Ling 5801: Lecture Notes 12
Probability

Weights for our parsers and other models are well defined as **probabilities**.

Probability in this view is a subjective measure of belief about some uncertain event (e.g. sentence).

Specifically, a **probability** p is a belief about a set of **outcomes** o in a **sample space** O.

Sample space: set of mutually exclusive possible propositions (e.g. FSA states / PDA store-states)

Belief: given an infinite number of trials of O, the set of o would happen p of the time.

Contents

12.1 Probability and probability spaces [Kolmogorov, 1933]

Probability is defined over a measure space $\langle O, \mathcal{E}, \mathsf{P} \rangle$ where the measure P (probability) sums to one.

This **probability measure space** $\langle O, \mathcal{E}, \mathsf{P} \rangle$ consists of:

1. a **sample space** O – a non-empty set of **outcomes**;
2. an **event space** ('sigma-algebra') $\mathcal{E} \subseteq 2^O$ – a set of **events** in the power set of O such that:
   (a) $\mathcal{E}$ contains O: $O \in \mathcal{E}$,
   (b) $\mathcal{E}$ is closed under complementation: $\forall_{A \in \mathcal{E}}\ O - A \in \mathcal{E}$,
   (c) $\mathcal{E}$ is closed under countable union: $\forall_{A_1..A_\infty \in \mathcal{E}} \bigcup_{i=1}^{\infty} A_i \in \mathcal{E}$

   (this set of events will serve as the **domain** of our probability function);
3. a **probability measure** $\mathsf{P} : \mathcal{E} \to \mathbb{R}_0^\infty$ – a function from events to non-negative reals such that:
   (a) the P measure is countably additive: $\forall_{A_1..A_\infty \in \mathcal{E} \text{ s.t. } \forall_{i,j}\ A_i \cap A_j = \emptyset}\ \mathsf{P}(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \mathsf{P}(A_i)$,
   (b) the P measure of entire space is one: $\mathsf{P}(O) = 1$.

These are the **Kolmogorov axioms of probability**.

This characterization is helpful because it unifies probability spaces that may seem very different:

1. **discrete** spaces – e.g. a coin:

   $\langle \underbrace{\{\mathrm{H},\mathrm{T}\}}_{O}, \underbrace{\{\emptyset,\{\mathrm{H}\},\{\mathrm{T}\},\{\mathrm{H},\mathrm{T}\}\}}_{\mathcal{E}}, \underbrace{\{\langle\emptyset,0\rangle,\langle\{\mathrm{H}\},.5\rangle,\langle\{\mathrm{T}\},.5\rangle,\langle\{\mathrm{H},\mathrm{T}\},1\rangle\}}_{\mathsf{P}} \rangle$

2. **continuous** spaces – e.g. a dart (here $2^{\mathbb{R}^2}$ is a Borel algebra: a set of all open subsets of $\mathbb{R}^2$):

   $\langle \underbrace{\mathbb{R}^2}_{O}, \underbrace{2^{\mathbb{R}^2}}_{\mathcal{E}}, \underbrace{\{\langle R,p\rangle \mid R \in 2^{\mathbb{R}^2}, p = \iint_{A\in R} \mathcal{N}_{\mathbf{0},\mathbf{1}}(x_A,y_A)\, dA\}}_{\mathsf{P}} \rangle$

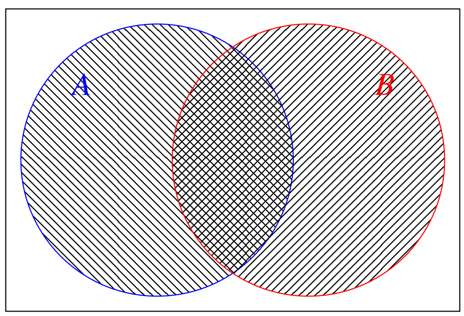
   (events must be open sets/ranges of outcomes because point outcomes have zero probability)

3. **joint** spaces using Cartesian products of sample spaces – e.g. two coins ($\{\mathrm{H},\mathrm{T}\} \times \{\mathrm{H},\mathrm{T}\}$):

   $\langle \underbrace{\{\mathrm{HH},\mathrm{HT},\mathrm{TH},\mathrm{TT}\}}_{O}, \underbrace{\{\emptyset,\{\mathrm{HH}\},\ldots,\{\mathrm{HH},\mathrm{HT},\mathrm{TH},\mathrm{TT}\}\}}_{\mathcal{E}}, \underbrace{\{\langle\emptyset,0\rangle,\langle\{\mathrm{HH}\},.25\rangle,\ldots,\langle\{\mathrm{HH},\mathrm{HT},\mathrm{TH},\mathrm{TT}\},1\rangle\}}_{\mathsf{P}} \rangle$

This axiomatization entails, for any events (sets of outcomes) $A, B \in \mathcal{E}$:

1. $\mathsf{P}(A) \in \mathbb{R}_0^1$
2. $\mathsf{P}(A \cup B) = \mathsf{P}(A) + \mathsf{P}(B) - \mathsf{P}(A \cap B)$



Minimal events – those used as base cases in the closure operations – are called **atomic events**.

Atomic events in continuous models can have any size you want (like even/odd die), but not points.

12.2 Probability notation

Though probabilities are defined over sets of outcomes, we often write them using **propositions**.

For example, if $O = X \times Y$ and therefore $\forall_{o\in O}\; o = \langle x_o, y_o\rangle$:

$$\begin{array}{lll l}
\mathsf{P}(x) & = \mathsf{P}(X{=}x) & = \mathsf{P}(\,\{o \mid o{\in}O \wedge x_o{=}x\}\,) & \text{(allow any value for } y_o \text{ component)} \\
\mathsf{P}(x \wedge y) & = \mathsf{P}(X{=}x \wedge Y{=}y) & = \mathsf{P}(\,\{o \mid o{\in}O \wedge x_o{=}x \wedge y_o{=}y\}\,) & \\
\mathsf{P}(\neg x) & = \mathsf{P}(X{\neq}x) & = \mathsf{P}(\,\{o \mid o{\in}O \wedge x_o{\neq}x\}\,) &
\end{array}$$

Random variables are functions from outcomes x_o, y_o to **values**, e.g. distance of point to origin.

Often we will simply use Cartesian factors of a joint sample space (X, Y) as random variables.

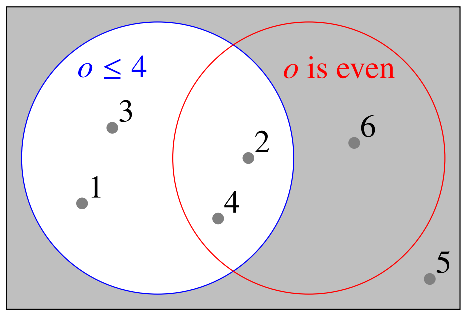
Distributions are sometimes written as probabilities over (all values of) random variables:

$$\mathsf{P}(X) = \mathsf{P}(Y) \quad \Leftrightarrow \quad \forall_{x \in X \cup Y}\ \mathsf{P}(X{=}x) = \mathsf{P}(Y{=}x).$$

We can also define **conditional probabilities** as ratios of these measures: $\mathsf{P}(x \mid y) = \frac{\mathsf{P}(x \wedge y)}{\mathsf{P}(y)}$.

(It's the probability of the joint $\{o \mid x_o{=}x\} \cap \{o \mid y_o{=}y\}$ over the probability of the condition $\{o \mid y_o{=}y\}$.)

For example, if we have $O = \{1, 2, 3, 4, 5, 6\}$, then $\mathsf{P}(o \text{ is even} \mid o \leq 4) = \frac{\mathsf{P}(o \text{ is even } \wedge\ o \leq 4)}{\mathsf{P}(o \leq 4)} = \frac{2}{4} = \frac{1}{2}$.



12.3 Estimating probabilities from data

We can estimate these probabilities from data!

First, define a **frequency space** $\langle O, \mathcal{E}, \mathsf{F} \rangle$ – same measure space with no $\mathsf{P}(O) = 1$ constraint.

We can define a frequency space using **counts** of some set of atomic events in some **training data**.

For example a model of sentence expansions ($O = \mathit{Root} \times \mathit{LeftChild} \times \mathit{RightChild}$ in a tree):

$\langle\ \{\langle$**V,N,V-aN**$\rangle, \langle$**V,N,V-gN**$\rangle, \langle$**V,R-aN,V**$\rangle, \langle$**S,V,R-aN**$\rangle, \ldots\},$
$\{\emptyset, \{\langle$**V,N,V-aN**$\rangle\}, \{\langle$**V,N,V-gN**$\rangle\}, \{\langle$**V,R-aN,V**$\rangle\}, \{\langle$**S,V,R-aN**$\rangle\}, \ldots\}$
$\{\langle\emptyset, 0\rangle, \langle\{\langle$**V,N,V-aN**$\rangle\}, 2\rangle, \langle\{\langle$**V,N,V-gN**$\rangle\}, 1\rangle, \langle\{\langle$**V,R-aN,V**$\rangle\}, 0\rangle, \langle\{\langle$**S,V,R-aN**$\rangle\}, 2\rangle, \ldots\}\ \rangle$

(Counts for larger sets are simply sums, according to axiom 3a.)

We can now define a very simple probability model (probability space) based on these counts:

$$\mathsf{P}(A) = \frac{\mathsf{F}(A)}{\mathsf{F}(O)}$$

$\langle\ \{\langle$**V,N,V-aN**$\rangle, \langle$**V,N,V-gN**$\rangle, \langle$**V,R-aN,V**$\rangle, \langle$**S,V,R-aN**$\rangle, \ldots\},$
$\{\emptyset, \{\langle$**V,N,V-aN**$\rangle\}, \{\langle$**V,N,V-gN**$\rangle\}, \{\langle$**V,R-aN,V**$\rangle\}, \{\langle$**S,V,R-aN**$\rangle\}, \ldots\}$
$\{\langle\emptyset, 0\rangle, \langle\{\langle$**V,N,V-aN**$\rangle\}, .4\rangle, \langle\{\langle$**V,N,V-gN**$\rangle\}, .2\rangle, \langle\{\langle$**V,R-aN,V**$\rangle\}, 0\rangle, \langle\{\langle$**S,V,R-aN**$\rangle\}, .4\rangle, \ldots\}\ \rangle$

(Counts for larger sets are simply sums, according to axiom 3a.)

This is called **relative frequency estimation**.

Probabilities of grammar rule expansions are more commonly notated:

$\mathsf{P}(c \to d\; e \,|\, c)$ probability speaker decided to expand c into d followed by e

It is a **branching process model** that assigns probability to any tree / sentence

These are/were widely used in computational linguistics.

References

[Kolmogorov, 1933] Kolmogorov, A. N. (1933). *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer. Second English Edition, *Foundations of Probability* 1950, published by Chelsea, New York.