DEPARTAMENTO DE COMPUTACION

Facultad de Ciencias Exactas y Naturales - UBA

Trabajo Práctico Número 1

28 de septiembre de 2016 — Aprendizaje Automático

Integrante	LU	Correo electrónico
Bordón, Pablo	794/07	bordonpablo@gmail.com
Gasco, Emilio	171/12	gascoe@gmail.com
Gatti, Mathias	477/14	mathigatti@gmail.com

Facultad de Ciencias Exactas y Naturales
Universidad de Buenos Aires
Ciudad Universitaria - (Pabellón I/Planta Baja)
Intendente Güiraldes 2160 - C1428EGA
Ciudad Autónoma de Buenos Aires - Rep. Argentina
Tel/Fax: (54 11) 4576-3359
http://www.fcen.uba.ar

Índice

0. Extracción de atributos

1. A continuación enumeramos los atributos que escogimos como potencialmente utiles y que luego implementamos para extraer automaticamente de los mails.

   - Los correos de spam suelen ser enviados a un único destinatario, para capturar esta característica se extraen 3 atributos del encabezado del correo: *recipient_count*, *has_cc* y *has_cc* para extraer cantidad de destinatarios del correo, si hay destinatarios en copia y si hay destinatarios en copia oculta respectivamente.
   - *headers_count* cantidad de encabezados.
   - *mailer* Software utilizido para envió de correo.
   - *has_body* Nos dice si el correo tiene cuerpo o si solo consta de encabezados
   - *content_type* Tipo del contenido del cuerpo de correo. Por ejemplo: text/plain, text/html, multipart/related, multipart/alternative,etc.
   - *content_transfer_encoding* la codificación utilizada para la transferencia del correo
   - *is_mulipart* Nos dice si el cuerpo consta de varias partes
   - *subject_length* Largo del titulo del correo
   - *raw_mail_len* Largo del cuerpo del mensaje.
   - *raw_body_count_spaces* Cantidad de espacios en cuerpo de correo
   - *has_dollar* Nos dice si aparece el símbolo $ en el cuerpo del correo.
   - *has_link* Indica presencia de link http dentro del cuerpo del correo.
   - *has_html* Indica presencia de html dentro del cuerpo del correo.
   - *has_attachement* Indica la presencia de archivos adjuntos analizando content-type de las partes de correos con múltiples partes. Se consideran archivos adjuntos a las partes que no sean del tipo text/*.
   - *uppercase_count* Frecuencia de caracteres de letras mayúsculas en cuerpo de correo.
   - *has_non_english_chars* Indica presencia de caracteres de idiomas diferentes al ingles dentro del cuerpo.
   - *spaces_over_len* frecuencia de espacio en cuerpo de correo
   - En correos de tipo ham se puede observar alta frecuencia de conjunciones y artículos. Por lo que tenemos atributos por para calcular la frecuencia de los mismos. Por ejemplo: a, and, for, of, to, in, the . La frecuencia se mide por separado a cada uno de los listados. La lista surgió de analizar palabras mas frecuentes en correos de ham en comparación con correos de spam.
   - *parts_count* Cantidad de partes en correo de múltiples partes.
   - *spell_error_count* Cantidad de errores ortográficos en cuerpo de correo.
   - Por último analizamos base de prueba de correos de spam, para extraer las 100 palabras mas utilizadas. A partir de cada palabra se genera un atributos que nos indica la presencia o no de la misma.

El atributo content_type es el atributo con mas ganancia de información, posicionando en la raíz de los clasificadores de arboles cuando no se limitaba la selección de atributos a un subconjunto aleatorio de atributos. El atributo *headers_count* no resulto ser muy efectivo, la cantidad de encabezados suele ser uniforme entre correos spam y ham,variando por la inclusión de encabezados cc y bcc que ya son capturados por otros atributos.

2. El conjunto original de mails fue dividido para tener por un lado un set de entrenamiento con el cual trabajar y un set de testing para probar al final si realmente nuestros clasificadores generalizaban bien y podían clasificar correctamente instancias nuevas. Al entrenar los árboles de decisión surgió la necesidad de medir su performance de alguna manera, para el dominio del problema en particular no pareció valido utilizar F0.5 como unidad de medida, ya que la precisión es lo que tiene mayor peso en un filtro de spam debido a que se busca evitar que el clasificador catalogue como spam un mail importante.

3. Experimentamos con distintos hiper-parámetros utilizando al técnica de grid search. Por limitaciones de computo no pudimos hacer una búsqueda demasiado exhaustiva, de todas maneras logramos obtener resultados considerablemente mejores que los que hubiéramos teníamos al principio cuando probamos con los valores seteados por defecto en los clasificadores.

1. Modelos

2. Reducción de dimensionalidad

Utilizamos distintas técnicas de reduccion de la dimensionalidad PCA, Linear SVC con penalidad L1, ExtraTreesClassifier y simplemente seleccionar los 100 mejores.

3. Resultados

Luego de experimentar con distintos parametros, clasificadores e hiperparametros terminamos escogiendo -COMPLETAR- con esta versión final de nuestro clasificador de Spam, logramos un f0.5 de -COMPLETAR- sobre el conjunto de testing.

4. Discusión

A. Apendice