Teoria de la Informació GCED-UPC curs 2019/20
Problemes; full número 1

6 de setembre de 2019

L'objectiu d'aquesta pràctica és estudiar propietats estocàstiques de textos des del punt de vista de la teoria de la informació[1]. Un "text" serà aquí una cadena de caràcters d'un alfabet finit. Es considera una mostra representativa[2] d'un determinat llenguatge: una novel·la representa la llengua en què està escrita; els píxels d'una fotografia representen una imatge; el genoma d'un individu representa la genètica d'una espècie, etc.

Per tal d'evitar problemes al començament i al final en considerar lletres consecutives se suposarà que, en acabar-se, el text torna a començar pel començament; això equival a treballar amb textos infinits obtinguts repetint el text donat indefinidament.

A Atenea trobareu fitxers amb textos en català, castellà, anglès, francès, alemany i italià, simplificats per tal que l'alfabet es redueixi a les lletres i l'espai. També un text en castellà sense simplificar, amb majúscules i minúscules, accents, signes de puntuació i caràcters de control. Useu aquests textos per fer els problemes. També podeu usar-ne altres que us semblin interessants.

En la majoria de problemes haureu de fer servir `Python`. En els enunciats se suposa que el text a estudiar s'ha posat en una variable `txt` de tipus string. Convé que us familiaritzeu amb aquest tipus de variables buscant funcions, mètodes i mòduls per treballar amb elles.

Es denota $\mathcal{X}$ el conjunt dels caràcters (lletres) que apareixen a `txt` i $\mathcal{X}^n$ el conjunt de les paraules de longitud n escrites amb lletres de $\mathcal{X}$. Es consideren les variables aleatòries següents:

- X pren valors a $\mathcal{X}$ i correspon a agafar aleatòriament una lletra de `txt`;
- $\boldsymbol{X}_n$ pren valors a $\mathcal{X}^n$ i correspon a agafar aleatòriament n lletres consecutives de `txt` (recordi's que al final el text torna a començar pel principi);
- X_k, a valors en $\mathcal{X}$, és la lletra k-èsima de $\boldsymbol{X}_n$;
- Y_r i Z_s, a valors en $\mathcal{X}^r$ i $\mathcal{X}^s$ respectivament, són les primeres r i les últimes s lletres de $\boldsymbol{X}_n$, amb $r+s=n$.

Així, per exemple, les variables Y_1 i Z_1 corresponen a agafar dues lletres consecutives i les variables Y_r i Z_1 corresponen a agafar r lletres i la lletra que hi ha a continuació. La variable $\boldsymbol{X}_2$ és la concatenació de Y_1 i Z_1 i la variable $\boldsymbol{X}_{r+1}$ és la concatenació de Y_r i Z_1.

[1] Podeu llegir l'article fundacional de Shannon Prediction and entropy of printed English

[2] Vegeu ergodicitat i també corpus.

1.1. Compareu les distribucions de probabilitat de X i de X_k per a diferents valors de k.

1.2. Compareu les distribucions de probabilitat de Y_r i Z_s quan $r = s$.

1.3. Considereu els textos següents sobre l'alfabet binari $\mathcal{X} = \{\texttt{0}, \texttt{1}\}$:

1. $\texttt{txt} = \texttt{0101}\cdots\texttt{0101}$ de longitud $2n$;
2. $\texttt{txt} = \texttt{00110011}\cdots\texttt{0011}$ de longitud $4n$;
3. $\texttt{txt} = \texttt{000111000111}\cdots\texttt{000111}$ de longitud $6n$;
4. el mateix amb blocs alternats de zeros i uns de mida k en lloc de 2 o 3;
5. $\texttt{txt} = \texttt{0111}\cdots\texttt{1}$ de longitud n;
6. $\texttt{txt} = \texttt{010011000111}\ \cdots\ \texttt{0}\overset{\texttt{n)}}{\cdots}\texttt{01}\overset{\texttt{n)}}{\cdots}\texttt{1}$ (quina longitud té?);
7. $\texttt{txt} = \texttt{010010001}\cdots\texttt{0}\overset{\texttt{n)}}{\cdots}\texttt{01}$ (quina longitud té?);
8. $\texttt{txt} = \texttt{11.}0010010000111111011010101000100\cdots$ (representació binària de π).

Per a cadascun d'ells calculeu la distribució de probabilitats del vector $\boldsymbol{X}_2 = (Y_1, Z_1)$, les distribucions marginals de Y_1 i Z_1 i la taula de les probabilitats condicionades $Z_1|Y_1$.

1.4. Implementeu una funció que dibuixi l'histograma d'una variable aleatòria discreta; per exemple, en forma de diagrama de barres.

1. Dibuixeu els histogrames de la variable X i compareu-los per a textos en llengües diferents;
2. estudieu la divergència de les distribucions de probabilitat corresponents respecte la distribució uniforme i entre llengües diferents;
3. compareu els histogrames de les variables aleatòries $Z_1|Y_1 = y$ per a diferents lletres $y \in \mathcal{X}$ i també de les variables $Y_1|Z_1 = z$.

Quines conclusions en traieu sobre diferències entre els diferents idiomes? Què proposeu per detectar en quin idioma està escrit un text?

1.5. Considereu la variable aleatòria que pren valors en les paraules dels textos, en comptes de les lletres. Podeu separar els textos en paraules amb el mètode `split`.

- Compareu longituds mitjanes de paraules en idiomes diferents.
- Busqueu informació sobre la llei de Zifp, i discutiu fins a quin punt les variables s'ajusten a aquesta distribució.

1.6. Considereu la variable aleatòria que només té en compte si les lletres són vocals o consonants. Una manera senzilla de fer això és construir un nou text substituint cada vocal per un mateix símbol i cada consonant per un altre símbol. Estudieu i compareu les probabilitats i les probabilitats condicionades de vocals, consonants i espais en textos d'idiomes diferents. Quin idioma té més proporció de vocals? En quin és més probable que després d'una consonant en vingui una altra? Quin té seqüències de vocals seguides més llargues? i de consonants?

1.7. Programeu una funció `minr(txt)` que calculi el valor mínim de r tal que la variable Z_1 és funció de la variable Y_r. Podeu usar `Counter` del mòdul `collections`.

1.8. Programeu una funció `crea_text(txt,n,N)` que creï nous textos amb les mateixes propietats estocàstiques que el text donat `txt` fins a ordre `n`: les paraules de `n` lletres dins del nou text es generen seguint la distribució de probabilitats de la variable $\boldsymbol{X}_{\mathtt{n}}$. El paràmetre `N` indica la longitud del text que s'ha de crear. Podeu usar el mètode `choices` del mòdul `random`.

O sigui, per a $\mathtt{n} = 1$ les lletres del nou text s'obtenen avaluant la variable X i per tant apareixeran amb la mateixa probabilitat que a `txt` (aproximadament, si `N` és gran); en canvi en els textos creats així els parells de lletres consecutius no obeiran la variable $\boldsymbol{X}_2$. Per a $\mathtt{n} = 2$ els parells de lletres consecutives del nou text tindran la mateixa probabilitat que tenen a `txt` (aproximadament), però ternes consecutives no tindran la distribució de $\boldsymbol{X}_3$, etc.

1.9. Demostreu que hi ha un `n` tal que la funció del problema anterior (amb `N` igual a la longitud de `txt`) dóna com a resultat sempre el mateix text `txt`, i digueu quin és el mínim valor de `n` amb aquesta propietat.

1.10. Implementeu una funció `entr(txt,n)` que retorna l'entropia $H(\boldsymbol{X}_{\mathtt{n}})$ de la variable aleatòria $\boldsymbol{X}_{\mathtt{n}}$. Recordeu que l'entropia d'una variable aleatòria X és l'esperança de la variable $-\log_2 p(X)$. Compareu entropies en diferents idiomes.

1.11. Implementeu una funció `cond_entr_word(txt,n,w)` que retorna l'entropia condicionada $H(Z_n|Y_r = w)$ on w és una paraula de longitud r. Compareu entropies condicionades a diferents paraules i traieu-ne conclusions.

Implementeu una funció `cond_entr(txt,r,s)` que retorna l'entropia $H(Z_s|Y_r)$ condicionada de la variable Z_s respecte la variable Y_r. Compareu per a diferents idiomes.