Product Space Embeddings

Hutter RNN Project

2026-01-31

1 Construction

Given two event spaces A (size m) and B (size n), we construct an embedding of A into $\mathbb{R}^n$ via the product space.

1.1 Recipe

1. **Product space:** Form $A \times B$ with $m \times n$ cells.
2. **Count:** For sequence data, count bigram transitions $a \to b$.
3. **Normalize:** Compute $P(b \mid a)$ for each $a \in A$.
4. **Embed:** Each a becomes a point in $\mathbb{R}^n$:

$$\phi(a) = (P(b_1 \mid a), P(b_2 \mid a), \ldots, P(b_n \mid a))$$

1.2 Properties

The embedding $\phi : A \to \mathbb{R}^n$ satisfies:

- $\phi(a)$ lies on the $(n-1)$-simplex (probabilities sum to 1)
- Distance in embedding space = distributional similarity
- $\|\phi(a_1) - \phi(a_2)\|$ small $\Rightarrow$ a_1, a_2 predict similar b

2 Example: Vowel $\times$ Consonant-Group

Let $A = \{a, e, i, o, u\}$ (5 vowels) and $B = 7$ consonant groups:

- Stops voiced: b, d, g
- Stops unvoiced: p, t, k
- Fricatives: f, v, s, z
- Nasals: m, n

- Liquids: l, r
- Glides: w, y
- Other: c, h, j, q, x

Product space: $5 \times 7 = 35$ pairs.

2.1 Normalized Embeddings

From 100M bytes of enwik9:

	stops_v	stops_uv	fric	nasal	liquid	glide	other
a	0.087	0.178	0.118	0.277	0.258	0.026	0.056
e	0.146	0.065	0.196	0.187	0.297	0.025	0.084
i	0.108	0.153	0.214	0.318	0.092	0.001	0.115
o	0.056	0.135	0.190	0.312	0.236	0.039	0.032
u	0.117	0.182	0.184	0.194	0.257	0.002	0.064

2.2 Distance Matrix (L2)

	a	e	i	o	u
a	0	0.180	0.208	0.102	0.113
e		0	0.265	0.187	0.131
i			0	0.180	0.216
o				0	0.151
u					0

Clusters:

- (a, o): 0.102 — both favor nasals + liquids ("an", "on", "al", "ol")
- (a, u): 0.113 — similar profiles
- (e, u): 0.131 — both high liquid ("er", "ur")
- i is unique: high nasal ("in", "im"), almost no glide

3 Information Content

$$H(\text{cgroup}) = 2.55 \text{ bits} \tag{1}$$

$$H(\text{cgroup} \mid \text{vowel}) = 2.48 \text{ bits} \tag{2}$$

$$I(\text{cgroup}; \text{vowel}) = 0.07 \text{ bits} \tag{3}$$

The embedding captures 2.8% of consonant-group entropy—not much, but interpretable.

4 Generalization

The same construction works at any scale:

Product Space	Size	Interpretability
Byte × Byte	65,536	Low
ES × ES	25	High
Vowel × CGroup	35	High
Word × Word	V^2	Medium

4.1 Connection to Neural Embeddings

Word2Vec and similar methods learn embeddings via:

$$\phi(w) = \arg\max P(\text{context} \mid w)$$

Our construction is the **exact** version: we compute $P(b \mid a)$ directly from counts, no learning required.

The neural network learns a compressed representation; we get the full distribution.

5 Hierarchical Embeddings

For the RNN hidden state $\mathbf{h} \in \mathbb{R}^H$:

$$\phi(\text{context}) = \mathbf{h}$$

This is an embedding of variable-length history into fixed-dimensional space.

Key insight: The RNN hidden state IS an embedding of the memory trace.