BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

FACULTAD DE CIENCIAS DE LA COMPUTACIÓN

Clasificador Naïve Bayes

EQUIPO:

Alvarez Santos Fredy
Flores Velazquez David
Juárez Escárcega Ilse Lilian
Torres Timal Martín Rafael

Procesamiento del Lenguaje Natural

Dr. David Eduardo Pinto Avendaño

Otoño 2012

Clasificador Naïve Bayes

En términos generales y matemáticos, el teorema de Bayes es de enorme relevancia puesto que vincula la probabilidad de A dado B con la probabilidad de B dado A. Es decir que sabiendo la probabilidad de tener un dolor de cabeza dado que se tiene gripe, se podría saber -si se tiene algún dato más-, la probabilidad de tener gripe si se tiene un dolor de cabeza, muestra este sencillo ejemplo la alta relevancia del teorema en cuestión para la ciencia en todas sus ramas, puesto que tiene vinculación íntima con la comprensión de la probabilidad de aspectos causales dados los efectos observados.
Sea $\{ A_1, A_2, A_3, ..., A_1, ..., A_n \}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero (0). Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B|A_i)$. Entonces, la probabilidad $P(A_i|B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)}$$

donde:
$P(A_i)$ son las probabilidades a priori.
$P(B|A_i)$ es la probabilidad de B en la hipótesis A_i.
$P(A_i|B)$ son las probabilidades a posteriori.

Tomando en cuenta esto podemos para el desarrollo de Naïve Bayes decir que:

$$P(D|C) = \frac{P(F_1...F_n|C)P(C)}{P(F_1...F_n)}$$

Y por regla de la cadena

$$= P(F_1|C)P(F_1|CF_1)...P(F_n|CF_1...F_n - 1)$$

Asumimos ingenuamente

$$= P(F_1|C)P(F_2|C)...P(F_n|C)$$

$$= \prod_{i=1}^{n} P(F_i|C)$$

En este caso usamos el clasificador Naïve Bayes para identificar un texto y a los distintos idiomas al que puede pertenecer.

Código
Naïve Bayes training

```
awk '{
  gsub(/[,_;:()+[\]\x2E-]/, " ", $0);
  for (i=3; i<=NF; i++) frecuencia[tolower($2),tolower($i)]++;
  Total[tolower($2)] = Total[tolower($2)] + (NF - 2);
}
END {
  for (x in frecuencia) {
    split(x, a, SUBSEP);
    print a[1], a[2], frecuencia[x]/Total[a[1]];
  }
} ' $*
```

Línea de comandos para correrlo:

$ bash NBTraining.awk training.col model.out

En donde *training.col* es nuestro corpus de entrenamiento y *model.out* es el archivo que se genera como modelo entrenado, una vez teniendo esto podemos hacer el test de consulta y demostrar que todo corre perfectamente.

Código
Naïve Bayes Test

```
awk '
FILENAME==model{
 modelo[$1,$2]=$3;
 classes[$1]=1;
 next;
}
FILENAME==goldstandard{
 gold[$1]=tolower($2);
 next;
}
BEGIN {
 EPSILON = 0.00001;
}
{
 maximo = 0;
 clase = "NINGUNA";

 for (x in classes) {
  probabilidad=0;
  for (i=2; i<=NF; i++) {
   if ((x,$i) in modelo) probabilidad = probabilidad + log(modelo[x,$i]+1);
   else probabilidad = probabilidad + log(EPSILON+1);
  }
  if (probabilidad > maximo) {
   maximo = probabilidad;
   clase = x;
  }
 }

 print $1, clase, maximo, " : ", $2, $3, $4, $5"...";

 if (gold[$1]==clase) exactitud++;
 else print "<<<<< MAL CLASIFICADO >>>>";
 total++;
}

END{
```

```
  print "-------------------------------------------------------"
  print "El metodo obtuvo una exactitud de " (exactitud/total)*100 "%";
  print "-------------------------------------------------------"
}' model=$1 goldstandard=$2 $*
```

Línea de comandos para correrlo:

$ bash NBTest.awk model.out test.col

En donde usamos al archivo *model.out* generado anteriormente y con un nuevo corpus el cual llamamos *test.col* comprobamos que todo funciona según lo acordado.