	212	177	179	306	344	208	275	106



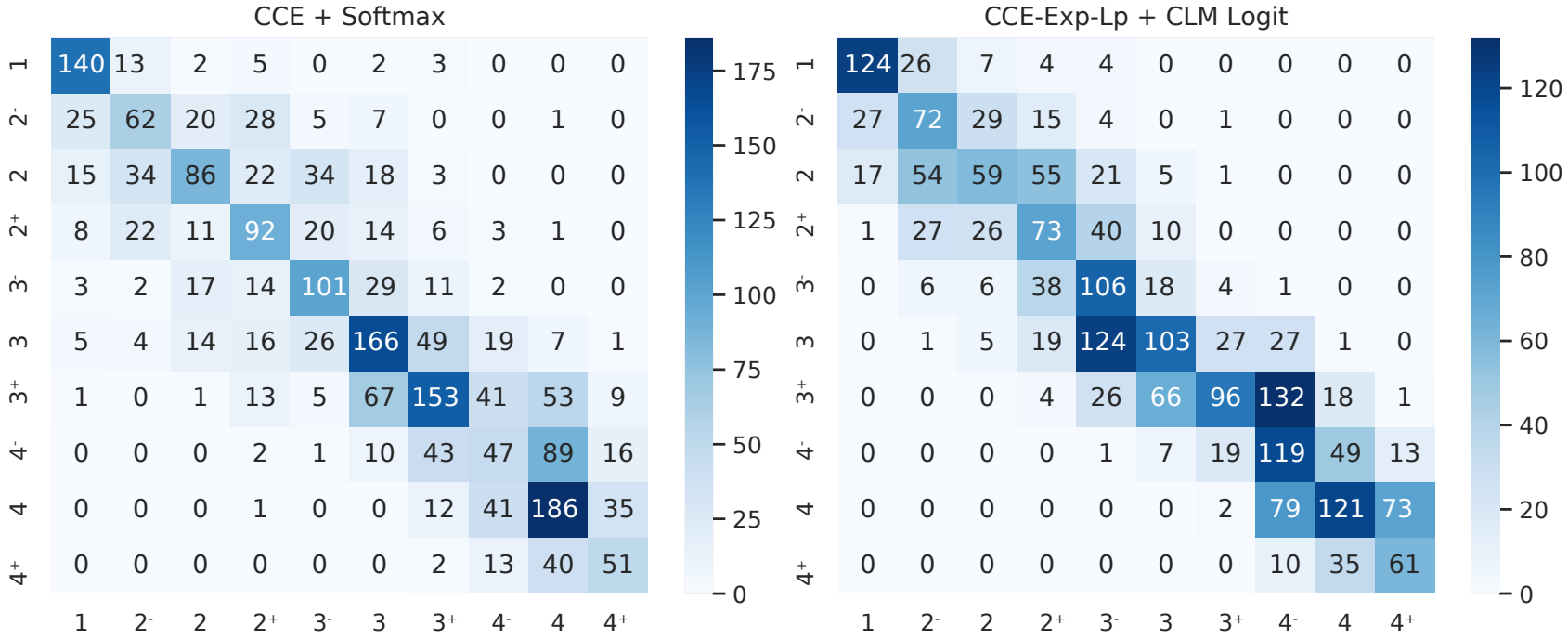
Benelli DSS [Vargas-Yun et al., 2023b]

Para abordar el problema utilizamos:

- Cumulative Link Model para la salida con distintas funciones de enlace.
- *Soft labelling* con distintas distribuciones:
  - Proponemos el uso de la exponencial generalizada (norma p).

Método	Ranking
CCE + Softmax (Baseline)	5,43
CCE-Exp-L_p + CLM CLogLog	8,20
CCE-Exp-L_p + CLM Logit	10,87
CCE-Exp-L_p + CLM Probit	**10,90**
CCE-Exp-L_o + CLM CLogLog	5,77
CCE-Exp-L_o + CLM Logit	10,17
CCE-Exp-L_o + CLM Probit	8,73
CCE-Poisson + CLM Probit	2,00
CCE-Poisson + CLM Logit	1,97
CCE-Poisson + CLM CLogLog	2,07
CCE-Binomial + CLM Probit	9,70
CCE-Binomial + CLM Logit	10,13
CCE-Binomial + CLM CLogLog	5,07

Benelli DSS [Vargas-Yun et al., 2023b]

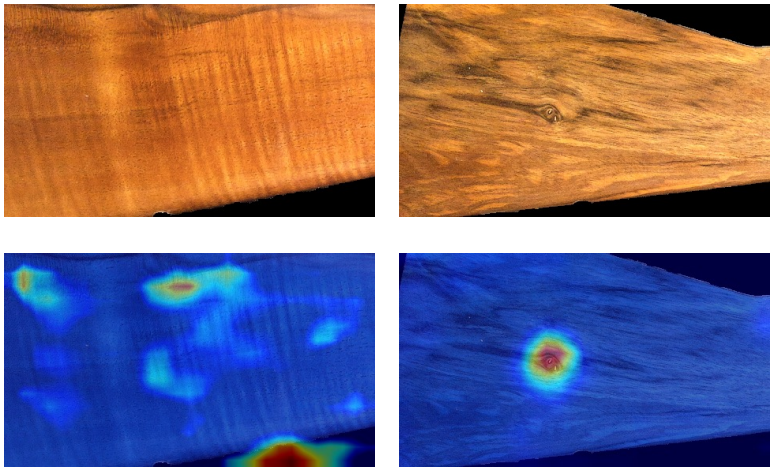


Benelli DSS [Rosati et al., 2022]

Mapas GradCam

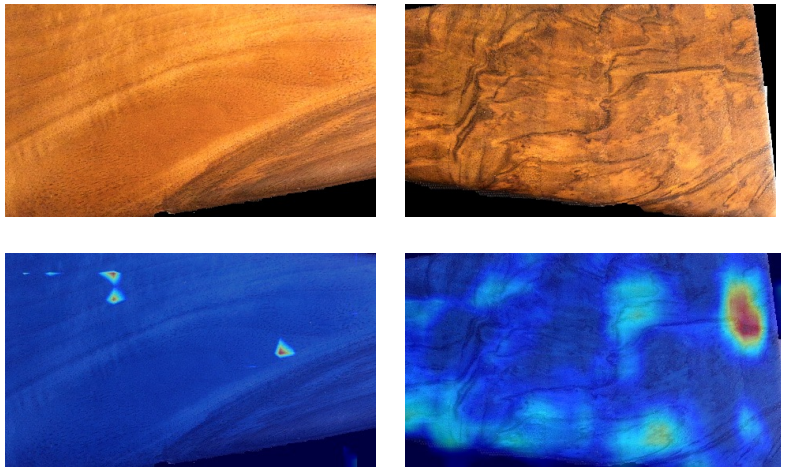
Nominal approach

Class 1 | Class 3-



Proposed approach

Class 1 | Class 3+



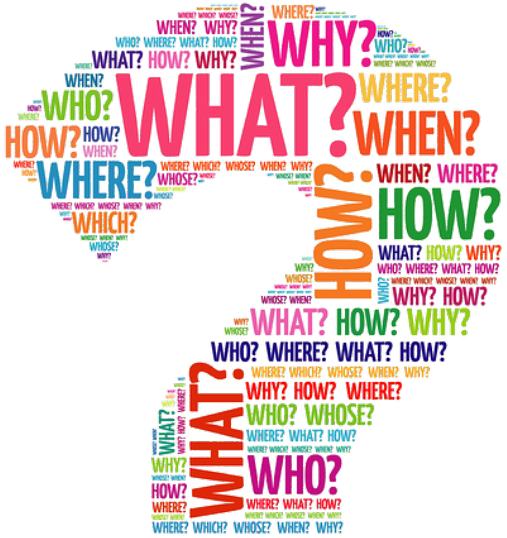
Conclusiones

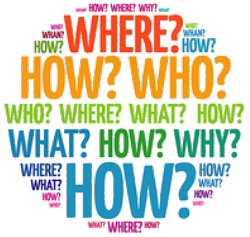
Clasificación ordinal: oportunidades y retos

- Oportunidades:
  - Campo relativamente nuevo, con un conjunto considerable de algoritmos, pero muchas ideas por adaptar.
  - Clasificación ordinal de series temporales.
  - Muchos problemas reales en los que se aplican, erróneamente, clasificadores nominales.
- Retos:
  - Usar métricas adecuadas para el problema.
  - Necesidad de evaluar la ordinalidad de un conjunto de datos.
  - Métricas de evaluación que no presupongan una distancia entre clases.
  - Desequilibrio provocado por clases extremas $\Rightarrow$ Remuestreo ordinal.

¿Preguntas?

¡Gracias!





Referencias

Referencias i

Agresti, A. (2010).
Analysis of ordinal categorical data.
Wiley Series in Probability and Statistics. Wiley.

Beckham, C. and Pal, C. (2017).
Unimodal probability distributions for deep ordinal classification.
In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning (PMLR)*, volume 70, pages 411–419. JMLR.org.

Chu, W. and Keerthi, S. S. (2007).
Support Vector Ordinal Regression.
Neural Computation, 19(3):792–815.

Referencias ii

Harrington, E. F. (2003).
Online ranking/collaborative filtering using the perceptron algorithm.
In *Proceedings of the Twentieth International Conference on Machine Learning (ICML2003)*.

Herbrich, R., Graepel, T., and Obermayer, K. (2000).
Large margin rank boundaries for ordinal regression.
In Smola, A., Bartlett, P., Schölkopf, B., and Schuurmans, D., editors, *Advances in Large Margin Classifiers*, pages 115–132, Cambridge, MA. MIT Press.

Referencias iii

Kotsiantis, S. B. and Pintelas, P. E. (2004).
A cost sensitive technique for ordinal classification problems.
In *Methods and applications of artificial intelligence (Proc. of the 3rd Hellenic Conference on Artificial Intelligence, SETN)*, volume 3025 of *Lecture Notes in Artificial Intelligence*, pages 220–229.

Kramer, S., Widmer, G., Pfahringer, B., and Groeve, M. D. (2010).
Prediction of ordinal classes using regression trees.
In Ras, Z. and Ohsuga, S., editors, *Proceedings of the 12th International Symposium, ISMIS 2000*, volume 1932 of *Lecture*

Notes in Computer Science, Foundations of Intelligent Systems, pages 665–674. Springer Berlin / Heidelberg, Charlotte, NC, USA.

Liu, X., Fan, F., Kong, L., Diao, Z., Xie, W., Lu, J., and You, J. (2020).
Unimodal regularized neuron stick-breaking for ordinal classification.
Neurocomputing, 388:34–44.

McCullagh, P. and Nelder, J. A. (1989).
***Generalized Linear Models*.**
Monographs on Statistics and Applied Probability. Chapman & Hall/CRC, 2nd edition.

Referencias v

Pinto da Costa, J. F., Alonso, H., and Cardoso, J. S. (2008).
The unimodal model for the classification of ordinal data.

Neural Networks, 21:78–91.

Rosati, R., Romeo, L., Vargas-Yun, V. M., Gutiérrez, P. A., Hervás-Martínez, C., and Frontoni, E. (2022).
A novel deep ordinal classification approach for aesthetic quality control classification.
Neural Computing and Applications, 34:11625–11639.
JCR (2021): 5.102 Position: 45/144 (Q2) Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE.

Referencias vi

Sánchez-Monedero, J., Gutiérrez, P. A., Tino, P., and Hervás-Martínez, C. (2013).
Exploitation of Pairwise Class Distances for Ordinal Classification.
Neural Computation, 25(9):2450–2485.
JCR (2013): 1.694 (category COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE, position 43/121 Q2).

Shashua, A. and Levin, A. (2003).
Ranking with large margin principle: two approaches.
In *Proceedings of the Seventeenth Annual Conference on Neural Information Processing Systems (NIPS2003)*, number 16 in Advances in Neural Information Processing Systems, pages 937–944. MIT Press.

Referencias vii

Sun, B.-Y., Li, J., Wu, D. D., Zhang, X.-M., and Li, W.-B. (2010).
Kernel discriminant learning for ordinal regression.
IEEE Transactions on Knowledge and Data Engineering, 22(6):906–910.

Vargas-Yun, V. M., Gutiérrez, P. A., Barbero-Gómez, J., and Hervás-Martínez, C. (2023a).
Soft labelling based on triangular distributions for ordinal classification.
Information Fusion, 93:258–267.
JCR(2021): 17.564 Position: 1/110 (Q1) Category: COMPUTER SCIENCE, THEORY & METHODS.

Referencias viii

Vargas-Yun, V. M., Gutiérrez, P. A., and Hervás-Martínez, C. (2020).
Cumulative link models for deep ordinal classification.
Neurocomputing, 401:48–58.
JCR(2020): 5.719 Position: 30/140 (Q1) Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE.

Vargas-Yun, V. M., Gutiérrez, P. A., and Hervás-Martínez, C. (2022).
Unimodal regularisation based on beta distribution for deep ordinal regression.
Pattern Recognition, 122:108310.
JCR(2021): 8.518 Position: 22/144 (Q1) Category: COMPUTER SCIENCE, ARTIFICIAL INTELLIGENCE.

Referencias ix

Vargas-Yun, V. M., Gutiérrez, P. A., Rosati, R., Romeo, L., Frontoni, E., and Hervás-Martínez, C. (2023b). **Exponential loss regularisation for encouraging ordinal constraint to shotgun stocks quality assessment.** *Applied Soft Computing*, page 110191.

Waegeman, W., De Baets, B., and Boullart, L. (2008). **Roc analysis in ordinal regression learning.** *Pattern Recognition Letters*, 29(1):1–9.