600.465/665 Natural Language Processing **Fall 2025**
Homework #2: Probability and Vector Exercises **Due:** Mon 22 September, 2 pm
Prof. Jason Eisner

Collaborator(s): [None]

1 Problem 1: Probability Axioms (20 points)

1.1 Problem 1(a)

Prove from the axioms that if $Y \subseteq Z$, then $p(Y) \leq p(Z)$.

Proof: Let $Y \subseteq Z$. We can decompose Z as the union of two disjoint sets: $Z = Y \cup (Z \setminus Y)$
Note that $Y \cap (Z \setminus Y) = \emptyset$ by definition of set difference.
Since Y and $(Z \setminus Y)$ are disjoint: $p(Z) = p(Y \cup (Z \setminus Y)) = p(Y) + p(Z \setminus Y)$
Since p maps to $[0, 1]$, we have $p(Z \setminus Y) \geq 0$.
Therefore: $p(Z) = p(Y) + p(Z \setminus Y) \geq p(Y) + 0 = p(Y)$
Hence $p(Y) \leq p(Z)$. ■

Note on the hints: The statement "$p(A) = 0$ does not imply $A = \emptyset$" reminds us that probability zero doesn't mean impossibility in continuous spaces. For example, the probability of hitting exactly one point on a continuous interval is 0, but that point exists. Similarly, "$p(B) = p(C)$ does not imply $B = C$" means equal probabilities don't imply equal sets and different events can have the same probability. For example, consider a biased coin where heads appears with probability 1. Let $B = \{\text{heads}\}$ and $Z = \{\text{heads, tails}\}$. Then $B \subseteq Z$ and $p(B) = p(Z) = 1$ (since $p(\text{tails}) = 0$), yet $B \neq Z$ because Z contains an additional element even though that element has zero probability.

1.2 Problem 1(b)

By definition, $p(X \mid Z) = \frac{p(X \cap Z)}{p(Z)}$ (assuming $p(Z) > 0$).

Lower bound: Since $X \cap Z \subseteq \mathcal{E}$ and probabilities are non-negative, we have $p(X \cap Z) \geq 0$. Also, $p(Z) > 0$ by assumption. Therefore: $p(X \mid Z) = \frac{p(X \cap Z)}{p(Z)} \geq \frac{0}{p(Z)} = 0$

Upper bound: Note that $X \cap Z \subseteq Z$. By part (a), since $X \cap Z \subseteq Z$, we have $p(X \cap Z) \leq p(Z)$. Therefore: $p(X \mid Z) = \frac{p(X \cap Z)}{p(Z)} \leq \frac{p(Z)}{p(Z)} = 1$

Hence $0 \leq p(X \mid Z) \leq 1$.

1.3 Problem 1(c)

Prove from the axioms that $p(\emptyset) = 0$.

Proof: Since $\emptyset$ and $\mathcal{E}$ are disjoint and $\emptyset \cup \mathcal{E} = \mathcal{E}$:
By the axiom: $p(\mathcal{E}) = p(\emptyset \cup \mathcal{E}) = p(\emptyset) + p(\mathcal{E})$.
Since $p(\mathcal{E}) = 1$, we have $1 = p(\emptyset) + 1$.
Therefore $p(\emptyset) = 0$. ■

1.4 Problem 1(d)

Let $\bar{X}$ denote $\mathcal{E} - X$. Prove from the axioms that $p(X) = 1 - p(\bar{X})$.

Proof: Since X and $\bar{X} = \mathcal{E} - X$ are disjoint and $X \cup \bar{X} = \mathcal{E}$:
By the axiom: $p(\mathcal{E}) = p(X \cup \bar{X}) = p(X) + p(\bar{X})$.
Since $p(\mathcal{E}) = 1$, we have $1 = p(X) + p(\bar{X})$.
Therefore $p(X) = 1 - p(\bar{X})$. ■

1.5 Problem 1(e)

Prove from the axioms that $p(\textbf{singing AND rainy} \mid \textbf{rainy}) = p(\textbf{singing} \mid \textbf{rainy})$.

Proof: By definition of conditional probability:

$$p(\text{singing AND rainy} \mid \text{rainy}) = \frac{p(\text{singing AND rainy} \cap \text{rainy})}{p(\text{rainy})}$$

Since (singing AND rainy) $\cap$ rainy $=$ singing AND rainy:

$$p(\text{singing AND rainy} \mid \text{rainy}) = \frac{p(\text{singing AND rainy})}{p(\text{rainy})} = p(\text{singing} \mid \text{rainy})$$

■

1.6 Problem 1(f)

Prove from the axioms that $p(X \mid Y) = 1 - p(\bar{X} \mid Y)$, where $\bar{X}$ denotes $\mathcal{E} - X$.

Proof: By definition of conditional probability:

$$p(X \mid Y) = \frac{p(X \cap Y)}{p(Y)} \quad \text{and} \quad p(\bar{X} \mid Y) = \frac{p(\bar{X} \cap Y)}{p(Y)}$$

Since X and $\bar{X} = \mathcal{E} - X$ partition $\mathcal{E}$, we have $(X \cap Y) \cup (\bar{X} \cap Y) = Y$ and $(X \cap Y) \cap (\bar{X} \cap Y) = \emptyset$.
By the axiom: $p(Y) = p(X \cap Y) + p(\bar{X} \cap Y)$.
Therefore:

$$p(X \mid Y) + p(\bar{X} \mid Y) = \frac{p(X \cap Y) + p(\bar{X} \cap Y)}{p(Y)} = \frac{p(Y)}{p(Y)} = 1$$

Hence $p(X \mid Y) = 1 - p(\bar{X} \mid Y)$. ■

1.7 Problem 1(g)

Simplify: $p(X \mid Y) \cdot p(Y) + p(X \mid \bar{Y}) \cdot p(\bar{Y}) \cdot \frac{p(\bar{Z} \mid X)}{p(\bar{Z})}$

Proof: By definition of conditional probability:

$$p(X \mid Y) \cdot p(Y) = p(X \cap Y)$$

$$p(X \mid \bar{Y}) \cdot p(\bar{Y}) = p(X \cap \bar{Y})$$

Since Y and $\bar{Y}$ partition $\mathcal{E}$, we have $(X \cap Y) \cup (X \cap \bar{Y}) = X$ and $(X \cap Y) \cap (X \cap \bar{Y}) = \emptyset$.

Therefore:

$$p(X \mid Y) \cdot p(Y) + p(X \mid \bar{Y}) \cdot p(\bar{Y}) = p(X \cap Y) + p(X \cap \bar{Y}) = p(X)$$

The expression simplifies to:

$$p(X) \cdot \frac{p(\bar{Z} \mid X)}{p(\bar{Z})}$$

By definition of conditional probability: $p(\bar{Z} \mid X) = \frac{p(\bar{Z} \cap X)}{p(X)}$.
Therefore:

$$p(X) \cdot \frac{p(\bar{Z} \mid X)}{p(\bar{Z})} = p(X) \cdot \frac{p(\bar{Z} \cap X)}{p(X) \cdot p(\bar{Z})} = \frac{p(\bar{Z} \cap X)}{p(\bar{Z})} = p(X \mid \bar{Z})$$

■

1.8 Problem 1(h)

Under what conditions is it true that $p($singing OR rainy$) = p($singing$) + p($rainy$)$?

Proof: By the axiom, $p(A \cup B) = p(A) + p(B)$ if and only if $A \cap B = \emptyset$.

Therefore, $p(\text{singing OR rainy}) = p(\text{singing}) + p(\text{rainy})$ if and only if the events "singing" and "rainy" are disjoint, i.e., $\text{singing} \cap \text{rainy} = \emptyset$.

This means it is impossible to be both singing and experiencing rainy weather simultaneously. ■

1.9 Problem 1(i)

Under what conditions is it true that $p($singing AND rainy$) = p($singing$) \cdot p($rainy$)$?

Proof: By definition of conditional probability: $p(\text{singing AND rainy}) = p(\text{singing} \mid \text{rainy}) \cdot p(\text{rainy})$.

Therefore, $p(\text{singing AND rainy}) = p(\text{singing}) \cdot p(\text{rainy})$ if and only if:

$$p(\text{singing} \mid \text{rainy}) = p(\text{singing})$$

This condition means that knowing whether it is rainy does not change the probability of singing. In other words, the events "singing" and "rainy" are independent. ■

1.10 Problem 1(j)

Suppose you know that $p(X \mid Y) = 0$. Prove that $p(X \mid Y, Z) = 0$.

Proof: By definition: $p(X \mid Y) = \frac{p(X \cap Y)}{p(Y)} = 0$.

Since $p(Y) > 0$ (otherwise the conditional probability would be undefined), we must have $p(X \cap Y) = 0$.

Now, $X \cap Y \cap Z \subseteq X \cap Y$.

By part (a), since $X \cap Y \cap Z \subseteq X \cap Y$ and $p(X \cap Y) = 0$, we have $p(X \cap Y \cap Z) \leq p(X \cap Y) = 0$.

Since probabilities are non-negative, $p(X \cap Y \cap Z) = 0$.

Therefore: $p(X \mid Y, Z) = \frac{p(X \cap Y \cap Z)}{p(Y \cap Z)} = \frac{0}{p(Y \cap Z)} = 0$ (assuming $p(Y \cap Z) > 0$). ■

1.11 Problem 1(k)

Suppose you know that $p(W \mid Y) = 1$. Prove that $p(W \mid Y, Z) = 1$.

Proof: By definition: $p(W \mid Y) = \frac{p(W \cap Y)}{p(Y)} = 1$.

Since $p(Y) > 0$ (otherwise the conditional probability would be undefined), we have $p(W \cap Y) = p(Y)$.

This means $W \cap Y = Y$ (up to sets of measure zero), so $Y \subseteq W$.

Since $Y \subseteq W$, we have $Y \cap Z \subseteq W \cap Y \cap Z$.

But also $W \cap Y \cap Z \subseteq Y \cap Z$.

Therefore $W \cap Y \cap Z = Y \cap Z$.

Hence: $p(W \mid Y, Z) = \frac{p(W \cap Y \cap Z)}{p(Y \cap Z)} = \frac{p(Y \cap Z)}{p(Y \cap Z)} = 1$ (assuming $p(Y \cap Z) > 0$). ■

2 Problem 2: Bayes' Theorem Application (25 points)

2.1 Problem 2(a)

Write an equation relating the following quantities and perhaps other quantities:

$$p(\text{Actual} = \text{blue}) \tag{1}$$

$$p(\text{Actual} = \text{blue} \mid \text{Claimed} = \text{blue}) \tag{2}$$

$$p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{blue}) \tag{3}$$

Proof: By Bayes' Theorem:

$$p(\text{Actual} = \text{blue} \mid \text{Claimed} = \text{blue}) = \frac{p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{blue}) \cdot p(\text{Actual} = \text{blue})}{p(\text{Claimed} = \text{blue})}$$

where $p(\text{Claimed} = \text{blue})$ can be computed using the law of total probability:

$$p(\text{Claimed} = \text{blue}) = p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{blue}) \cdot p(\text{Actual} = \text{blue}) + p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{red}) \cdot p(\text{Ac}$$

■

2.2 Problem 2(b)

Events:

- Actual = blue → the hypothesis you are evaluating
- Claimed = blue → the evidence you have observed

Probabilities:

- p(Actual = blue) → prior probability of the hypothesis
- p(Claimed = blue | Actual = blue) → likelihood of the hypothesis
- p(Actual = blue | Claimed = blue) → posterior probability of the hypothesis

2.3 Problem 2(c)

From the problem: $p(\text{Actual} = \text{blue}) = 0.1$, $p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{blue}) = 0.8$, and $p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{red}) = 0.2$.

First, compute $p(\text{Claimed} = \text{blue})$:

$$p(\text{Claimed} = \text{blue}) = 0.8 \times 0.1 + 0.2 \times 0.9 = 0.08 + 0.18 = 0.26$$

Using Bayes' Theorem:

$$p(\text{Actual} = \text{blue} \mid \text{Claimed} = \text{blue}) = \frac{0.8 \times 0.1}{0.26} = \frac{0.08}{0.26} = \frac{4}{13} \approx 0.31$$

The judge should care about the posterior probability $p(\text{Actual} = \text{blue} \mid \text{Claimed} = \text{blue}) = \frac{4}{13}$, which represents the probability that the car was actually blue given the witness's claim.

2.4 Problem 2(d)

Proof: By definition of conditional probability:

$$p(A \mid B, Y) = \frac{p(A \cap B \cap Y)}{p(B \cap Y)}$$

We can rewrite the numerator using conditional probability:

$$p(A \cap B \cap Y) = p(B \mid A, Y) \cdot p(A \cap Y) = p(B \mid A, Y) \cdot p(A \mid Y) \cdot p(Y)$$

Similarly for the denominator:

$$p(B \cap Y) = p(B \mid Y) \cdot p(Y)$$

Substituting:

$$p(A \mid B, Y) = \frac{p(B \mid A, Y) \cdot p(A \mid Y) \cdot p(Y)}{p(B \mid Y) \cdot p(Y)} = \frac{p(B \mid A, Y) \cdot p(A \mid Y)}{p(B \mid Y)}$$

■

2.5 Problem 2(e)

Proof: We want to prove:

$$p(A \mid B, Y) = \frac{p(B \mid A, Y) \cdot p(A \mid Y)}{p(B \mid A, Y) \cdot p(A \mid Y) + p(B \mid \bar{A}, Y) \cdot p(\bar{A} \mid Y)}$$

From part (d), we have:

$$p(A \mid B, Y) = \frac{p(B \mid A, Y) \cdot p(A \mid Y)}{p(B \mid Y)}$$

We need to express $p(B \mid Y)$ using the law of total probability. Since A and $\bar{A}$ partition the event space:

$$p(B \mid Y) = p(B \mid A, Y) \cdot p(A \mid Y) + p(B \mid \bar{A}, Y) \cdot p(\bar{A} \mid Y)$$

Substituting this into the expression from part (d):

$$p(A \mid B, Y) = \frac{p(B \mid A, Y) \cdot p(A \mid Y)}{p(B \mid A, Y) \cdot p(A \mid Y) + p(B \mid \bar{A}, Y) \cdot p(\bar{A} \mid Y)}$$

■

2.6 Problem 2(f)

The formula from 2(e) with specific propositions:

$$p(\text{Actual} = \text{blue} \mid \text{Claimed} = \text{blue}, \text{Baltimore}) = \frac{\text{numerator}}{\text{denominator}}$$

Where:

$$\begin{aligned}
\text{numerator} &= p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{blue}, \text{Baltimore}) \times p(\text{Actual} = \text{blue} \mid \text{Baltimore}) && (4)\\
\text{denominator} &= p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{blue}, \text{Baltimore}) \times p(\text{Actual} = \text{blue} \mid \text{Baltimore}) && (5)\\
&\quad + p(\text{Claimed} = \text{blue} \mid \text{Actual} = \text{red}, \text{Baltimore}) \times p(\text{Actual} = \text{red} \mid \text{Baltimore}) && (6)
\end{aligned}$$

Substituting values:

$$= \frac{0.8 \times 0.1}{0.8 \times 0.1 + 0.2 \times 0.9} = \frac{0.08}{0.26} = \frac{4}{13}$$

3 Problem 3: Conditional Probability Tables (20 points)

3.1 Problem 3(a)

For each situation, the probabilities sum to 1:

$$p(\text{bwa} \mid \text{Predator!}) + p(\text{bwee} \mid \text{Predator!}) + p(\text{kiki} \mid \text{Predator!}) = 1 \tag{7}$$

$$p(\text{bwa} \mid \text{Timber!}) + p(\text{bwee} \mid \text{Timber!}) + p(\text{kiki} \mid \text{Timber!}) = 1 \tag{8}$$

$$p(\text{bwa} \mid \text{I need help!}) + p(\text{bwee} \mid \text{I need help!}) + p(\text{kiki} \mid \text{I need help!}) = 1 \tag{9}$$

Or more generally: $\sum_{\text{cry}} p(\text{cry} \mid \text{situation}) = 1$ for each situation.

3.2 Problem 3(b)

Given: $p(\text{predator}) = 0.2$, $p(\text{timber}) = 0$, and $p(\text{I need help}) = 0.8$.

Using $p(\text{cry}, \text{situation}) = p(\text{cry} \mid \text{situation}) \cdot p(\text{situation})$:

$p(\text{cry}, \text{situation})$	Predator!	Timber!	I need help!	TOTAL
bwa	$0 \times 0.2 = 0$	$0.1 \times 0 = 0$	$0.8 \times 0.8 = 0.64$	0.64
bwee	$0 \times 0.2 = 0$	$0.6 \times 0 = 0$	$0.1 \times 0.8 = 0.08$	0.08
kiki	$1.0 \times 0.2 = 0.2$	$0.3 \times 0 = 0$	$0.1 \times 0.8 = 0.08$	0.28
TOTAL	0.2	0	0.8	1.0

3.3 Problem 3(c)

i. This probability is written as: $p(\text{predator} \mid \text{kiki})$

ii. It can be rewritten without the | symbol as: $\frac{p(\text{predator,kiki})}{p(\text{kiki})}$

iii. Using the above tables, its value is: $\frac{0.2}{0.28} = \frac{5}{7}$

iv. Alternatively, Bayes' Theorem allows you to express this probability as:

$$\frac{p(\text{kiki} \mid \text{predator}) \cdot p(\text{predator})}{p(\text{kiki} \mid \text{predator}) \cdot p(\text{predator}) + p(\text{kiki} \mid \text{timber}) \cdot p(\text{timber}) + p(\text{kiki} \mid \text{I need help}) \cdot p(\text{I need help})}$$

v. Using the above tables, the value of this is:

$$\frac{1.0 \cdot 0.2}{1.0 \cdot 0.2 + 0.3 \cdot 0 + 0.1 \cdot 0.8} = \frac{0.2}{0.2 + 0 + 0.08} = \frac{0.2}{0.28} = \frac{5}{7}$$

4 Problem 4: N-gram Language Models (25 points)

4.1 Problem 4(a)

Using the chain rule and naive MLE parameters:

$$p(w_1 w_2 w_3 w_4) = p(w_1 \mid \text{BOS}, \text{BOS}) \cdot p(w_2 \mid \text{BOS}, w_1) \cdot p(w_3 \mid w_1, w_2) \cdot p(w_4 \mid w_2, w_3) \cdot p(\text{EOS} \mid w_3, w_4) \tag{10}$$

$$= \frac{c(\text{BOS BOS } w_1)}{c(\text{BOS BOS})} \cdot \frac{c(\text{BOS } w_1 w_2)}{c(\text{BOS } w_1)} \cdot \frac{c(w_1 w_2 w_3)}{c(w_1 w_2)} \cdot \frac{c(w_2 w_3 w_4)}{c(w_2 w_3)} \cdot \frac{c(w_3 w_4 \text{ EOS})}{c(w_3 w_4)} \tag{11}$$

What each count represents:

- c(BOS BOS): Total number of sentences in the corpus
- c(BOS BOS i): Number of sentences starting with word "i"
- c(bagpipe music EOS): Number of sentences ending with "bagpipe music"

The naive estimates often yield zero probabilities for unseen trigrams, making smoothing necessary in practice.

4.2 Problem 4(b)

Under any good language model of English, $p(\vec{w}) = p(\text{do, you, think, the})$ should be extremely low because this is not a complete, grammatical sentence. The sequence ends abruptly with "the" without completing the thought.

In the trigram model, the parameter responsible for making this probability low is $p(\text{EOS} \mid \text{think, the})$. Since sentences rarely end with the word "the" (a determiner that typically precedes a noun), this conditional probability would be very small in any corpus of well-formed English sentences.

The low probability of $p(\text{EOS} \mid \text{think, the})$ reflects the fact that "the" almost always requires a following noun or adjective, making sentence termination at this point highly unlikely.

4.3 Problem 4(c)

Matching expressions with descriptions:

- Expression (A) $p(\text{do}) \cdot p(\text{you} \mid \text{do}) \cdot p(\text{think} \mid \text{do, you})$ represents **(2)** the first 3 words you hear are do you think, all as part of a single sentence.
- Expression (B) $p(\text{do} \mid \text{BOS}) \cdot p(\text{you} \mid \text{BOS, do}) \cdot p(\text{think} \mid \text{do, you}) \cdot p(\text{EOS} \mid \text{you, think})$ represents **(1)** the first complete sentence you hear is do you think.
- Expression (C) $p(\text{do} \mid \text{BOS}) \cdot p(\text{you} \mid \text{BOS, do}) \cdot p(\text{think} \mid \text{do, you})$ represents **(3)** the first complete sentence you hear starts with do you think.

Explanation: Expression (A) has no BOS/EOS markers, so it's about any 3-word sequence. Expression (B) includes both BOS and EOS, making it a complete sentence. Expression (C) starts with BOS but has no EOS, so it represents the beginning of a sentence that continues beyond "think."

4.4 Problem 4(d) - Extra Credit

Forward model:

$$p(\vec{w}) = \prod_{i=1}^{n+1} p(w_i \mid w_{i-2}, w_{i-1})$$

Reversed model:

$$p_{\text{reversed}}(\vec{w}) = \prod_{i=0}^{n} p(w_i \mid w_{i+1}, w_{i+2})$$

For $\vec{w} = (i, love, bagpipe, music)$:
Forward:

$$p(\vec{w}) = \frac{c(\text{BOS BOS i})}{c(\text{BOS BOS})} \cdot \frac{c(\text{BOS i love})}{c(\text{BOS i})} \cdot \frac{c(\text{i love bagpipe})}{c(\text{i love})} \cdot \frac{c(\text{love bagpipe music})}{c(\text{love bagpipe})} \cdot \frac{c(\text{bagpipe music EOS})}{c(\text{bagpipe music})}$$

Reversed:

$$p_{\text{rev}}(\vec{w}) = \frac{c(\text{BOS i love})}{c(\text{i love})} \cdot \frac{c(\text{i love bagpipe})}{c(\text{love bagpipe})} \cdot \frac{c(\text{love bagpipe music})}{c(\text{bagpipe music})} \cdot \frac{c(\text{bagpipe music EOS})}{c(\text{music EOS})} \cdot \frac{c(\text{music EOS EOS})}{c(\text{EOS EOS})}$$

Both expressions contain identical trigram counts in numerators and identical bigram counts in denominators. Since multiplication is commutative, $p(\vec{w}) = p_{\text{reversed}}(\vec{w})$.

5 Problem 5: Topic Models (15 points)

Write a formula for $p(w_1 w_2 w_3 w_4)$ under this better bigram model that includes topic information.

Using the chain rule and conditional independence assumptions:

$$\begin{aligned} p(w_1 w_2 w_3 w_4) &= \sum_a p(w_1 w_2 w_3 w_4, a) && (12) \\ &= \sum_a p(a) \cdot p(w_1 \mid a) \cdot p(w_2 \mid w_1, a) \cdot p(w_3 \mid w_2, a) \cdot p(w_4 \mid w_3, a) && (13) \end{aligned}$$

where a ranges over all possible topics (POLITICS, CELEBRITIES, ANIMALS, SPORTS), and we assume:

- The topic a is chosen first with probability $p(a)$
- Each word depends on both the previous word and the persistent topic a
- The topic remains constant throughout the sentence

This model captures topic persistence while maintaining the bigram dependency structure.

6 Problem 6: Log-linear Models (15 points)

Can't wait to pass the quiz!

7 Problem 7: Word Embeddings Application (20 points)

As is posted online.

8 Problem 8: Programming - Word Similarity (30 points)

8.1 Problem 8(a)

Implementation

The main challenge was completing the `Integerizer` API integration. After examining `integerize.py`, I found that word retrieval uses bracket notation: `word_to_index[idx]` rather than a method call. The implementation normalizes embeddings using `torch.nn.functional.normalize` and computes similarities via `torch.mv` for efficient vectorized operations.

Results Analysis

Testing revealed clear patterns in embedding quality. Strong semantic clustering appeared for content words:

- **dog**: dogs, badger, cat, hound, puppy, dachshund (animal grouping)
- **communist**: socialist, bolshevik, trotskyist, leftist (political ideology)
- **seattle**: seahawks, tacoma, spokane (geographic + cultural associations)

Weaker results occurred with technical terms (**jpg**: mixed relevant formats with random words) and function words (**the**: grammatical particles with unclear semantic similarity).

Dimensional Effects

Dimension size critically affects quality. At 10 dimensions, **dog** returned nonsensical results: "turnip, coronets, ass, pig." At 200 dimensions, results became highly coherent: "dogs, hound, inu, sighthound, mastiff"—all canine-related terms. Similarly, **seattle** progressed from mixed cities (50-dim) to Washington state localities (200-dim: tacoma, spokane, bremerton).

The pattern suggests 10 dimensions are insufficient for meaningful semantic representation, 50+ dimensions provide reasonable quality, and 200 dimensions yield optimal semantic precision. Higher dimensions enable finer-grained distinctions but show diminishing returns beyond a threshold.

Key Observations

Content words with clear semantic categories cluster effectively, while function words and rare technical terms show weaker patterns. Embedding quality correlates with word frequency and semantic richness. The model successfully captures geographic clustering, ideological groupings, and taxonomic relationships when given sufficient dimensional capacity.

8.2 Problem 8(b)

Algorithm Implementation

To extend the program for analogies, I modified the `find_similar_words` method to handle the analogy case when both `--minus` and `--plus` arguments are provided. The algorithm computes the target vector as:

```
query_vector = embeddings[word_idx] - embeddings[minus_idx] + embeddings[plus_idx]
```

The resulting vector is then normalized and used to find the 10 most similar words, excluding all three input words from the results.

Test Results and Analysis

Gender Analogies:

- `king - man + woman`: queen, marries, isabella, sibylla, consort, anjou, vasa, heiress, valois, betrothed
- `doctor - man + woman`: nurse, doctors, obstetrics, gynecology, physician, forrester, dentists, dentist, priscilla, marge
- `uncle - man + woman`: aunt, daughter, mother, niece, wife, cousin, husband, father, married, sister

The gender analogies work exceptionally well. `king→queen` is the perfect result, while `uncle→aunt` correctly identifies the female equivalent. The `doctor→nurse` result reflects gender stereotypes in the training data.

Geographic Analogies:

- `paris - france + uk`: london, leeds, molview, glasgow, reissue, cafe, bluerhinos, promo, gig, premiered
- `tokyo - japan + china`: shanghai, guangdong, beijing, kobe, jiangxi, hangzhou, hubei, hunan, nanjing, chongqing

Geographic relationships show mixed success. `paris→london` works correctly, while `tokyo` maps to multiple Chinese cities, demonstrating the model's understanding of major urban centers.

Historical Analogies:

- `hitler - germany + italy`: mussolini, himmler, rome, loos, speer, petacci, naples, nazi, cesare, ribbentrop

The fascist leader analogy succeeds perfectly with `mussolini` as the top result, showing the model captures historical political relationships.

Linguistic Analogies:

- `child - goose + geese`: children, infants, parents, mothers, sexually, young, pups, parenting, infanticide, midwife
- `goes - eats + ate`: went, got, going, go, coming, gets, gone, walked, came, wakes

Morphological patterns show partial success. `child→children` captures the singular/plural relationship, while `goes→went` demonstrates past tense transformation.

Conceptual Analogies:

- `car - road + air`: helicopter, aircraft, pressurized, airframe, psa, interceptor, usaf, curtiss, usmc, parachutes

The transportation medium analogy successfully transitions from ground to air vehicles, with `helicopter` and `aircraft` as top results.

Dimensional Comparison

Testing `king - man + woman` with different dimensions reveals quality degradation at lower dimensions:

- 10-dim: anjou, flavia, pepin, desiderius, kampaku, quarrelled, magister, pretender, highness, chatillon
- 200-dim: queen, marries, isabella, sibylla, consort, anjou, vasa, heiress, valois, betrothed

Higher dimensions clearly provide superior analogy performance, with 200-dimensional embeddings consistently producing more semantically coherent results.

Why Vector Arithmetic Works

The `king - man` operation creates a vector representing ”royalty without maleness.” This direction vector encodes the semantic transformation from commoner to royal status. Adding `woman` applies this transformation to the female concept, yielding vectors close to female royalty terms. The success demonstrates that consistent semantic relationships manifest as parallel vector offsets in the embedding space, enabling geometric computation of analogical reasoning.

9 Problem 9: Extra Credit - Circular Models (10 points)

Analysis of Besag's Approximation

Besag's approximation $p(A, B, C, D) \approx p(A|B) \cdot p(B|C) \cdot p(C|D) \cdot p(D|A)$ cannot be justified directly by chain rule plus backoff due to the circular dependency structure.

The Problem with Circular Dependencies

The chain rule requires a directed acyclic factorization, but Besag's formula creates a cycle: $A \to B \to C \to D \to A$. This violates the fundamental assumption that we can order variables in a sequence where each depends only on previous ones.

Solution: Breaking the Cycle

We can fix this by arbitrarily choosing a starting point and applying the standard chain rule. For example, starting with variable D:

$$p(A, B, C, D) = p(D) \cdot p(A|D) \cdot p(B|A) \cdot p(C|B)$$

This is a valid probability decomposition that can be justified by the chain rule. We can then apply backoff approximations: - $p(A|D) \approx p(A|D)$ (no backoff needed for bigram) - $p(B|A) \approx p(B|A)$ (no backoff needed for bigram) - $p(C|B) \approx p(C|B)$ (no backoff needed for bigram)

Comparison with Besag's Formula

Our corrected version differs from Besag's original in structure:

$$\begin{aligned} \text{Our version:} \quad & p(D) \cdot p(A|D) \cdot p(B|A) \cdot p(C|B) & (14) \\ \text{Besag's version:} \quad & p(A|B) \cdot p(B|C) \cdot p(C|D) \cdot p(D|A) & (15) \end{aligned}$$

The key difference is that our version includes the marginal $p(D)$ and maintains a consistent directional flow, while Besag's creates a cycle. Both capture pairwise dependencies, but only our version represents a valid probability distribution.

Conclusion

Besag's approximation cannot be derived from chain rule plus backoff due to circular dependencies. The corrected approach breaks the cycle by choosing an arbitrary starting variable, yielding a valid probability decomposition that maintains the spirit of local pairwise dependencies while conforming to probabilistic principles.

10 Problem 10: Extra Credit - Logic as Probability (10 points)

Proving Logical Implication Through Probability

We need to show that (e) follows from (a)-(d) using probability theory, treating logical statements as events with probability 1.

Given Premises:

$$p(\neg\text{shoe} \mid \neg\text{nail}) = 1 \tag{a}$$
$$p(\neg\text{horse} \mid \neg\text{shoe}) = 1 \tag{b}$$
$$p(\neg\text{race} \mid \neg\text{horse}) = 1 \tag{c}$$
$$p(\neg\text{fortune} \mid \neg\text{race}) = 1 \tag{d}$$

To Prove: $p(\neg\text{fortune} \mid \neg\text{nail}) = 1$

Proof Strategy

Consider the joint probability $p(\neg\text{fortune}, \neg\text{race}, \neg\text{horse}, \neg\text{shoe} \mid \neg\text{nail})$.

Using the chain rule:

$$\begin{aligned}
&p(\neg\text{fortune}, \neg\text{race}, \neg\text{horse}, \neg\text{shoe} \mid \neg\text{nail}) && (16)\\
&= p(\neg\text{shoe} \mid \neg\text{nail}) && (17)\\
&\quad \times p(\neg\text{horse} \mid \neg\text{shoe}, \neg\text{nail}) && (18)\\
&\quad \times p(\neg\text{race} \mid \neg\text{horse}, \neg\text{shoe}, \neg\text{nail}) && (19)\\
&\quad \times p(\neg\text{fortune} \mid \neg\text{race}, \neg\text{horse}, \neg\text{shoe}, \neg\text{nail}) && (20)
\end{aligned}$$

Key Insight from Problem 1(k)

Since $p(\neg\text{shoe} \mid \neg\text{nail}) = 1$, by Problem 1(k), we have:

$$p(\neg\text{horse} \mid \neg\text{shoe}, \neg\text{nail}) = p(\neg\text{horse} \mid \neg\text{shoe}) = 1$$

Similarly, since $p(\neg\text{horse} \mid \neg\text{shoe}) = 1$:

$$p(\neg\text{race} \mid \neg\text{horse}, \neg\text{shoe}, \neg\text{nail}) = p(\neg\text{race} \mid \neg\text{horse}) = 1$$

And since $p(\neg\text{race} \mid \neg\text{horse}) = 1$:

$$p(\neg\text{fortune} \mid \neg\text{race}, \neg\text{horse}, \neg\text{shoe}, \neg\text{nail}) = p(\neg\text{fortune} \mid \neg\text{race}) = 1$$

Final Calculation

Therefore:

$$p(\neg\text{fortune}, \neg\text{race}, \neg\text{horse}, \neg\text{shoe} \mid \neg\text{nail}) = 1 \times 1 \times 1 \times 1 = 1 \tag{21}$$

Since this joint probability equals 1, and since $\{\neg\text{fortune}, \neg\text{race}, \neg\text{horse}, \neg\text{shoe}\} \subseteq \{\neg\text{fortune}\}$, by Problem 1(a):

$$p(\neg\text{fortune} \mid \neg\text{nail}) \geq p(\neg\text{fortune}, \neg\text{race}, \neg\text{horse}, \neg\text{shoe} \mid \neg\text{nail}) = 1$$

Since probabilities cannot exceed 1, we conclude $p(\neg\text{fortune} \mid \neg\text{nail}) = 1$.

Conclusion

This demonstrates that logical reasoning is indeed a special case of probabilistic reasoning where all conditional probabilities are either 0 or 1, representing certainty rather than uncertainty.