On the Uniform Distribution of Strings

Sébastien Rebecchi and Jean-Michel Jolion

PSC 2008, Prague, Czech Republic
September 2, 2008

Introduction

How to describe the data?

- Structures:
  - $+$ representational capabilities,
  - $-$ lack of mathematical tools;
- feature vectors:
  - $+$ powerful statistical algorithms,
  - $-$ representational capabilities;

$\Longrightarrow$ reconcile the two approaches;

$\Longrightarrow$ need to define a statistical characterization of spaces of structures.

We introduce the uniform distribution of strings.

Notations

- A alphabet;
- $|A|$ cardinal of A.

Notations

- $|X|$ length of the string X over A;
- A^n set of strings of length n over A;
- $A^{\leqslant n}$ set of strings of length at most n over A;
- X_i i-th letter of X.

Uniform distribution of strings

First approach: equiprobability.

- U over A^n = concatenation of n U over A

$$\mathrm{P}(X) = |A|^{-n};$$

- generation in $O(n)$;
- probability in $O(1)$;
- preservation under concatenation:

$$(X \sim U \text{ over } A^n) \wedge (l \sim U \text{ over } A) \Longrightarrow Xl \sim U \text{ over } A^{n+1}.$$

Uniform distribution of strings

Second approach: normalized measure.

- Let S be a set;
- $E \subseteq S$:

$$P(E) = \mu(E)/\mu(S);$$

- examples:
  - $S \subset \mathbb{N}$, S finite, $\mu =$ cardinality,
  - $S \subset \mathbb{R}$, S bounded, $\mu =$ Lebesgue measure.

σ-algebra

σ-algebra over S = set of subsets of S that is

- non empty;
- closed under complements;
- closed under countable unions.

If S is countable, then powerset(S) is the only σ-algebra over S containing all singletons $\{x\}, x \in S$.

Measure

Measure μ over σ = function $\sigma \to \mathbb{R}^+ \cup \{\infty\}$ that is

- 0 for $\{\}$;
- additive under countable disjoint unions.

$\mu(\{x\}) =_{\text{notation}} \mu(x)$, $\{x\} \in \sigma$.

Uniform distribution

Uniform distribution w.r.t. μ: $\forall E \in \sigma$:

$$\mathsf{P}(E) = \mu(E)/\mu(S).$$

$\mathsf{P}(\{x\}) =_{\text{notation}} \mathsf{P}(x)$, $\{x\} \in \sigma$.

String measure

- $\lambda \notin A$ denotes the empty letter;
- we assume the measure μ_A over powerset$(A \cup \{\lambda\})$;
- for $n \in \mathbb{N}$, we define the measure μ^n over powerset$(A^{\leqslant n})$.

String measure

- String of length at most n over A = cannonical representation of a set of sequences composed of n elements of $A \cup \{\lambda\}$;
- example:
  - $A = \{a, b\}$,
  - $n = 3$,
  - ab "=" $\{\lambda ab, a\lambda b, ab\lambda\}$.

String measure

$\forall X \in A^{\leqslant n}$:

$$\mu^n(X) = \binom{n}{|X|} \times \prod_{i=1}^{|X|} \mu_A(X_i) \times \mu_A(\lambda)^{n-|X|}.$$

Probability

Total measure:

$$\mu^n(A^{\leqslant n}) = \mu_A(A \cup \{\lambda\})^n.$$

$\Longrightarrow$ Probability of a string in $O(n)$.

Preservation, generation

Preservation under concatenation:

$$(X \sim U \text{ w.r.t. } \mu^n) \land (l \sim U \text{ w.r.t. } \mu_A) \Longrightarrow Xl \sim U \text{ w.r.t. } \mu^{n+1}.$$

$\Longrightarrow$ Generation of a string in $O(n)$.

Generation

Input: $n \in \mathbb{N}$.
Output: A string uniform w.r.t. μ^n.
begin
    $\mathrm{P}_A \longleftarrow$ uniform distribution w.r.t. μ_A: $\forall l \in A \cup \{\lambda\}$:

$$\mathrm{P}_A(l) = \mu_A(l) / \mu_A(A \cup \{\lambda\});$$

    $X \longleftarrow$ empty string;
    for $i \longleftarrow 1$ **à** n **do**
        $l \longleftarrow$ random choice according to P_A;
        $X \longleftarrow Xl$;
    end
    return X;
end

Unification

First approach = second approach with:

- $\mu_A(\lambda) = 0$ ($\Longrightarrow \mathrm{P}^n(A^{\leqslant n-1}) = 0$ if $n > 0$);
- $\mu_A(l) = \mu_A(m), \forall l, m \in A$.

Conclusion

- Uniform string = concatenation of uniform letters;
- simple but relevant measure;
- easy to extend to ordered trees;
- statistical test;
- how to *sum* for CLT?