Centro de Investigación en Matemáticas

Unidad Monterrey

Proyecto No. 1.

Métodos multivariados de Análisis de Datos

Filtro de spam personalizado.

Ricardo Cruz Sánchez
Rolando Corona Jiménez

Índice

1. Introducción.

Aplicación de modelos de clasificación para obtener un filtro (personalizado) para correos electrónicos spam.

2. Análisis exploratorio.

2.1. Descripción del conjunto de datos.

El conjunto de datos proviene de una serie de correos electrónicos del personal de la empresa HP. Los correos etiquetados como spam fueron proporcionados por el administrador del servidor de correo de la empresa, mientras que los correos que no están etiquetados como spam corresponden a correos personales y de trabajo de George Forman, es por ello que palabras como *george* o código de área 650 son indicadores de no spam. La base de datos fue creada por Mark Hopkins, Erik Reeber, George Forman y Jaap Suermondt de Hewlett-Packard Labs.

En total se cuentan 4601 observaciones, de las cuales 1813 fueron etiquetadas como spam, que corresponde al 39.4 % del total. Las observaciones están representadas a través de un conjunto de 58 atributos: 57 variables cuantitativas y una variable cualitativa nominal de clase. Ninguno de los atributos presenta datos faltantes.

Spam	1813	39.4 %
Non-Spam	2788	60.6 %

Tabla 1: Distribución de clases.

2.2. Diccionario de datos

Los 58 atributos se pueden agrupar en:

- 48 variables cuantitativas continuas con rango $[0, 100]$, de la forma word_freq_WORD, que indica el porcentaje de palabras en el correo que coinciden con WORD, es decir:

$$word\_freq\_WORD = 100 \times \frac{\#\text{apariciones de WORD en el correo}}{\#\text{total de palabras en el correo}}$$

- 6 variables cuantitativas continuas con rango $[0, 100]$, de la forma char_freq_CHAR, que indica el porcentaje de caracteres en el correo que coinciden con CHAR, es decir:

$$char\_freq\_CHAR = 100 \times \frac{\#\text{apariciones de CHAR en el correo}}{\#\text{total de caracteres en el correo}}$$

- 1 variable cuantitativa continua capital_run_length_average con rango $[0, \infty)$, que es igual a longitud media de las secuencias contiguas de letras mayúsculas que aparecen en el correo.

- 1 variable cuantitativa discreta capital_run_length_longest con rango $[0, \infty)$, que es igual a la longitud de la secuencia contigua de letras mayúsculas más larga que aparece en el correo.
- 1 variable cuantitativa discreta capital_run_length_total con rango $[0, \infty)$, que es igual a la suma de las longitudes de las secuencias contiguas de letras mayúsculas que aparecen en el correo.
- 1 variable nominal de clase con valores en $\{0, 1\}$, que indica si el correo se considera spam (1) o no (0).

La documentación del conjunto de datos no indica los criterios para la selección de las 48 palabras y 6 caracteres usados para la definición de sus correspondientes variables.

make	order	business	hp	data	cs
address	mail	email	hpl	415	meeting
all	receive	you	george	85	original
3d	will	credit	650	technology	project
our	people	your	lab	1999	re
over	report	font	labs	parts	edu
remove	addresses	000	telnet	pm	table
internet	free	money	857	direct	conference

Tabla 2: Palabras que corresponden a las variables de tipo word_freq_WORD.

;
(
[
!
$
#

Tabla 3: Caracteres que corresponden a las variables de tipo char_freq_CHAR.

2.3. Matriz de correlación.

2.4. Palabras más frecuentes por clase.

A modo de resumen se muestra una representación gráfica de las palabras más frecuentes por cada clase, para obtener la frecuencia de cada palabra, se realizó la suma de cada variable sobre todas las observaciones de cada clase.

2.5. Representación en componentes principales.

3. Métricas para la clasificación binaria.

3.1. Accuracy

3.2. F1-score

3.3. Selección de métricas para el problema del spam

4. Modelos de clasificación.

4.1. Regresión logística.

4.2. Máquinas de soporte vectorial.

4.3. Árboles de decisión.

4.4. Random Forest.

4.5. Modelos con reducción de dimensión y selección de variables

4.5.1. Representación en componentes principales tras la selección de variables.

4.6. Comparativa con otros modelos

de spambase.DOCUMENTATION 7 % misclassification error. False positives (marking good mail as spam) are very undesirable. If we insist on zero false positives in the training/testing set, $20 - 25$ % of the spam passed through the filter.

revisar papers donde trabajan con el ejemplo

5. Conclusiones.

Yo digo que no hay pedo

Referencias

[1] Lorrie Faith Cranor and Brian A. LaMacchia. 1998. Spam!. Commun. ACM 41, 8 (August 1998), 74-83. DOI: https://doi.org/10.1145/280324.280336

[2] Emilio Ferrara. 2019. The history of digital spam. Commun. ACM 62, 8 (July 2019), 82-91. DOI: https://doi.org/10.1145/3299768

[3] Hastie, T., Tibshirani, R.,, Friedman, J. (2001). The Elements of Statistical Learning. New York, NY, USA: Springer New York Inc..