Sesión 6: Análisis no supervisado

ML-IA: Módulo de Machine learning

Jacob Muñoz-Castro

Universidad de los Andes

19 de noviembre de 2020

Agenda

1. Recap
2. Métricas de evaluación
3. Medidas de similitud: distancias
4. Análisis de clústeres

Recap: Árboles

Métricas de evaluación

- Para regresión teníamos como medida MSE.
- En nuestro caso discreto, tenemos entonces una tabla de frecuencia cruzada entre la variable observada y la proyectada.
- A esto le llamamos **Matriz de Confusión**.

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

- a es un verdadero positivo (TP)
- b es un falso negativo (FN)
- c es un falso positivo (FP)
- d es un verdadero negativo (TN)

Exactitud (accuracy)

		PREDICTED CLASS	
		Class=Yes	Class=No
ACTUAL CLASS	Class=Yes	a	b
	Class=No	c	d

- Definimos **Exactitud** de la predicción como el número de aciertos respecto al total predicho

$$Exactitud = \frac{a+d}{a+b+c+d}$$

Problemas con Exactitud

- Esta medida parece adecuada(?)
- Considere un problema de dos clases
- El número de casos de la clase 0 es 9990
- El número de casos de la clase 1 es 10
- Creemos un modelo ingenuo: todo va a clase 0!
- Entonces, la exactitud es 99.9 %.
- Ahora no resulta tan buena medida del poder predictivo de un modelo pensando en que naturalmente estos casos extremos son los que en la realidad son más costosos (liquidaciones, siniestros, identificación de cáncer,...).

Precisión

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

$$p = \frac{a}{a + c}$$

Sensibilidad o Recall

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

$$r = \frac{a}{a+b}$$

Puntaje F1

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

Usamos el promedio armónico entre recall y precisión

$$F1 = \frac{rp}{r+p} = \frac{2a}{2a+b+c}$$

TPR y FPR

	PREDICTED CLASS		
ACTUAL CLASS		Class=Yes	Class=No
	Class=Yes	a	b
	Class=No	c	d

Tasa de verdaderos positivos (TPR)

$$TPR = \frac{a}{a+b} = r$$

Tasa de falsos positivos (FPR)

$$FPR = \frac{c}{c+d}$$

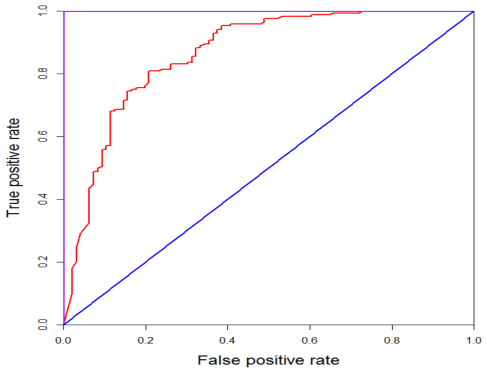
Umbrales

- Si tenemos probabilidades reportadas, podemos variar el umbral al que se deciden las clases
- Este umbral se vuelve un parámetro

Pid	Prediction	T=0.5	T=0.25	T=0.75
2	.95	Survived	Survived	Survived
3	.86	Survived	Survived	Survived
5	.02	Dead	Dead	Dead
7	.15	Dead	Dead	Dead
13	.48	Dead	Survived	Dead
14	.35	Dead	Survived	Dead
21	.12	Dead	Dead	Dead
24	.01	Dead	Dead	Dead
34	.74	Survived	Survived	Dead
54	.63	Survived	Survived	Dead

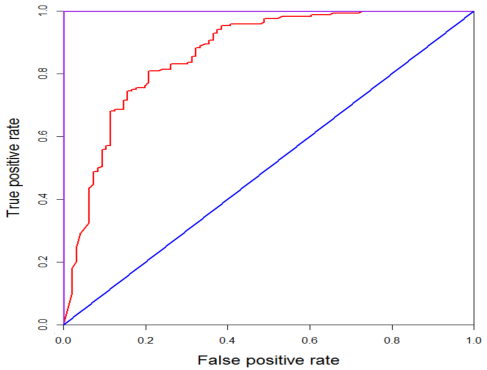
ROC (Reciever Operating Characteristic)

- Diseñada para analizar señales ruidosas
- Eje x: Predicción errónea para $\hat{y} = 0$
- Eje y: Predicción acertada para $\hat{y} = 1$



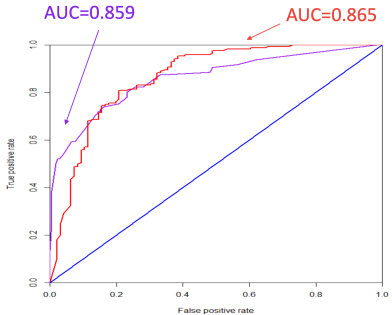
ROC (Reciever Operating Characteristic)

- La curva ideal está en morado: 100 % de aciertos y 0 % de error (falsos positivos)
- La curva azul es el peor de los casos (random chance)
- Entonces nuestra curva ROC representa la calidad del modelo



Área bajo la curva (AUC)

- En este caso, ningún modelo se desempeña mejor que el otro consistentemente
- El morado es mejor en umbrales bajos
- El rojo es mejor en umbrales altos
- Mientras más área haya bajo la curva, mejor es nuestro modelo...directamente es mejor el rojo



Medidas de similitud:

- Distancia de Minkowski

$$D(x, y) = \left(\sum_i |x_i - y_i|^p \right)^{1/p}$$

Deriva en:

- Distancia Euclindiana

$$d(x, y) = \sqrt{\sum_i (x_i - y_i)^2}$$

- Distancia Manhattan

$$M(x, y) = \sum_i |x_i - y_i|$$

Medidas de similitud: índices

- Similitud coseno

$$\cos(\theta) = \frac{x \cdot y}{||x||\ ||y||} \in [-1, 1]$$

- Índice Jaccard

$$J(x, y) = \frac{x \cdot y}{||x|| + ||y|| - x \cdot y} \in [0, 1]$$

Intuición gráfica

Recordemos tipos de aprendizaje

Aprendizaje supervisado

- Valores de clase conocidos
- Datos de entrenamiento marcados con valores de clase
- Meta: Encontrar una forma de mapear atributos a valores de clase
- Clasificación y regresión

Aprendizaje no supervisado

- Valores de clase desconocidos
- Datos de entrenamiento desmarcados
- Meta: descubrir información escondida en los datos.
- Agrupamiento (clustering)

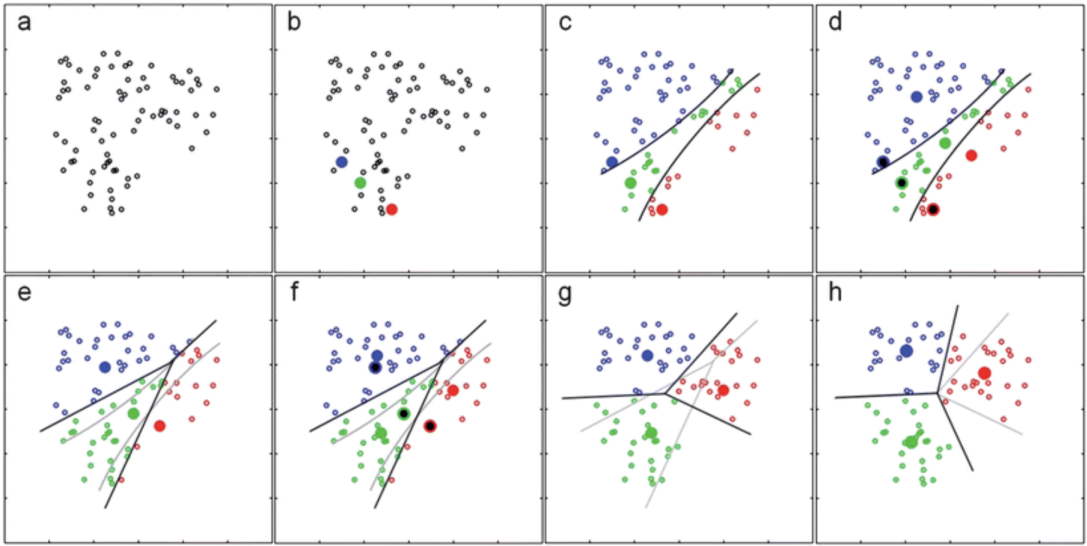
Clustering

- Esencialmente: grupos de elementos tales que dentro del cluster estos elementos son similares entre sí, pero muy diferentes de los elementos fuera del cluster.
- **Clasificación no supervisada**: cluster no es para asociar atributos a clases o variables latentes, sino para estimar membresía de distintos grupos
- Los grupos reciben etiquetas por interpretación post-estimación.
- Típicamente usado cuando no sabemos (o nunca sabremos) la "verdaderaçlase.
- Issues: cómo ponderar distancia resulta arbitrario

K-medias

- Esencia: asignar cada elemento a uno de los k clusters, donde la meta es minimizar las diferencias within-cluster y maximizar diferencias between-cluster.
- Usa posiciones iniciales aleatorias e itera hasta estabilizar
- Como con kNN, k-means clustering trata los valores de los atributos como coordenadas en un espacio multidimensional.
- Ventajas: simplicidad, flexibilidad, eficiente.
- Desventajas: nno hay una regla para determinar k, usa un elemento de aleatoriedad para valores iniciales.

K-medias



Agrupamientos jerarquizados

El algoritmo:

- Start by considering each item as its own cluster, for n clusters
- calculate the $N(N-1)/2$ pairwise distances between each of the n clusters, store in a matrix D_0
- Find smallest (off-diagonal) distance in D_0, and merge the items corresponding to the i, j indexes in D_0 into a new "cluster"
- Recalculate distance matrix D_1 with new cluster(s). Options for determining the location of a cluster include:
  - Centroids (mean)
  - Most dissimilar objects
  - Ward's measure(s) based on minimising variance
- Repeat 3–4 until a stopping condition is reached e.g. all items have been merged into a single cluster
- To plot the dendrograms, need decisions on ordering, since there are $2(N-1)$ possible orderings

Agrupamientos jerarquizados