Predicción estructurada estadística, parte 1

Marcos Esteve Casdemunt

Diciembre 2019

Contents

1 Ejercicio 1

Un modelo disciminativo busca aprender la diferencia entre las entradas y las diferentes salidas. Un ejemplo de los modelos discriminativos podría ser las maquinas de vectores soporte en un problema de clasificación.

La principal diferencia con un modelo generativo, radica en que este trata de estimar la probabilidad conjunta p(x,y), proporcionando de esta forma un modelo que explica como los datos se generan. Un ejemplo de un modelo generativo podría ser las gramáticas probabílisticas que hemos estudiado en las clases neuronales o las redes neuronales.

2 Ejercicio 4

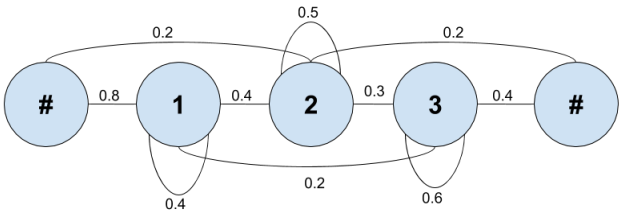

Figure 1: Modelo de Markov asociado al ejercicio propuesto

Un posible arbol de derivación para la cadena a b a podría ser:

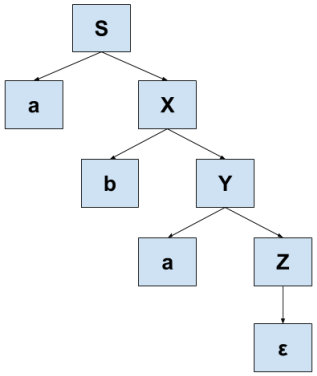

Figure 2: Árbol de derivación para la cadena a b a

La probabilidad de que se genere la cadena a b a es por tanto:

$$P(aba) = 0.8 * 0.2 * 0.4 * 0.2 * 0.3 * 0.5 * 0.4$$

3 Ejercicio 5

El objetivo de este ejercicio radica en conseguir el número de arboles de derivación distintos que tiene una gramática para una cadena de entrada. Para ello, podemos fijarnos en la diapositiva 24 del tema 1.

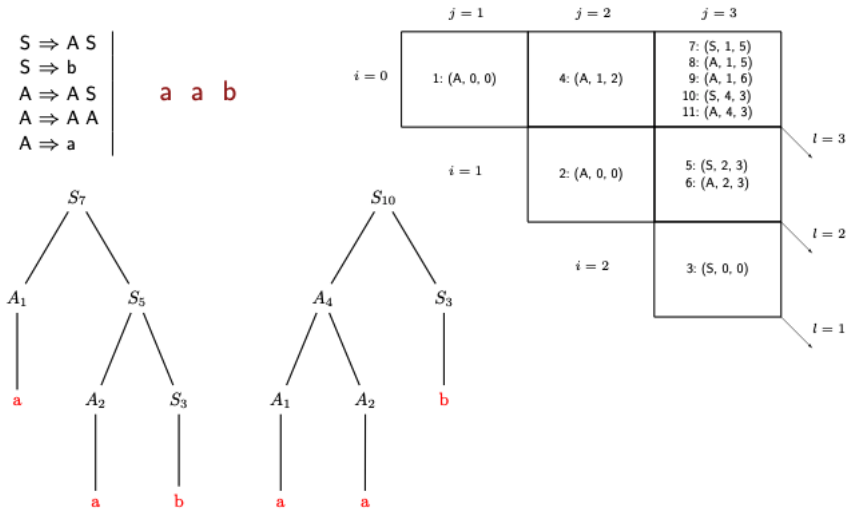

Figure 3: Aplicación del algoritmo CYK

El algoritmo CYK permite que dada una cadena de entrada y una gramática obtener si la cadena pertenece al lenguaje generado por la gramática. Para ello, tal y como se puede observar en la Figura 4, se construye una tabla con los resultados de las producciones y si, en la esquina superior derecha de esa tabla se encuentra el axioma de la gramática entonces la cadena pertenece al lenguaje generado por dicha gramática.

Input: $G = (N, \Sigma, \mathcal{P}, S)$ in FNC and $\mathbf{x} = x_1 \dots x_T \in \Sigma^*$
Output: Parsing table $t[i,j]$ $(1 \leq i, j \leq T)$;
$A \in t[i, i+l]$ **iff** $A \stackrel{*}{\Longrightarrow} x_{i+1} \dots x_{i+l}$

for $i : 0 \dots T-1$ **do**
  $t[i,i+1] = t[i,i+1] \cup \{A : (A \to b) \in P;\ b = x_{i+1}\}$

for $l : 2 \dots T$ **do**
  for $i : 0 \dots T-l$ **do**
    for $k : 1 \dots l-1$ **do**
      $t[i,i+l] = t[i,i+l] \cup \{A : (A \to BC) \in P$;
      $B \in t[i,i+k];\ C \in t[i+k,i+l]\}$;

if $S \in t[0,T]$ **then** $x \in L(G)$ **else** $x \notin L(G)$;

Figure 4: Algoritmo CYK

Si sustituimos el ultimo if por:

$$count(S \in t[0,T])$$

Conseguimos que si el axioma no se encuentra en la esquina superior derecha del trellis se devuelva 0 y sino devolver la suma de veces que aparece el axioma de la gramática en dicha posición de la tabla.

4 Ejercicio 6

Dado una secuencia de entrada $x = x_1, ..., x_T$ y $A \in N$

En primer lugar, la inicialización seguirá siendo la misma:

$$\forall A \in N \quad \forall i : 0...T-1$$

$$e(A < i, i+1 >) = p(A \to B)\delta(b, x_{i+1})$$

La recursión $\forall A \in N \quad \forall l : 2...T \quad \forall i : 0...T-l$ pasará a ser:

$$e(A < i, i+L >) = \max_{B,C \in N} P(A \to BC) * e(B < i, i+1 >) * e(C < i+1, i+l >)$$

y por último el resultado final será:

$$P_\theta(x) = e(S, 0, T)$$

el coste temporal de este algoritmo sera (siendo n la longitud de la secuencia o cadena):

$$O(n^2)$$

5 Ejercicio Práctico

Se propone realizar una evaluación utilizando el toolkit SCFG a una tarea de reconocimiento de triángulos equiláteros, isósceles y escalenos.

Para ello se proporcionan 3 modelos distintos G1, G2 y G3 cada uno dividido en tres submodelos distintos para los distintos tipos de triángulos. Además se proporciona un corpus de test para cada tipo de triangulo con 1000 muestras cada uno.

5.1 Perplejidad

En esta primera tarea se propone realizar un análisis de la perplejidad de los distintos modelos. Para ello se ha realizado un analisis comparando las perplejidades de todas las posibles combinaciones de submodelos con corpus de test.

Podemos entender la perplejidad de forma intuitiva como cuanto se "sorprende" un determinado modelo al suministrarle un conjunto de muestras que no ha visto durante el entrenamiento. Además guarda una relación con como de compleja es la tarea para el modelo.

Table 1: Análisis de perplejidades para el modelo G1

submodelo	**testSet**	**perplejidad**
G1-EQ	TS-EQ	99095.20
G1-EQ	TS-IS	191.88
G1-EQ	TS-SC	245.91
G1-IS	TS-EQ	47334.82
G1-IS	TS-IS	26581.81
G1-IS	TS-SC	21621.97
G1-SC	TS-EQ	48369.44
G1-SC	TS-IS	27508.97
G1-SC	TS-SC	33618.35

A la vista de los resultados en la tabla superior se observa que el modelo G1 manifiesta una alta perplejidad en todos los casos salvo en en el caso de G1-EQ con los test de isosceles y escalenos.

Table 2: Análisis de perplejidades para el modelo G2

submodelo	**testSet**	**perplejidad**
G2-EQ	TS-EQ	635136.42
G2-EQ	TS-IS	389521.31
G2-EQ	TS-SC	520221.73
G2-IS	TS-EQ	102498.66
G2-IS	TS-IS	52217.59
G2-IS	TS-SC	49786.00
G2-SC	TS-EQ	68152.55
G2-SC	TS-IS	42723.11
G2-SC	TS-SC	43543.30

A la vista de los resultados en la tabla superior se observa que el modelo G2 manifiesta una alta perplejidad en todos los casos, por lo que para este modelo la tarea de clasificación resulta más compleja

Table 3: Análisis de perplejidades para el modelo G3

submodelo	**testSet**	**perplejidad**
G3-EQ	TS-EQ	986.20
G3-EQ	TS-IS	550208.12
G3-EQ	TS-SC	1081980.36
G3-IS	TS-EQ	1283.37
G3-IS	TS-IS	1254.89
G3-IS	TS-SC	1218.74
G3-SC	TS-EQ	1272.72
G3-SC	TS-IS	1258.35
G3-SC	TS-SC	1218.04

A la vista de los resultados en la tabla superior se observa que el modelo G3 manifieste perplejidades más reducidas que los anteriores modelos salvo los casos del submodelo G3-EQ con los corpus de isósceles y escalenos. Será por tanto el modelo que se comporte mejor de los tres propuestos.

5.2 Análisis de la matriz de confusión

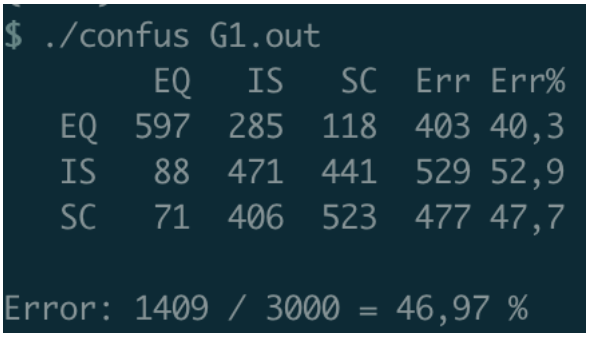

Figure 5: Matriz de confusión para el modelo G1

A la vista de los resultados expuestos en la figura 5 se observa que el modelo G1 tiene problemas a la hora de detectar las muestras de triángulos isosceles ya que en muchos casos las confunde con escalenos, lo mismo ocurre con la detección de los triángulos escalenos, los cuales son confundidos con triángulos isósceles.

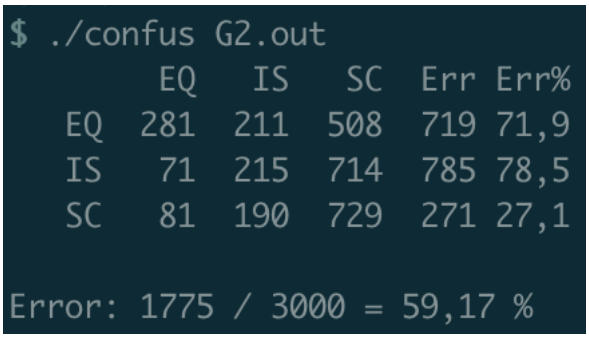

Figure 6: Matriz de confusión para el modelo G2

El modelo G2 tal y como se puede observar en la figura 6 confunde en gran medida los triángulos equiláteros y isósceles con escalenos. Esto hace que la tasa de error del modelo sea elevada ya que tiene una alta predisposición a que la clase asignada sea la de triangulo escaleno.

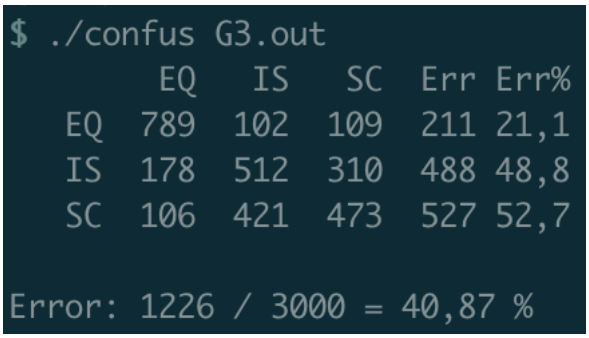

Figure 7: Matriz de confusión para el modelo G3

Por último el modelo G3, tal y como puede observarse en la figura 7 es el que mejor se comporta ya que es el que menor tasa de error posee. Aun siendo el mejor modelo la tasa de error sigue siendo elevada y como se puede observar confunde gran cantidad de las muestras de isósceles con escalenos y los escalenos con isósceles.