Quiz: Embeddings and Attention Mechanisms

STA414/2104 - Winter 2026

1. How does the attention mechanism dynamically bridge the gap to non-parametric kernel methods?

   (a) It acts as a fixed prior kernel covariance matrix similar to Gaussian Processes, forcing fixed similarity.

   (b) It relies entirely on linear combinations of fixed inputs without using any multiplicative interactions.

   (c) It replaces a fixed similarity kernel with a soft attention score computed from learned Query and Key embeddings to weight the Values.

   (d) It eliminates the need for any internal state by memorizing the entire training corpus in the Value matrix.

   Correct Answer: (c)
   Rationale: Just as kernel methods use a kernel function to compare an input query to training examples, attention compares an input Query to Keys to generate dynamic weights for the Output Values, replacing the fixed kernel with a soft attention score.

2. Consider a corpus of N documents. A word w appears in every single document. Which of the following correctly describes its TF-IDF score for any document x_i?

   (a) It is very high, because appearing in every document makes it universally important.

   (b) It is zero for every document, because $\text{IDF}(w) = \log\left(\frac{N}{N}\right) = 0$ regardless of term frequency.

   (c) It equals the raw term frequency of w in x_i, since the IDF factor normalises to 1.

   (d) It is undefined, because the denominator of the IDF formula becomes zero.

   Correct Answer: (b)
   Rationale: IDF measures how rare a word is across documents. When a word appears in every document, $N_j = N$, so $\text{IDF}(w) = \log(N/N) = \log(1) = 0$. Multiplying any term frequency by zero yields zero, meaning the word carries no discriminative information and is effectively down-weighted to nothing.

3. What is the primary motivation for using Multi-Head Attention rather than a single attention head?

   (a) It fundamentally reduces the computational complexity of the sequence modeling from $O(T^2)$ to $O(T \log T)$.

   (b) It allows the model to compute attention scores without needing the scaling factor $\sqrt{d}$.

   (c) It creates distinct sets of W^Q, W^K, and W^V matrices, allowing the network to jointly attend to different representation subspaces.

   (d) It forces the attention matrix to become symmetric, mimicking a standard valid positive semi-definite kernel.

   Correct Answer: (c)
   Rationale: By having multiple heads with different learned projections, the model concatenates the outputs of all of the attention heads together. This allows the model to capture different types of relationships in parallel.

4. What fundamental limitation of Recurrent Neural Networks (RNNs) does the attention mechanism directly solve when processing sequences?

   (a) RNNs require an exhaustive $O(2^T)$ search to backpropagate through time.

   (b) RNNs force all past sequence information to be compressed into a single hidden state S_t, causing an information bottleneck.

   (c) RNNs can only process numerical data, whereas attention allows the direct input of raw string characters.

   (d) RNNs cannot process sequences of variable lengths.

   Correct Answer: (b)
   Rationale: Instead of relying on one state vector to carry all historical context, attention directly computes scores linking the current token to all previous hidden states simultaneously.

5. In the Word2Vec skip-gram model, what is the unsupervised learning task that the model is trained to solve?

   (a) Predict a masked center word given all the surrounding words in a fixed context window.

   (b) Predict the TF-IDF score of a word given its one-hot encoding and position in the document.

   (c) Predict the surrounding context words given a single center word, by maximising the average log-probability over all context pairs.

   (d) Predict the document-level label (e.g. spam / not spam) from the embedding of each word in the document.

   Correct Answer: (c)
   Rationale: Skip-gram is built on the distributional hypothesis—words with similar meanings appear in similar contexts. Formally, the model maximises $\frac{1}{T}\sum_{t=1}^{T}\sum_{j\in\text{context}(t)}\log p(w_j \mid w_t)$*, predicting each context word from the center word. Option (a) describes the Continuous Bag-of-Words (CBoW) objective, which is the exact reverse.*

6. You embed two documents using binary Bag-of-Words and compute their cosine similarity. Document A contains exactly the words {"cat", "mat"} and document B contains exactly {"cat", "mat", "sat"} (assume a vocabulary of at least these three words). What is the cosine similarity between A and B?

   (a) $\frac{2}{3}$

   (b) $\frac{\sqrt{2}}{\sqrt{3}}$

   (c) $\frac{\sqrt{2}}{\sqrt{6}}$

   (d) 1

   Correct Answer: (b)
   Rationale: Let $h_A = (1, 1, 0)$ *and* $h_B = (1, 1, 1)$. *The dot product is* $1 \cdot 1 + 1 \cdot 1 + 0 \cdot 1 = 2$. $\|h_A\| = \sqrt{2}$ *and* $\|h_B\| = \sqrt{3}$. *Cosine similarity* $= \frac{2}{\sqrt{2} \cdot \sqrt{3}} = \frac{2}{\sqrt{6}} = \frac{\sqrt{2}}{\sqrt{3}}$.