Universidad Andrés Bello
Facultad de Ingeniería
Ingeniería Civil Informática

Sistemas Inteligentes

Trabajo 2 - Clasificación

Profesor:	Alejandro Figueroa	**Fecha de publicación:**	13/08/15
Ayudantes:	Alexander Espina - Daniel Palomera	**Fecha de entrega:**	28/08/15

Aspectos generales

- La entrega del informe debe ser realizada, de manera impresa, en la fecha indicada, en la secretaría.
- Se debe utilizar los datos etiquetados por todos los integrantes (máx. dos).

Objetivos

El objetivo apunta a aprender un modelo de predicción, ocupando los datos etiquetados en el trabajo anterior. Este modelo permitirá predecir si una respuestas es informacional, basándose en el cuerpo de la respuesta.

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^{light} como modelo base y F-Score como métricas de desempeño. En cuanto al espacio vectorial, utilizaremos el cuerpo de las respuestas 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 son un atributo, entonces podríamos crear una representación vectorial que indique frecuencia de esta palabra en la respuesta. Para la etiqueta se utilizará un 1 en caso de que la respuesta sea informativa (ejemplo positivo) y un -1 en caso de que no lo sea (ejemplo negativo)

Consideremos el siguiente ejemplo:

Pregunta:	**how do i open my yahoo address?**	
	Cuerpo	**Etiqueta (E)**
Respuesta 1 (R1)	Just go to the Yahoo! main page and click on Yahoo! services. When the page comes up, click on Address Book	1
Respuesta 2 (R2)	Go to the Yahoo! main page and click on mail. When the page comes up, type in your User Name and Password.	1
Respuesta 3 (R3)	Use gmail.	-1

El primer paso sería construir una lista de palabras para toda la colección. Para este ejemplo, tendríamos;

ID	Palabra	ID	Palabra	ID	Palabra	ID	Palabra	ID	Palabra
1	address	**6**	gmail	**11**	main	**16**	services	**21**	use
2	and	**7**	go	**12**	name	**17**	the	**22**	user
3	book	**8**	in	**13**	on	**18**	to	**23**	when
4	click	**9**	just	**14**	page	**19**	type	**24**	yahoo
5	comes	**10**	mail	**15**	password	**20**	up	**25**	your

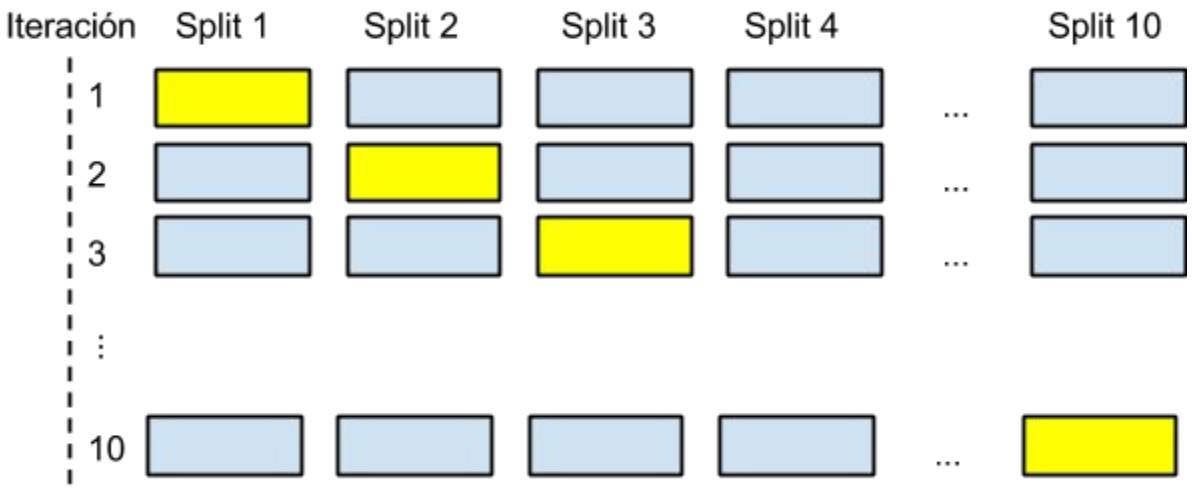
Luego, podríamos representar las respuestas, como vectores, indicando la presencia (con un 1) o ausencia (con un 0).

	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**	**16**	**17**	**18**	**19**	**20**	**21**	**22**	**23**	**24**	**25**	**E**
R1	1	0	1	1	1	0	1	0	1	0	1	0	2	2	0	1	2	1	0	1	0	0	1	1	0	**1**
R2	1	2	0	1	1	0	1	1	0	1	1	1	1	2	1	0	2	1	1	1	0	1	1	1	1	**1**
R3	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	**-1**

Ahora, estamos en una posición de generar los vectores. En esta tarea, utilizaremos vectores con el formato SVM^{Ligth}, el que será presentado más adelante.

Cross-Validation

Una vez obtenido los vectores hay que dividir el set de entrenamiento en 10 partes iguales (por ejemplo, si en total se tienen 1000 respuestas, cada *split* (sub-set) consta de 100 respuestas etiquetadas). Luego, se debe repetir 10 veces el proceso de asignar un split para evaluar y los restantes 9 como entrenamiento.



Imágen 1: 10-Fold Cross-Validation. Los cuadrados amarillos representan al set de evaluación y los grises al de entrenamiento.

Si llamamos a cada split S_i, con i=1...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^{Light}

SVM^{Light} [1] es un clasificador supervisado, lineal, 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 SVMLight, 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** es +1 si la respuesta es informativa o -1 si no lo es
- **feature** es el id de la palabra
- **value** indica la frecuencia de esta palabra en la respuesta.

Aquellas palabras que no se encuentren presentes en alguna respuesta (lo que equivaldría a value = 0) se deben omitir.

Para el ejemplo descrito en la sección anterior, tendríamos:

```
+1 1:1 4:1 6:1 8:1 9:1 13:1 14:1 #comentario
+1 3:1 4:1 6:1 9:1 15:1 #comentario
-1 2:1 5:1 6:1 7:1 10:1 11:1 12:1 #comentario
```

como nuestro archivo de entrada para SVM^{Light} .

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^{Light} exige que se escriban en orden ascendente (en términos de ID).

Una vez construidos los conjuntos de entrenamiento, se debe generar los modelos. Para ésto, se debe utilizar el comando "**svm_light E $\mathbf{M_i}$**", con lo que se genera un modelo (M_i) por cada uno de los E_i.

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_light_classify $\mathbf{S_i}$ $\mathbf{M_i}$ $\mathbf{R_i}$**" donde R_i es el archivo en donde se almacenará el resultado.

Salida

La salida, en donde se almacenará el resultado, contiene una línea por cada instancia, describiendo el valor de la función de decisión para esta instancia. Para clasificación, se puede tomar el signo para determinar si la clase es positiva (informativa) o negativa (no informativa).

[1] http://svmlight.joachims.org/

Desarrollo

Cuando se tengan los diez archivos de resultados, se debe calcular el desempeño del clasificador SVM para predecir si una respuesta es o no informativa. Con este objetivo, examine los archivos de resultados y compare las etiquetas asignadas manualmente y las que entrega el clasificador.

Calcule la accuracy y la matriz de confusión. Además, la precisión, recall y F-Score de ambas clases por split. 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? ¿Hay alguna correlación entre el número de palabras de las respuestas y su tipo? ¿Existe alguna relación entre el tipo de la pregunta y la distribución de las clases de sus respuesta? Comente que observa por categoría (de primer nivel): ¿Hay categorías con una mayor tasa de error? Tome en cuenta sólo las categorías más frecuentes en su conjunto de datos.