Relazione Tecnica: Selezione del Numero Ottimale di Cluster per Embedding LLM

Team di Ricerca

26 maggio 2025

Contents

1 Introduzione

In questo documento presentiamo un'analisi approfondita del processo di clustering applicato alle embedding generate da un modello BERT fine-tuned (`Teto03/Bert_base_fineTuned`), con l'obiettivo di separare le risposte di un LLM tra quelle conformi alle policy di allineamento e quelle che rappresentano potenziali jailbreak. La sfida principale consiste nella scelta del numero di cluster, k, che permetta di rilevare possibili sfumature nelle risposte, evitando sia l'underfitting (troppi pochi cluster) sia l'overfitting (cluster non significativi).

2 Descrizione del Codice Base

Il workflow del codice si articola in due macro-fasi principali:

1. **Estrazione delle embedding**: ogni risposta JSON viene passata attraverso il tokenizer e il modello BERT, estraendo il vettore corrispondente al token `[CLS]` dell'ultimo layer.
2. **Clustering e analisi**: applicazione di K-Means e indici di validazione (Elbow Method, Silhouette Analysis), seguita da visualizzazioni 2D (PCA, t-SNE, UMAP) e valutazione finale.

2.1 Estrazione delle embedding

Listing 1: Blocco di caricamento delle embedding

```
with open("response.json", "r") as f:
responses = json.load(f)

embedding_list = []
for item in responses:
text = item.get("response", "")
inputs = tokenizer(text, return_tensors="pt", truncation
   =True, padding=True)
outputs = model(**inputs, output_hidden_states=True)
emb = outputs.hidden_states[-1][:, 0, :].squeeze(0).
   detach().cpu().numpy()
embedding_list.append(emb)

X = np.array(embedding_list)
```

In questa fase raccoglie vettori a dimensione dell'embedding (tipicamente 768 per BERT-base).

3 Selezione del numero di cluster

La determinazione di avviene tramite due principali indici:

- **Elbow Method**
- **Silhouette Analysis**

Di seguito una descrizione formale di entrambi.

3.1 Elbow Method

Definizione. Per ciascun nel range , si calcola l'inertia:
dove:

- = insieme dei punti assegnati al cluster
- = centroide geometric
- = norma Euclidea

In pratica l'inertia misura la compattezza interna dei cluster: più è bassa, più i punti restano vicini ai propri centroidi.

Interpretazione del gomito. Il grafico vs. inizia con curve ripide (forte riduzione di inertia) e poi si appiattisce. Il punto di flesso (gomito) è considerato il ottimale, poiché rappresenta un buon compromesso tra riduzione dell'errore e complessità del modello.

3.2 Silhouette Analysis

Silhouette di un singolo punto. Data un'assegnazione di al cluster :

$$a(x) = \frac{1}{|A|-1} \sum_{y \in A, , y \neq x} d(x,y) \; b(x) \quad = \min_{B \neq A} \frac{1}{|B|} \sum_{y \in B} d(x,y) \; s(x) = \frac{b(x) - a(x)}{\max a(x), b(x)},$$

con distanza Euclidea. Il valore :

- : ottima compattezza interna e separazione esterna
- : punto al confine tra due cluster
- : assegnazione potenzialmente sbagliata

Silhouette media. Si definisce:

il cui massimo indica il numero di cluster che ottimizza simultaneamente compattezza e separazione.

4 Implementazione nel codice

Il blocco principale per la selezione automatica di k è il seguente:

Listing 2: Elbow & Silhouette loop

```
ks = range(2, 11)
inertias, sil_scores = [], []
for k in ks:
km = KMeans(n_clusters=k, random_state=42).fit(X)
inertias.append(km.inertia_)
sil_scores.append(silhouette_score(X, km.labels_))

Plot combinato e scelta di best_k = argmax(sil_scores)
```

Dopo il loop:

- Si traccia un grafico con `plt.plot(ks, inertias)` e `plt.plot(ks, sil_scores)` su assi y distinti.
- Si seleziona $k = \arg\max(\{\text{sil\_scores}\})$, ovvero il valore di k che massimizza la silhouette media.

Il clustering finale con consente di procedere alle visualizzazioni 2D (PCA, t-SNE, UMAP) e alle metriche finali (silhouette media, silhouette plot dettagliato, distribuzione dei cluster).

5 Visualizzazioni 2D

Per validazione qualitativa, proiettiamo su uno spazio 2D utilizzando:

- PCA (*Principal Component Analysis*)
- t-SNE (*t-Distributed Stochastic Neighbor Embedding*)
- UMAP (*Uniform Manifold Approximation and Projection*)

Ognuna di queste tecniche riduce la dimensionalità preservando differenti proprietà (varianza globale, strutture locali, geometria del manifold).

6 Risultati e conclusioni

Dall'analisi congiunta dei due indici, siamo in grado di:

1. Determinare in maniera automatica e robusta
2. Valutare la qualità del clustering con metriche numeriche e grafici diagnostici
3. Identificare eventuali outlier o punti borderline tramite silhouette plot

Questa metodologia garantisce un bilanciamento tra rigore quantitativo e verifica qualitativa tramite visualizzazioni.

Relazione generata automaticamente dal sistema di analisi clustering, 26 maggio 2025.