Universidade Vigo
Departamento de Estatística e Investigación Operativa
Grao en Intelixencia Artificial
E.S.E. Informática
Ed. Politécnico
Campus de Ourense
Tel. 986 387 000
http://esei.uvigo.es

Practica: Filtro Anti-Spam con Naive Bayes

El Naive Bayes es muy popular entre los filtros anti-spam del correo-e. Principalmente por su sencillez en la implementación y los buenos resultados que se obtienen. El objetivo de esta práctica es un ejercicio de implementación de procedimiento más sencillo anti-spam bayesiano. En el artículo [1] aparecen diferentes modificaciones al Naive Bayes además de la explicación del procedimiento (lectura recomendada!!).

1 Naive Bayes

El teorema de Bayes nos permite obtener, dado un correo-e entrante, la probabilidad de clasificarlo con *Spam* o *Ham* ($\overline{C_{Spam}}$), dado por la igualdad:

$$P\left(C_{Spam}/\,\text{mensaje}_0\right) = \frac{P\left(C_{Spam}\cap \text{mensaje}_0\right)}{P\left(\text{mensaje}_0\right)} = \frac{P\left(\text{mensaje}_0/\,C_{Spam}\right)P\left(C_{Spam}\right)}{P\left(\text{mensaje}_0/\,C_{Spam}\right)P\left(C_{Spam}\right)+P\left(\text{mensaje}_0/\,\overline{C_{Spam}}\right)P\left(\overline{C_{Spam}}\right)} \tag{1}$$

donde C_{Spam} es la categoría Spam, (frente a la categoría complementaria $\overline{C_{Spam}}$) y $mensaje_0$ es el mensaje que queremos clasificar. Intuitívamente si

$$P\left(C_{Spam}/\,\mathbf{x}\right) > Umbral \tag{2}$$

para un valor alto de $Umbral \in [0,1]$ entonces clasificaríamos el mensaje entrante como *Spam*. Para clasificadores binarios, un valor razonable de $Umbral$ es el que da la opción mayoritaria ($Umbral = 0.5$).

1.1 Obtención de las probabilidades

Por lo tanto, dado lo anterior, se deben especificar:

- Probabilidades incondicionales, expresadas como frecuencias relativas (del conjunto de datos) de aparición del suceso de interés, i.e.:

$$P(C_{Spam}) \approx \frac{\#\left\{\text{mensaje spam}\right\}}{\#\left\{\text{total de mensajes}\right\}} \tag{3}$$

$$P(\overline{C_{Spam}}) = 1 - P(C_{Spam}) \tag{4}$$

- Probabilidades condicionales. Procederemos con la versión *naive* en donde se debe transformar cada *mensaje* en un vector de longitud fija (digamos m-dimensional) indicando para cada coordenada el resultado de Bernoulli (0/1) si el atributo i-ésimo está presente o no en el mensaje. Entonces para el $mensaje_i$, transformado en el vector de características m-dimensional dado por $(t_{i,1}, t_{i,2}, \ldots, t_{i,m})$ con valores $t_{i,j} \in \{0,1\}$, $j = 1, 2, \cdots, m$, podemos expresar la probabilidad condicionada como:

$$P\left(\text{mensaje}_0/\,C_{Spam}\right) = P\left(\mathbf{x}_0/\,C_{Spam}\right) = \prod_{j=1}^{m} P\left(x_j/\,C_{Spam}\right)^{t_{0,j}}\left(1-P\left(x_j/\,C_{Spam}\right)\right)^{1-t_{0,j}}$$

con $P\left(X_j = x_j/\,C_{Spam}\right)$ la probabilidad de que la j-ésima variable de Bernoulli tome el atributo $j-$ésimo condicionado a que el mensaje pertenece a la clase C_{Spam}. Dicha probabilidad se puede estimar a partir de los datos como:

$$P\left(X_j = x_j/\,C_{Spam}\right) \approx \frac{\#\left\{\text{mensajes spam con la característica } X_j\right\}}{\#\left\{\text{total de mensajes spam}\right\}}$$

Análogamente se obtendrían las probabilidades condicionadas por la categoría $\overline{C_{Spam}}$:

$$P\left(\text{mensaje}_0/\,\overline{C_{Spam}}\right) = P\left(\mathbf{x}_0/\,\overline{C_{Spam}}\right) = \prod_{j=1}^{m} P\left(x_j/\,\overline{C_{Spam}}\right)^{t_{0,j}}\left(1-P\left(x_j/\,\overline{C_{Spam}}\right)\right)^{1-t_{0,j}}$$

y

$$P\left(X_j = x_j/\,\overline{C_{Spam}}\right) \approx \frac{\#\left\{\text{mensajes no-spam con la característica } X_j\right\}}{\#\left\{\text{total de mensajes no-spam}\right\}}$$

2 Datos

Genéricamente, se dispondrá de una colección de correos-e, digamos de tamaño M, clasificados entre correo-e Spam y No-Spam. Cada uno de los mensajes se debe transformar en un vector m-dimensional el cual debe recoger las características necesarias para que pueda el sistema clasificar correos nuevos como Spam/No-Spam.

2.1 Spambase

En el *UC Irvine Machine Learning Repository* está el conjunto de datos Spambase el cual contiene 4601 correo-e transformados y clasificados en Spam/No-Spam. Ver la descripción de los 57 atributos/columnas asociados a cada mensaje en el archivo *spambase.names*.

Este conjunto se debe dividir aleatoriamente en dos, uno de entrenamiento (90% de las filas) y el otro de test (10% restante). El conjunto de entrenamiento para usará para calcular las probabilidades necesarias para obtener la $P\left(C_{Spam}/\text{ mensaje}_j\right)$ evaluadas en cada $j \in test$

3 Resultados y análisis

- Generar una función que dé la probabilidad de que dado un mensaje se clasifique como Spam (1), dada la matriz de datos de entrenamiento y el vector de clasificación.
- Obtener las probabilidades de clasificación para los mensajes dados en el conjunto *test*.
- Para un valor de *Umbral=0.5*, computar la tabla 2x2 de clasificación correcta/incorrecta (matriz de confusión en terminología IA):

  - Verdadero positivo: El mensaje es positivo y la prueba predijo tambien que era positivo.
  - Verdadero negativo: El valor real es negativo y la prueba predijo tambien que el resultado era negativo.
  - Falso negativo: El valor real es positivo, y la prueba predijo que el resultado es negativo. Esto es lo que en estadística se conoce como error tipo II
  - Falso positivo: El valor real es negativo, y la prueba predijo que el resultado es positivo. Esto es lo que en estadística se conoce como error tipo I

  Identificando Positivo/Negativo como Spam/No-Spam y valores reales la clasificación real del mensaje y valores predicción como los valores que el naive bayes da como clasificación.

- Obtener la matriz de confusión para los valores de Umbral $0.01, 0.02, ..., 0.98, 0.99$. Realizar un gráfico donde el eje de abcisas sea el umbral, y como eje de ordenadas:
  - Proporción de verdaderos positivos
  - Proporción de verdaderos negativos
  - Proporción de mensajes clasificados correctamente.

  A la vista del gráfico anterior, ¿qué conclusiones se pueden obtener para el valor de umbral?

References

[1] Metsis, V. Androutsopoulos, I., Paliouras G. (2006). Spam filtering with Naive Bayes - Which Naives Bayes?. https://www2.aueb.gr/users/ion/docs/ceas2006_paper.pdf