
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.

# 3. Métricas para clasificación binaria.

Las métricas de evaluación permiten medir y resumir la calidad de un modelo entrenado al ser probado con nuevas observaciones. La exactitud (*accuracy*) es una de las métricas más comunes para evaluar la capacidad de generalización de un clasificador, sin embargo no siempre resulta ser la ideal, y esto depende específicamente del problema en cuestión. Entre la distintas métricas que existen para el problema de clasificación binaria, a continuación se mencionan algunas, para finalmente discutir sobre cuál es la ideal para la clasificación de spam, con el fin de tener un criterio de preferencia entre los clasificadores que serán presentados más adelante.

Existen dos tipos de errores al asignar una clase a una observación: el *falso positivo* $fp$ (diagnóstico positivo, condición de interés ausente) y el *falso negativo* $fn$(diagnóstico negativo, condición de interés presente). De forma similar se definen los verdaderos positivos $tp$ y verdaderos negativos $tn$.

A partir de lo anterior, se definen las métricas de interés que se muestran en la tabla 4.

| **Métrica** | **Fórmula** | **Enfoque de la evaluación** |
|---|---|---|
| Accuracy | $\frac{tp+tn}{tp+fn+fp+tn}$ | Desempeño general |
| Sensitivity | $\frac{tp}{tp+fn}$ | Efectividad para identificar a la clase positiva |
| Specificity | $\frac{tn}{fp+tn}$ | Efectividad para identificar a la clase negativa |

Tabla 4: Métricas para clasificación binaria

## 3.1. Selección de métricas para el problema del spam.

Para la clasificación de spam, es fundamental que los correos que no son spam, en la medida de lo posible, no sea etiquetados como spam, pues de lo contrario, los usuarios podrían perder información de valor si no revisan periódicamente la bandeja de spam, lo cual sería contraproducente. Dicho en términos de las métricas mencionadas anteriormente, se desea un clasificador con alta especificidad (specificity), de modo que la probabilidad de que un correo que no es spam no sea marcado como spam, sea alta. Sin embargo, puede pasar que al aumentar la especificidad, la exactitud y la sensibilidad (sensitivity) se vean reducidas, lo que causaría que el clasificador no se capaz de identificar muchos correos que deberían ser marcados como spam, en cuyo caso el usuario los recibiría en su bandeja principal y tendría que marcar manualmente dichos correos como spam. Este último enfoque es preferido, pues se trata de garantizar que los usuarios no pierdan información de valor. Teniendo en cuenta estas consideraciones, se procede a ajustar una serie de clasificadores binarios.

# 4. Modelos de clasificación.

Los modelos a continuación presentados, se implementaron a través del paquete estadístico *R*. Para su análisis se requiere la separación del conjunto original de datos en un conjunto de entrenamiento (aquél con el que se ajustan los modelos) y un conjunto de prueba (conjunto con el que se validan los resultados).

En este caso, se optó por un muestreo aleatorio para generar ambas submuestras. El conjunto de entrenamiento consiste de 3680 observaciones (aproximadamente un 80 %) y el de prueba contiene 921 observaciones.

## 4.1. Regresión logística.

La regresión logística, es un caso particular de los modelos lineales generalizados. Su característica principal es el tener una variable dependiente dicotómica. Utilizá la función enlace *logit* y gracias a esto se modela:

$$logit(p) = ln(\frac{p}{1-p}) = \beta x$$

donde $p$ corresponde a la probabilidad asociada a la distribución binomial de la cual se generan los valores de $y$, es decir, $y \sim Bin(n, p)$

Una vez que se encuentran los parámetros $\beta$, solo es cuestión de evaluar la observación de las variables independientes $x$ y obtener el valor de $p$ aplicando la función inversa de *logit*.

El valor estimado de $y$ dependerá de $p$ y un *punto de corte*, que usualmente es el valor 0.5, ya que si la observación es más grande que el punto de corte quiere decir que $y$ se asemeja más a la distribución de los valores $y = 1$ y se asignará como 1. En caso de ser menor al punto de corte se determina que el valor estimado es 0.

Se ajustó la regresión considerando las 57 variables, lo cual arrojo las métricas mostradas en la tabla 5:

| | Entrenamiento | Prueba |
|---|---|---|
| Accuracy | .93 | .92 % |
| Recall | .90 | .88 % |
| Specificity | .92 | .94 % |

Tabla 5: Métricas modelo full.

Regresando al punto de corte, se esperaría que las distribuciones de los casos verdaderos positivos y verdaderos negativos, tenga un comportamiento similar al de la figura 1, tratando de minimizar las áreas de falsos positivos y negativos.

Se puede cambiar el punto de corte para tratar de minimizar las áreas de errores y por ende mejorar las métricas mencionadas. Cabe resaltar que el punto de corte se determina con

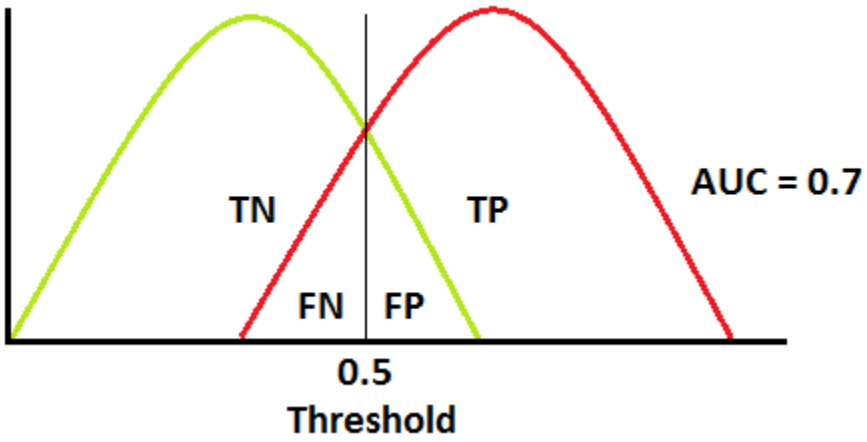
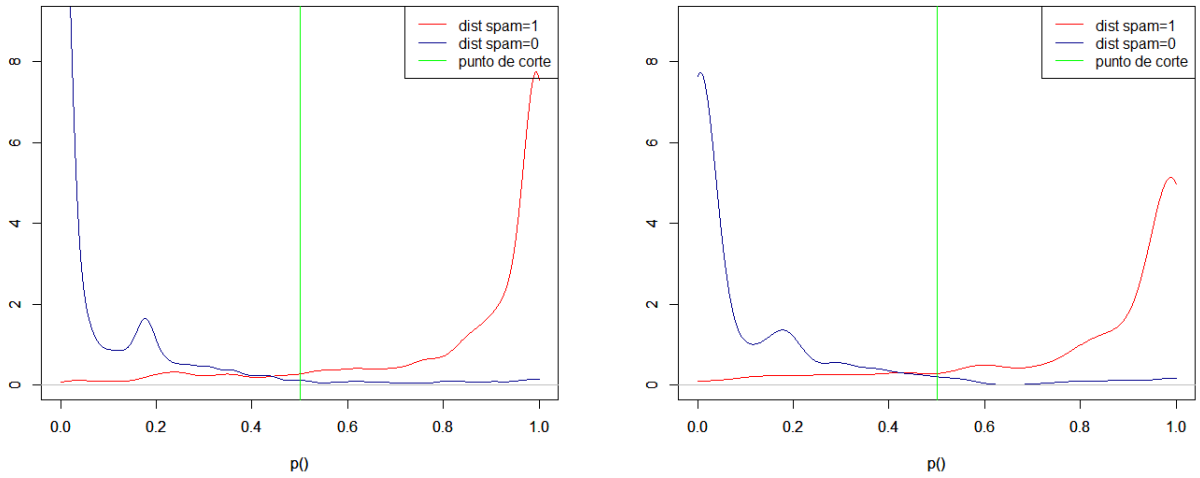



Figura 1: Distribuciones de TP y FP con relación al punto de corte.

el conjunto de entrenamiento.

La figura 2 muestra las distribuciones de verdaderos positivos y negativos para el conjunto de entrenamiento. En las distribuciones de entrenamiento se muestra como se puede desplazar el punto de corte a la izquierda para poder minimizar las áreas de falsos.




(a) Entrenamiento

(b) Prueba

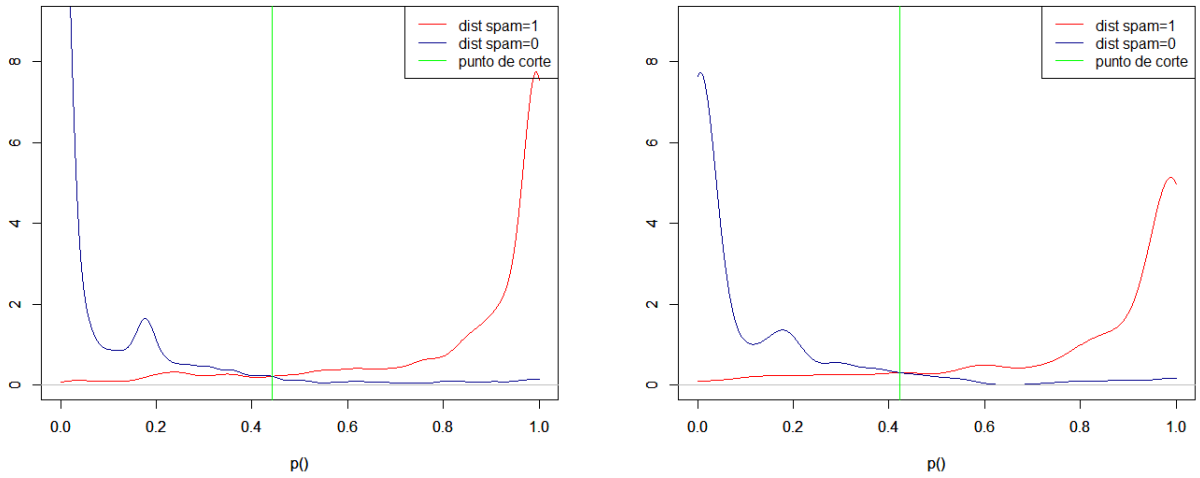
Figura 2: Densidades TP y TN

Por lo anterior, se plantea recorrer el punto de corte, a donde se intersectan las dos distribuciones (0.4227). El resultado de esto se muestra en la figura 3 y la tabla 6

Se puede apreciar que las métricas no se modificaron significativamente, por lo que se podría dejar el punto de corte donde inicialmente estaba. Esto depende totalmente de la estrategia que se quiera seguir y se debe estar consciente que las medidas de sensitivity y

| | Entrenamiento | Prueba |
|---|---|---|
| Accuracy | .93 % | .91 % |
| Recall | .91 % | .90 % |
| Specificity | .95 % | .93 % |

Tabla 6: Métricas modelo full, punto de corte 0.42




(a) Entrenamiento

(b) Prueba

Figura 3: Densidades TP y TN

specificity son complementarias, es decir, existe un *trade-off* entre estas dos medidas al cambiar el punto de corte.

Por ejemplo, si la estrategia a seguir fuese no dejar pasar como spam a ningún correo que realmente no sea spam, entonces el punto de corte debe despalzarse hasta la probabilidad máxima observada en los verdaderos negativos durante el entrenamiento. La tabla 7 muestra los resultados de plantear esta estratégia:

| | Entrenamiento | Prueba |
|---|---|---|
| Accuracy | .63 % | .63 % |
| Recall | .08 % | .09 % |
| Specificity | 1 % | 1 % |

Tabla 7: Métricas modelo full, punto de corte 0.9945

En las densidades mostradas se puede observar que las distribuciones se concentran demasiado en los extremos, es decir, las probabilidades generadas están muy cerca de 0 y 1. De hecho, esto aparece como un *warning* al generar los modelos, en consola se puede leer que se generaron probabilidades 0 ó 1.

Este comportamiento se conoce como *separación completa o quasi completa* y se genera cuando para una categoría en particular, se puede observar el valor constante de alguna o algunas variables, lo cual llevaría a que se pueda determinar el valor de la respuesta con solo ver las variables que son constantes, es decir, sería un modelo determinista.

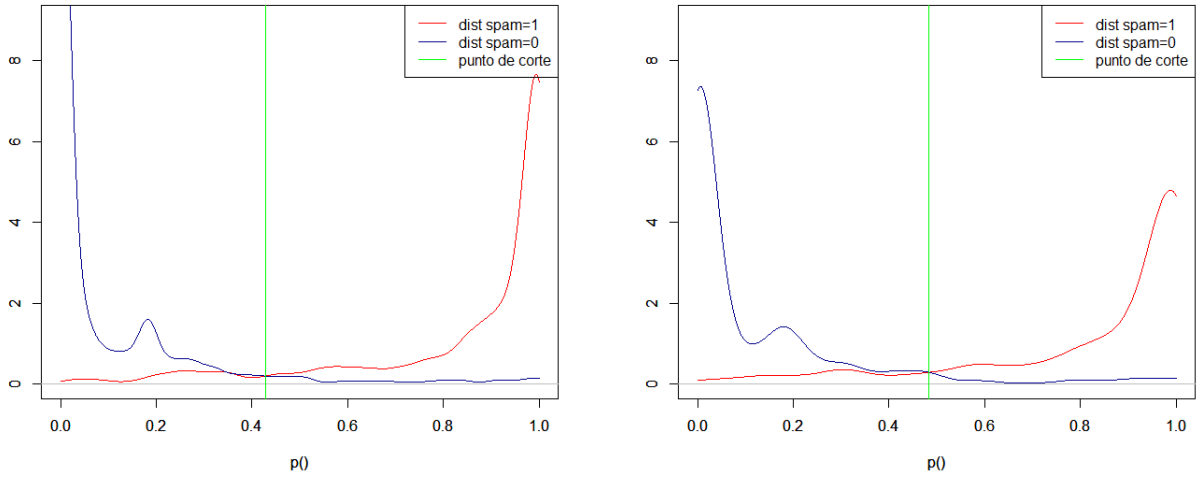
Esto puede generar sobreajuste o una mala definición del modelo para observaciones futuras. Sus posibles soluciones son considerar más observaciones en el conjunto original o realizar una selección de variables.

Se optará por la segunda alternativa. Al ejecutar un proceso de selección de variables a través de un método stepwise y AIC como criterio del modelo, se obtienen 41 variables.

Sin embargo, el desempeño es bastante similar y continua la separación completa. Así que, tal vez, no sea un problema de sobreajuste, sino la naturaleza misma del evento que se estudia. La tabla 8 contiene las métricas de este nuevo modelo y la figura 4 las respectivas densidades

| | Entrenamiento | Prueba |
|---|---|---|
| Accuracy | .93 % | .92 % |
| Recall | .91 % | .88 % |
| Specificity | .94 % | .94 % |

Tabla 8: Métricas modelo stepwisw




(a) Entrenamiento

(b) Prueba

Figura 4: Densidades TP y TN

Finalmente, en este método, se consideraron los coeficientes más importantes de acuerdo a la magnitud que poseían:

- char_freq_$: 6.66
- cs: -508.4
- george: -9.81
- conference: -4.52
- meeting: -2.69

## 4.2. Lasso y Ridge

Una de las posibles alternativas son los métodos de regularización, los cuales pueden seleccionar variables y mejorar las estimaciones de los coeficientes haciendolos tender a 0.

Para implementar estos dos métodos, primero se realiza validación cruzada, con lo que se determinará el valor de $\lambda$ en cada caso.

C-V sugiere utilizar $\lambda = .018$ para ridge y $\lambda = 0.00049$ para lasso. Con estos dos valores lasso selecciona 54 variables, mientras que ridge ocupa todas aunque las aproxima a 0.

La tabla 9 muestra el desempeño de ambos modelos en el conjunto de prueba:

| | Lasso | Ridge |
|---|---|---|
| Accuracy | .90 % | .87 % |
| Recall | .82 % | .73 % |
| Specificity | .96 % | .96 % |

Tabla 9: Métricas modelo lasso y ridge

No se mejoró el desempeño de la regresión original, de hecho, la logisitic full, sigue siendo la mejor de los modelos hasta ahora presentados.

## 4.3. Máquinas de soporte vectorial.

| | SVM | Bayes |
|---|---|---|
| Accuracy | .91 % | .91 % |
| Recall | .88 % | .91 % |
| Specificity | .93 % | .92 % |

Tabla 10: Métricas svm y regresion bayesiana

## 4.4. Árboles de decisión.

El índice de Gini se define como

$$G$$

gini podado matriz de costo

## 4.5. Random Forest.

## 4.6. Modelos con reducción de dimensión y selección de variables

## 4.7. Representación en componentes principales.

Se calculan las componentes principales y se grafica el screeplot, que muestra el porcentaje de varianza acumulada en función del número de componentes principales. La figura 5 muestra que, por ejemplo para explicar el 80 % de varianza, se requieren de al menos 30 componentes principales, lo que sugiere que los modelos entrenados en dimensión reducida pueden tener un ajuste deficiente.

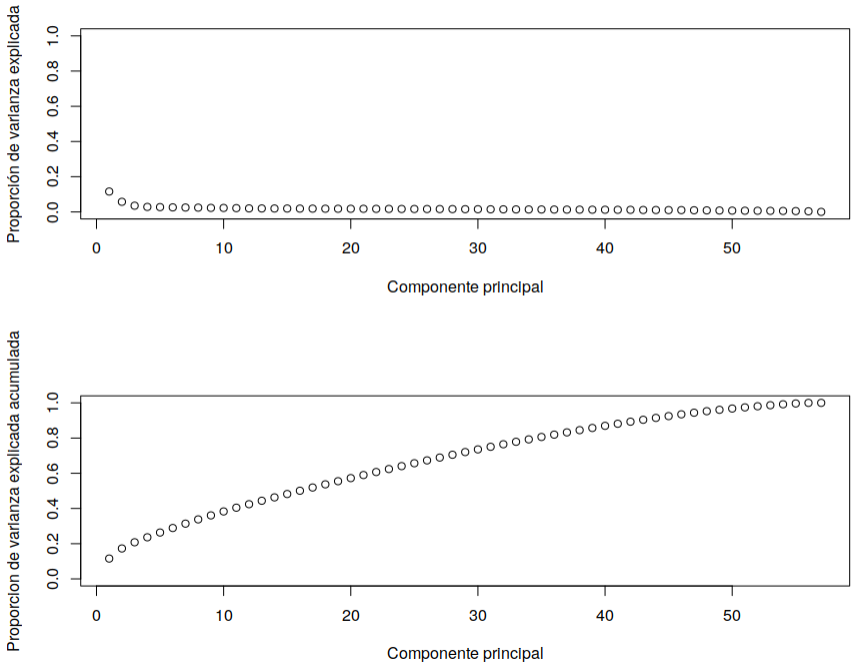



Figura 5: Screeplot

## 4.8. 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 trai-

ning/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.

[2] Emilio Ferrara. 2019. The history of digital spam. Commun. ACM 62, 8 (July 2019), 82-91.

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

[4] Marina Sokolova and Guy Lapalme. 2009. A systematic analysis of performance measures for classification tasks. Inf. Process. Manage. 45, 4 (July 2009), 427-437.