

Sistemas Inteligentes

Tarea 4 Aprendizaje Supervisado

Profesor: Alejandro Figueroa
alejandro.figueroa@unab.cl

Ayudante: Jean Contreras
j.contrerasleyton@uandresbello.edu

Horario: miércoles 14:10-15:40
Jueves 12:10-13:40

Fecha de Publicación: miércoles 19 de abril de 2017

Fecha de Entrega: viernes 26 de mayo de 2017

Lugar: Horario de clases/ayudantía/con Haydeé Vidal

Aspectos Generales

- El trabajo es individual.
- La entrega del informe impreso debe ser realizada de manera presencial, en horario de clases.
- Lea atentamente las indicaciones esbozadas en el syllabus del curso.

Objetivos

El objetivo es aprender la metodología para construir un modelo de predicción, ocupando los datos etiquetados en el primer trabajo. Para realizar aprendizaje supervisado se necesitan cuatro componentes: un espacio vectorial, una clase de modelo, una metodología experimental y una métrica. En esta tarea, utilizaremos [SVM Multiclass](#) como modelo base y MRR como métrica de desempeño. En cuanto al espacio vectorial, [utilizaremos el contenido del perfil del usuario](#) y como metodología de evaluación cross-validation.

Espacio Vectorial

Para hacer aprendizaje supervisado debemos modelar el problema mediante vectores. Si asumimos que cada una de las palabras es un atributo, entonces podríamos crear una representación vectorial que indique la presencia o ausencia de esta palabra en la pregunta. Para la etiqueta se utilizará un valor desde 1 hasta 9. Cada etiqueta debe estar asociada con un único valor.

Consideremos el siguiente ejemplo:

Título y Cuerpo (Unidos) Etiqueta (E)

Pregunta 1 (P1)	What type of ham is your favorite? My favorite...	3
Pregunta 2 (P2)	What website will let you ask questions? ...	1
Pregunta 3 (P3)	Test post please ignore? LOL...	6
Pregunta 4 (P4)	Spam, baloney, or ramon noodles? Which is ...	3

1er Semestre 2017 - Sistemas Inteligentes

Pregunta 5 (P5) Which restaurant has the best tasting onion ... 2

Pregunta 6 (P6) Is Argentina a first world country?... 3

Pregunta 7 (P7) What rugby position am I?... 5

El primer paso sería construir una lista de palabras (Bolsa de palabras) para toda la colección. Para este ejemplo (case-insensitive), tendríamos;

ID	Palabra	ID	Palabra	ID	Palabra
1	WHAT	6	FAVORITE	11	ASK
2	TYPE	7	MY	12	QUESTIONS
3	OF	8	WILL	13	TEST
4	HAM	9	LET	14	POST
5	YOUR	10	YOU	15	PLEASE

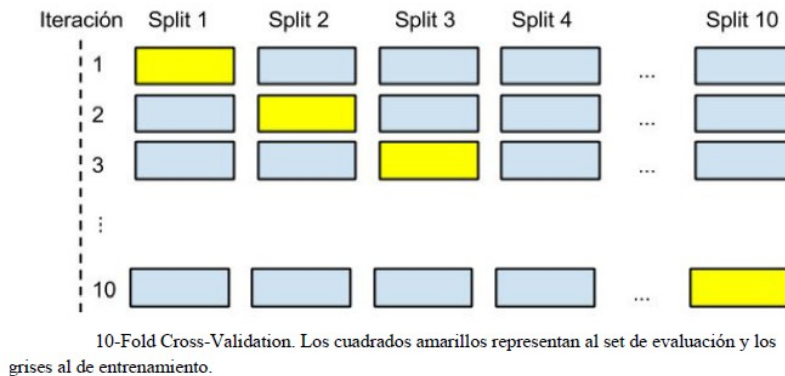
Luego, podríamos representar las preguntas, como vectores, indicando la frecuencia de la palabra en la pregunta en el índice correspondiente a la palabra. Ahora, estamos en una posición de generar los vectores. En esta tarea, utilizaremos vectores con el formato SVM multiclass , el que luego será descrito.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	E
P1	1	1	5	3	4	5	0	1	1	0	0	5	1	1	0	1
P2	4	1	1	0	4	1	1	6	6	0	3	0	1	2	2	3
P3	0	5	0	4	3	4	2	5	1	6	1	3	6	1	1	3
P4	4	2	4	1	2	4	4	3	0	5	3	2	1	2	3	1
P5	0	3	2	2	6	0	1	3	3	2	1	6	1	6	3	2
P6	0	2	1	5	5	6	2	4	0	4	0	4	1	5	5	1
P7	3	3	3	1	3	3	3	2	5	3	5	3	0	2	5	3

Cross-Validation

Una vez obtenido los vectores hay que dividir el set de entrenamiento en diez partes iguales (por ejemplo, si en total se tienen 1000 preguntas, cada split (subset) consta de 100 preguntas etiquetadas). Luego, se debe repetir diez veces el proceso de asignar un split para evaluar y los restantes nueve como entrenamiento.

1er Semestre 2017 - Sistemas Inteligentes



Si llamamos a cada split S_i , con $i=1\dots 10$, tenemos que cada set de entrenamiento se conformará de la siguiente forma:

$$E_1 = S_2 + S_3 + S_4 + S_5 + S_6 + S_7 + S_8 + S_9 + S_{10}$$

$$E_2 = S_1 + S_3 + S_4 + S_5 + S_6 + S_7 + S_8 + S_9 + S_{10}$$

$$E_3 = S_1 + S_2 + S_4 + S_5 + S_6 + S_7 + S_8 + S_9 + S_{10}$$

$$E_4 = S_1 + S_2 + S_3 + S_5 + S_6 + S_7 + S_8 + S_9 + S_{10}$$

$$E_5 = S_1 + S_2 + S_3 + S_4 + S_6 + S_7 + S_8 + S_9 + S_{10}$$

$$E_6 = S_1 + S_2 + S_3 + S_4 + S_5 + S_7 + S_8 + S_9 + S_{10}$$

$$E_7 = S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_8 + S_9 + S_{10}$$

$$E_8 = S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7 + S_9 + S_{10}$$

$$E_9 = S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7 + S_8 + S_{10}$$

$$E_{10} = S_1 + S_2 + S_3 + S_4 + S_5 + S_6 + S_7 + S_8 + S_9$$

SVM multiclass

SVM multiclass es un clasificador supervisado, con el que se generará el modelo de predicción. Al ser un clasificador supervisado, es necesario realizar un entrenamiento previo, en el que se le entrega cada uno de los vectores generados, con su correspondiente etiqueta.

El formato que sigue SVM multiclass, para cada uno de los datos de entrenamiento, para leer los archivos es el siguiente:

```
<target> <feature>:<value> <feature>:<value> ... <feature>:<value> #
<info>
```

En nuestro caso;

- **target** indica la clase de la pregunta correspondiente al vector
- **feature** es el id de la palabra
- **value** indica la frecuencia de la palabra con ese id, en la pregunta correspondiente al vector.

Aquellas palabras que no se encuentren presentes en la respuesta (lo que equivaldría a $\text{value} = 0$) se pueden omitir. Para el ejemplo descrito en la sección anterior, tendríamos:

1 1:1 2:1 3:5 4:3 5:4 6:5 8:1 9:1 12:5 13:1 14:1
3 1:4 2:1 3:1 5:4 6:1 7:1 8:6 9:6 11:3 13:1 14:2 15:2
3 2:5 4:4 5:3 6:4 7:2 8:5 9:1 10:6 11:1 12:3 13:6 14:1 15:1
1 1:4 2:2 3:4 4:1 5:2 6:4 7:4 8:3 10:5 11:3 12:2 13:1 14:2 15:3
2 2:3 3:2 4:2 5:6 7:1 8:3 9:3 10:2 11:1 12:6 13:1 14:6 15:3

Cabe destacar que cada ID debe referirse a la misma palabra a través de todas las instancias de entrenamiento y que el formato de SVM multiclass exige que se escriban en orden ascendente (en términos de ID).

Una vez contruidos los conjuntos de entrenamiento, se debe generar los modelos. Para ésto, se debe utilizar el comando “svm_multiclass_learn -c X E_i M_i”, con lo que se genera un modelo (M_i) por cada uno de los E_i. X corresponde a un factor de penalización que utiliza el algoritmo.

Para evaluar el modelo, debe ejecutarse sobre el split que no fue considerado en el respectivo E_i, es decir S_i, mediante el comando “svm_multiclass_classify S_i M_i R_i ” donde R_i es el archivo en donde se almacenará el resultado.

Por ejemplo,

```
$ ./svm_multiclass_learn -c 5000 train.dat model.dat  
$ ./svm_multiclass_classify test.dat model.dat predictions.dat
```

Como nuestro, el resultado de ejecutar el comando de prueba queda en el archivo **predictions.dat**

Desarrollo

Cuando se tengan los diez archivos de resultados, se debe calcular el desempeño del clasificador SVM para predecir cada una de las variables demográficas y sicográficas (exceptuando la última de valores múltiples). Para cada una de ellas calcule el MRR y la matriz de confusión. Entregue también los promedios de los diez splits. Comente acerca de los errores ¿Hay algún patrón? ¿Cómo podría mejorar? ¿Qué puede decir de los errores?

Haga lo mismo pero ahora considere en el vector de características sólo las variables demográficas y trate de predecir cada una de las variables sicográficas. Hagalo también al revés, considere sólo las variables sicográficas en el vector de características e intente predecir cada una de las variables demográficas.

Finalmente, ¿Qué variable demográfica combinada con el espacio vectorial inicial podría mejorar la predicción de alguna de las variables sicográficas? Y al revés, ¿Qué variable sicográfica combinada con el espacio vectorial inicial podría mejorar la predicción de alguna de las variables demográficas?