Predicción estructura estadística: 2º boletín

Marcos Esteve Casademunt

Enero 2020

Contents

1 Ejercicios teóricos

1.1 Ejercicio 3

El objetivo principal de este ejercicio consiste en calcular la estimación de la regla $Suj \longrightarrow ArtNomAdj$ para las muestras comentadas a continuación. Notar que al introducir una muestra parentizada estamos limitando los posibles árboles de derivación de esa muestra por lo que algunas combinaciones no serán aceptadas.

P_θ((la vieja) (demanda ayuda)) $= 9 * 10^{-4}$
P_θ(la mujer oculta pelea) $= 0.01266$
P_θ(la vieja ayuda) $= 0.007$

$$\overline{p}(Suj \longrightarrow ArtNomAdj) = \frac{\sum_{x \in D} \frac{1}{P_\theta(x)} \sum_{t_x} N(Suj \longrightarrow ArtNomAdj, t_x) P_\theta(x, t_x)}{\sum_{x \in D} \frac{1}{P_\theta(x)} \sum_{t_x} N(Suj, t_x) P_\theta(x, t_x)}$$

$$\frac{\frac{1}{.01266}.01176}{\frac{1}{.0009}(.0009) + \frac{1}{.01266}(.0009 + .01176) + \frac{1}{.007}(.007)} = \frac{.9289}{3} = .3096$$

1.2 Ejercicio 4

A la hora de reestimar la regla $Suj \longrightarrow ArtNomAdj$ por viterbi, podemos utilizar la siguiente ecuación:

$$\overline{p}(Suj \longrightarrow ArtNomAdj) = \frac{\sum_{x \in D} N(Suj \longrightarrow ArtNomAdj, \widehat{t_x})}{\sum_{x \in D} N(Suj, \widehat{t_x})} = \frac{1}{3}$$

Destacar que la estimación de la probabilidad es mas sencilla ya que únicamente se considera el mejor árbol de derivación para cada muestra y se realizan conteos sobre este. Además destacar, que la probabilidad obtenida no difiere en gran medida de la obtenida en el ejercicio 3 consiguiendo una estimación buena con una menor complejidad computacional.

1.3 Ejercicio 5

El objetivo principal de este ejercicio consiste en calcular la estimación de la regla $Suj \longrightarrow ArtNomAdj$ para las muestras comentadas en el enunciado. Como se puede observar la muestra "La vieja mujer oculta demanda ayuda" no puede ser generada por la gramática comentada en el enunciado del problema. Para solventar ese problema se ha decido crear una nueva regla y reasignar las probabilidades como se muestra a continuación:

$0.4 Suj \longrightarrow ArtNom$
$0.2 Suj \longrightarrow ArtAdjNom$
$0.2 Suj \longrightarrow ArtNomAdj$
$0.2 Suj \longrightarrow ArtAdjNomAdj$

Con las nuevas reglas podemos reestimar las probabilidades como:

P_θ(la vieja demanda ayuda) $= 1 * .1 * .4 * .3 * .2 * 0.3 + 1 * .3 * .2 * .2 * .2 * .7 =$

$7.2 * 10^{-4} + 1.68 * 10^{-3} = 2.4 * 10^{-3}$

$P_\theta($ la mujer oculta pelea$) = 1 * .3 * .4 * .1 * .2 * 0.3 + 1 * .3 * .7 * .2 * .4 * .7 = 7.2 * 10^{-4} + 0.01176 = 0.01248$

$P_\theta($ la vieja mujer oculta demanda ayuda $) = 1 * .3 * .3 * .7 * .2 * .3 * .2 * .3 = 2.268 * 10^{-4}$

A continuación se muestra el único árbol de derivación que existe para la cadena "la vieja mujer oculta demanda ayuda"

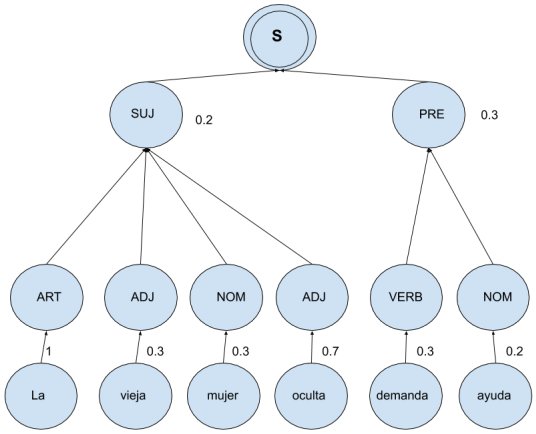


Figure 1: Árbol de derivación para la secuencia "La vieja mujer oculta demanda ayuda"

Por último podemos re-estimar la probabilidad de la regla mediante inside-outside:

$$\overline{p}(Suj \longrightarrow ArtNomAdj) = \frac{\sum_{x \in D} \frac{1}{P_\theta(x)} \sum_{t_x} N(Suj \longrightarrow ArtNomAdj, t_x) P_\theta(x, t_x)}{\sum_{x \in D} \frac{1}{P_\theta(x)} \sum_{t_x} N(Suj, t_x) P_\theta(x, t_x)}$$

$$\frac{\frac{1}{.01248} .01176}{\frac{1}{2.4*10^{-3}}(7.2 * 10^{-4} + 1.68 * 10^{-3}) + \frac{1}{.01248}(7.2 * 10^{-4} + .01176) + \frac{1}{2.268*10^{-4}}(2.268 * 10^{-4})} =$$

$$\frac{.9423}{3} = .3141$$

1.4 Ejercicio 6

Para la realización de este ejercicio es necesario obtener los 2 mejores árboles de derivación y realizar el mismo proceso de arriba con los dos mejores árboles de derivación. Como en ningún caso el número de arboles de derivación es superior a 2 el resultado del ejercicio será el mismo que el ejercicio 5.

2 Ejercicios prácticos

2.1 Ejercicio 8

Tal y como se puede observar en la tabla 1, se observa una tendencia a aumentar el número de triángulos rectángulos al aumentar el número de no terminales. Esto se puede deber a que al aumentar el número de no terminales permitimos un mayor número de reglas, permitiendo de esta forma, obtener modelos que se adapten de una forma adecuada a la estructura de dichos triángulos.

Table 1: Evolución del número de triángulos rectángulos al incrementar el número de terminales

# terminales	**# rectángulos**
5	29
10	63
15	61
20	84

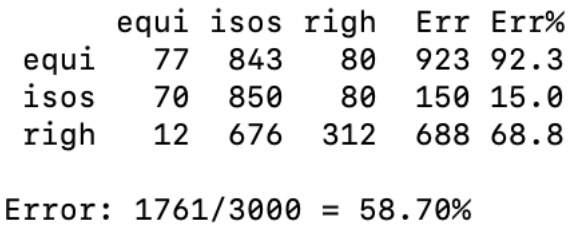
2.2 Ejercicio 9

```
      equi isos righ  Err Err%
equi   794  206    0  206 20.6
isos   531  225  244  775 77.5
righ   108  145  747  253 25.3

Error: 1234/3000 = 41.13%
```

Figure 2: Matriz de confusión calculada con los ejemplos con corchetes y inside-outside

A la vista de los resultados expuestos en la matriz de confusión superior, se observa que el algoritmo entrenado por inside-outside tiene una mejor tasa de acierto en el reconocimiento de triángulos equiláteros y rectángulos que en el reconocimiento de triángulos isósceles.

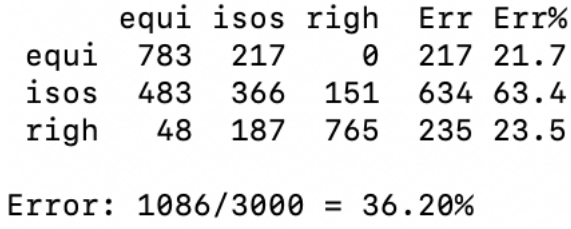


	equi	isos	righ	Err	Err%
equi	77	843	80	923	92.3
isos	70	850	80	150	15.0
righ	12	676	312	688	68.8

Error: 1761/3000 = 58.70%

Figure 3: Matriz de confusión calculada con los ejemplos con corchetes y viterbi

Como podemos ver en los resultados expuestos en la matriz de confusión superior, se observa que el algoritmo entrenado por viterbi tiene una mejor tasa de acierto en el reconocimiento de triángulos isósceles mientras que en los triángulos equiláteros y rectángulos comete un mayor número de fallos. Como podemos observar, el entrenamiento por viterbi podría complementar el entrenamiento de inside-outside mejorando así la tasa de reconocimiento.

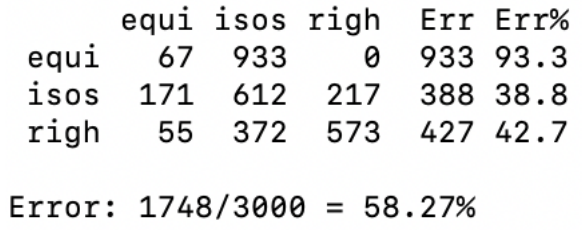


	equi	isos	righ	Err	Err%
equi	783	217	0	217	21.7
isos	483	366	151	634	63.4
righ	48	187	765	235	23.5

Error: 1086/3000 = 36.20%

Figure 4: Matriz de confusión calculada con los ejemplos sin corchetes y inside-outside

Como podemos ver al eliminar los brackets conseguimos mejorar la tasa de acierto al entrenar por inside-outside donde se observa por ejemplo, una mejor tasa de acierto en el reconocimiento de triángulos isósceles



	equi	isos	righ	Err	Err%
equi	67	933	0	933	93.3
isos	171	612	217	388	38.8
righ	55	372	573	427	42.7

Error: 1748/3000 = 58.27%

Figure 5: Matriz de confusión calculada con los ejemplos sin corchetes y viterbi

Por último, si entrenamos por viterbi y sin brackets se empeora el reconocimiento de triángulos equiláteros y isósceles pero se consigue mejorar el

reconocimiento de triángulos rectángulos frente a viterbi con brackets