UNIVERSIDAD ANDRÉS BELLO

FACULTAD DE INGENIERÍA

INGENIERÍA CIVIL INFORMÁTICA

Trabajo 4 – Aprendizaje Supervisado

GUSTAVO ANDRÉS VÉLIZ LÓPEZ

SANTIAGO-CHILE

MAYO, 2017

INTRODUCCIÓN

La tarea asignada al alumno tiene como objetivo el aprendizaje por parte del alumno de la construcción de un modelo de predicciones, para esto, el alumno debe re utilizar los datos obtenidos del etiquetado en la primera tarea.

Para la correcta realización de este problema, el aprendizaje supervisado, se tendrá que recurrir al uso de cuatro componentes. Estos son:

- Espacio Vectorial, el cual tendrá el formato:

<Etiqueta> <ID Palabra>:<Frecuencia Palabra>

- Clase de Modelo, el cual, para esta tarea será SVM MultiClass.

- Metodología Experimental, esta será Cross-Validation.

- Métrica, la cual para será MMR para medir el desempeño.

La finalidad del aprendizaje supervisado, como lo dice su nombre, es el aprendizaje en base a datos entregados (en este caso los de la tarea 1), de modo que a futuro se pueda lograr una predicción de resultados para casos similares.

DESCRIPCIÓN DEL PROBLEMA

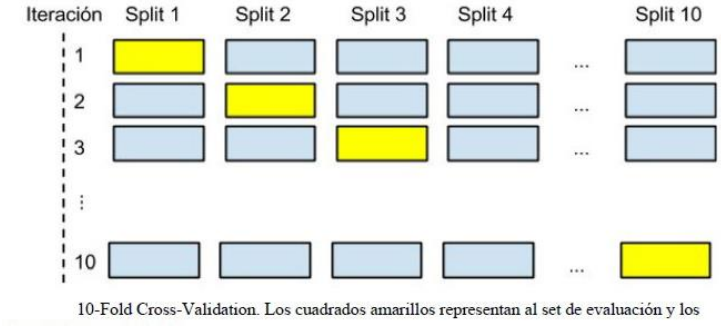
Como fue ya mencionado, el alumno deberá hacer uso de los cuatro componentes para la realización del Aprendizaje Supervisado. Para el caso del *Espacio Vectorial* se deben usar los 200 Perfiles que fueron obtenidos en la tarea 1, usando las preguntas y respuestas que estos daban, luego se debe seguir cierto formato para representar estar preguntas de modo que sean "legibles" para el programa. Para eso se recurre al uso de *números* a modo de identificadores, tanto como para la Clase (las distintas etiquetas disponibles en la tarea 1) y cada valor de ella, como para las palabras encontradas en la totalidad de los perfiles. Para esto se utiliza una *Bolsa de Palabras*, la cual se obtiene al asignar a cada una de las palabras existentes en los perfiles como un valor numérico, por ejemplo, 1=perro, 2=comida, 3=manuscrito, etc.

Luego de realizar esto, se debería obtener un archivo con todos los perfiles ordenados de la forma ya mencionada para cada clase, como en la siguiente figura.

```
10 36:2 80:2 110:25 113:11 115:16 116:28 119:23 12
2 4326:2 4435:1 4437:4 4439:3 4507:14 4508:20 4509
11232:6 11288:4 11367:1 11371:2 11465:37 11564:1 1
5:1 18239:2 18305:3 18934:3 18993:4 19029:2 19079:
26113:12 26132:1 26247:1 26411:2 26453:6 26496:3 2
58:1 31359:3 31360:2 31361:15 31362:15 31366:8 313
290:2 311:2 368:6 526:4 561:1 574:998 639:53 640:1
13 3004:9 3009:10 3010:88 3011:2 3012:78 3014:8 30
```

Figura 1: Clase Edad formateada como Espacio Vectorial

Luego de esto, se debe realizar el *Cross-Validation*, para esto el set de entrenamiento (el archivo que obtuvimos anteriormente) se debe dividir en 10 partes iguales, luego debe asignarse un Split para evaluar y los restantes como entrenamiento, esto se debe repetir 10 veces para su correcta realización.



10-Fold Cross-Validation. Los cuadrados amarillos representan al set de evaluación y los

Figura 2: Cross-Validation

Gracias al uso de *SVM MultiClass* podremos obtener el archivo con las predicciones, para esto se debe asignar un valor a la *variable de relajo*, el cual para esta tarea se usó 5000. La variación de este valor afecta positiva o negativamente las predicciones obtenidas.

DESARROLLO

En la tabla 1 a continuación se presentara el MRR (Mean Reciprocal Rank), el promedio de los diez splits de cada clase (Average Loss), el accuracy (precisión en el etiquetado), y posterior a la tabla 1 se presentaran las matrices de confusión de las clases.

Nombre Clase	MRR	Promedio Splits (Average Loss)	Accuracy
Edad	0.251095	45%	0.58
Etnia	0.74555	31%	0.59
Estado Civil	0.7819	37%	0.63
Clase Social	0.970085	4.5%	0.955
Religión	0.80098	31.5%	0.685
Estilo de Vida	0.105735	55.5%	0.445
Partido Político	0.958665	7%	0.93

Tabla 1: MMR y Promedio de Splits

Explicación:

MMR: Esta estadística sirve para la evaluación de procesos que generen posibles respuestas a una interrogante cualquiera. Estas respuestas estarán ordenadas por su probabilidad.

Se utilizó la siguiente fórmula para obtener el MMR de cada clase:

$$MRR = \frac{1}{|Q|}\sum_{i=1}^{|Q|}\frac{1}{rank_i}$$

Ecuación 1: Formula MMR

Para este conjunto de clases la que presento mejor MMR fue la Clase Social, y el peor fue Estilo de Vida.

Promedio Splits: En la realización del Cross-Validation el set fue dividido en 10 partes donde SVM MultiClass dio promedios por cada Split. Estos 10 promedios fueron promediados entre sí para dar el promedio total de la clase. Para este set se puede decir que la clase con más errónea es Estilo de Vida (Cabe mencionar que esta clase era la que tenía más variables, debido a que la etiqueta podía tomar hasta 16 valores).

Ahondando más en el porqué de estos errores, es posible a que se deba a lo poco que los usuarios escribían sobre sí mismos, sobre sus acciones día a día y sobre sus intereses marcados, además ciertos perfiles tampoco daban información para inferir estos datos, lo cual causo que la gran parte de los perfiles (más de 150 de un total de 200) no pudieran ser correctamente identificados.

Accuracy: Esta medida se realizó para medir el desempeño al momento de calcular variables demográficas desde las sicográficas y viceversa (Siendo para este caso Estilo de Vida y Partido Político las variables Sicográficas).

En este caso podemos concluir que la mejor variable demográfica para calcular variables sicográficas es *Partido Político*, y para el caso contrario es *Clase Social*

Matrices de Confusión

La matriz de confusión nos permite ver la representación del desempeño de SVM al momento de realizar el etiquetado. Esta permite ver los resultados *True Positive, True Negative, False Positive y False Negative*. Esta información es útil al momento de obtener otras métricas.

Edad

PREDICTED	Under 12 years old	12-17 years old	18-24 years old	25-34 years old	35-44 years old
ACTUAL					
Under 12 years old	0	0	0	0	0
12-17 years old	0	14	4	2	1
18-24 years old	0	2	1	1	1
25-34 years old	0	1	2	0	0
35-44 years old	0	2	0	0	0
45-54 years old	0	0	0	0	0
55-64 years old	0	0	0	0	0
65-74 years old	0	0	0	0	0
75 years or older	0	0	0	0	0
No identificable	0	18	5	3	3

PREDICTED	45-54 years old	55-64 years old	65-74 years old	75 years or older	No identificable
ACTUAL					
Under 12 years old	0	0	0	0	0
12-17 years old	0	0	0	0	18
18-24 years old	0	0	0	0	10
25-34 years old	0	0	0	0	4
35-44 years old	0	0	0	0	3
45-54 years old	0	0	0	0	1
55-64 years old	0	0	0	0	1
65-74 years old	0	0	0	0	0
75 years or older	0	0	0	0	2
No identificable	0	0	0	0	101

Etnia

PREDICTED	White	Hispanic or Latino	Black or African American
ACTUAL			
White	8	0	1
Hispanic or Latino	0	37	0
Black or African American	2	0	0
Native American or American Indian	0	0	0
Asian / Pacific Islander	1	0	0
Other	0	0	0
No identificable	22	6	7

PREDICTED	tive American or American Ind	Asian / Pacific Islander	Other	No identificable
ACTUAL				
White	0	0	0	24
Hispanic or Latino	0	0	0	8
Black or African American	0	0	0	7
Native American or American Indian	0	0	0	0
Asian / Pacific Islander	0	0	0	2
Other	0	0	0	2
No identificable	0	0	0	73

Estado Civil

PREDICTED	Single	Married or domestic partnership	Widowed	Divorced	Separated	No identificable
ACTUAL						
Single	20	0	1	0	0	27
Married or domestic partnership	5	3	0	0	0	10
Widowed	0	0	0	0	0	1
Divorced	0	0	0	0	0	1
Separated	0	0	0	0	0	0
No identificable	25	4	0	0	0	103

Clase Social

PREDICTED	TOP-UPPERS	BOTTOM-UPPERS	TOP-MIDDLES	BOTTLE-MIDDLES
ACTUAL				
TOP-UPPERS	0	0	0	0
BOTTOM-UPPERS	0	0	0	0
TOP-MIDDLES	0	0	0	0
BOTTLE-MIDDLES	0	0	0	0
TOP-LOWERS	0	0	0	0
BOTTLE-LOWERS	0	0	0	0
No identificable	1	0	0	0

PREDICTED	TOP-LOWERS	BOTTLE-LOWERS	No identificable
ACTUAL			
TOP-UPPERS	0	0	1
BOTTOM-UPPERS	0	0	2
TOP-MIDDLES	0	0	0
BOTTLE-MIDDLES	0	0	1
TOP-LOWERS	0	0	2
BOTTLE-LOWERS	0	0	1
No identificable	1	0	191

Religión

PREDICTED	Protestant/Other Christian	Catholic	Mormon	Jewish	Muslim
ACTUAL					
Protestant/Other Christian	0	0	0	0	0
Catholic	0	1	0	0	0
Mormon	0	0	0	0	0
Jewish	0	0	0	0	0
Muslim	0	0	0	0	0
Other non-Christian	0	0	0	0	0
No religion identity	1	0	0	0	0
No identificable	9	3	0	0	0

PREDICTED	Other non-Christian	No religion identity	No identificable
ACTUAL			
Protestant/Other Christian	0	1	9
Catholic	0	0	8
Mormon	0	0	0
Jewish	0	0	0
Muslim	0	2	1
Other non-Christian	0	0	0
No religion identity	0	4	12
No identificable	0	17	132

Estilo de Vida

PREDICTED	Activism	Asceticism	Modern Primitivism	Back to the land	Bohemianism
ACTUAL					
Activism	0	0	0	0	0
Asceticism	0	0	0	0	0
Modern Primitivism	0	0	0	0	0
Back to the land	0	0	0	0	0
Bohemianism	0	0	0	0	0
Clothes	0	0	0	0	0
Communal	0	0	0	0	0
Hippies	0	0	0	0	0
Nomadism	0	0	0	0	0
Quirky alone	0	0	0	0	0
Simple living	0	0	0	0	0
Groupie	0	0	0	0	0
Rural lifestyle	0	0	0	0	0
Traditional lifestyle	1	0	0	0	0
Other	0	0	0	0	0
No identificable	1	0	0	0	0

PREDICTED	Clothes	Communal	Hippies	Nomadism	Quirkyalone	Simple living
ACTUAL						
Activism	0	0	0	0	0	0
Asceticism	0	0	0	0	0	0
Modern Primitivism	0	0	0	0	0	0
Back to the land	0	0	0	0	0	0
Bohemianism	0	0	0	0	0	0
Clothes	0	0	0	0	0	0
Communal	0	0	0	0	0	0
Hippies	0	0	0	0	0	0
Nomadism	0	0	0	0	0	0
Quirkyalone	0	0	0	0	0	0
Simple living	0	0	0	0	0	0
Groupie	0	0	0	0	0	0
Rural lifestyle	0	0	0	0	0	0
Traditional lifestyle	0	0	0	0	0	0
Other	0	0	0	0	0	0
No identificable	0	0	0	0	0	0

PREDICTED	Groupie	Rural lifestyle	Traditional lifestyle	Other	No identificable
ACTUAL					
Activism	0	0	0	0	1
Asceticism	0	0	0	0	1
Modern Primitivism	0	0	0	0	0
Back to the land	0	0	0	0	0
Bohemianism	0	0	1	0	0
Clothes	0	0	0	0	0
Communal	0	0	0	0	0
Hippies	0	0	0	0	0
Nomadism	0	0	0	0	1
Quirkyalone	0	0	0	0	0
Simple living	0	0	0	0	0
Groupie	0	0	0	0	2
Rural lifestyle	0	0	0	0	1
Traditional lifestyle	1	0	46	1	48
Other	0	0	2	0	1
No identificable	0	0	49	0	43

Partido Político

PREDICTED	Republicano	Democrata	Libertarian/Libertario
ACTUAL			
Republicano	0	0	0
Democrata	0	0	0
Libertarian/Libertario	0	0	0
Independiente	0	0	0
Green/Verde	0	0	0
Other	0	0	0
No identificable	0	2	1

PREDICTED	Independiente	Green/Verde	Other	No identificable
ACTUAL				
Republicano	0	0	0	1
Democrata	0	0	0	1
Libertarian/Libertario	0	0	0	3
Independiente	0	0	0	0
Green/Verde	0	0	0	0
Other	0	0	0	5
No identificable	0	0	1	186

CONCLUSIÓN

La clase con mejor métrica MMR fue Clase Social y la peor fue Estilo de Vida, esto puede deberse a la ambigüedad en las preguntas y respuestas de los perfiles al momento de definirse a sí mismos lo cual hizo más difícil la obtención correcta del valor de la clase Estilo de Vida, por otro lado, un alto número de usuarios tiene la tendencia a mencionar los objetos materiales que son capaces de obtener, lo cual es útil para obtener la Clase Social del mismo.

En las matrices de confusión se puede apreciar que efectivamente para la mayoría de las clases (En especial la clase de Estilo de Vida) no se podía identificar el estilo de vida que llevaba el usuario.

Respecto a los errores, primero que todo está el error humano, tanto como en la realización de la tarea 1 (al momento de etiquetar cada perfil) como en el momento de la creación de la bolsa de palabras. Puede haberse dado el caso de errores de formato al momento de trabajar con Cross-Validation, lo cual entregaría valores erróneos.

Finalmente, al momento de obtener variables demográficas usando las sicográficas, la clase que mostro mejor desempeño fue Partido Político, mientras que, para el caso contrario, la clase fue Clase Social.