LEARNING A LANGUAGE MODEL

Assumiamo di avere un **training corpus**. Per tale corpus sia $P(s)$ il **LM perfetto**. C'è un modo per calcolare un LM approssimato $Q(s)$ più vicino possibile a $P(s)$, usando il training corpus?

Possiamo raggiungere il nostro obiettivo risolvendo il seguente problema di ottimizzazione:

$$\arg\min_{Q} \mathcal{L}(P, Q)$$

Dove $\mathcal{L}$ è la **loss function** che specifica l'errore che abbiamo usando $Q(s)$ invece di $P(s)$.

Dobbiamo trovare $\mathcal{L}$. Essa deve confrontare due distribuzioni di probabilità, per farlo usiamo l'**entropia**.

INFORMATION ENTROPY

Data una variabile aleatoria discreta X, con probabilità $P(x)$, definiamo l'**information content** $I[x]$ di un outcome x come:

$$I[x] = -\log P[x]$$

La $I[x]$ può essere vista come una funzione $I(x)$, così che possiamo considerare la variabile aleatoria (Trasformata) $I(x)$ e calcolarne l'**expectation** $\mathbb{E}[I(x)]$, che è anche chiamata **entropia** $H(P)$ della variabile aleatoria X con prob. P:

$$H(P) = \mathbb{E}[I(x)] = \mathbb{E}[-\log P(x)] = -\sum_{x} P[x] \log P[x]$$

Nel caso dei LM, questo concetto di $H(P)$ va adattato alla sequenza di parole dipendenti dalla perplexity PP. Bisogna tradurre l'entropia dell'informazione $H(P)$ con un'altra detta entropia del linguaggio.

CROSS-ENTROPY di $Q(x)$

Vogliamo capire quanta informazione $Q(x)$ cepra, rispetto al modello perfetto $P(x)$. Per farlo calcoliamo la cross-entropy di $Q(x)$ rispetto a $P(x)$

$$H(P,Q) = -\sum_x P[x] \log Q[x]$$

$Q[x]$ è usata solo come sorgente dell'informazione. Dobbiamo calcolare la perdita di informazione nell'usare Q al posto di P. Calcoliamo la seguente differenza detta Loss-Information Content:

$$H(P,Q) - H(P) =$$

$$-\sum_x P[x] \log Q[x] + \sum_x P[x] \log P[x] =$$

$$= \sum_x P[x] \left(\log P[x] - \log Q[x]\right)$$

$$= \sum_x P[x] \log \frac{P[x]}{Q[x]} \equiv D_{KL}(P \| Q)$$

Questa $D_{KL}(P\|Q)$ è detta relative entropy o anche KULLBACK-LEIBLER (KL) DIVERGENCE.

Non è una metrica ma una semi-metrica.

Disuguaglianza di Gibb

$$D_{KL}(P\|Q) \geq 0$$

Dove l'uguaglianza si verifica solo se $P \equiv Q$.

In generale D_{KL} non è simmetrica:

$$D_{KL}(P\|Q) \neq D_{KL}(Q\|P)$$

Dunque NON È UNA DISTANZA

ENTROPIA DI UN LM

Abbiamo visto come il nostro obiettivo sia quello di confrontare due distribuzioni di probabilità P e Q.

Assumiamo di avere n dati, e consideriamo la sequenza random di parole $S_n = w_1 \dots w_n$.
Assumiamo poi $P(S_n)$ come LM di S_n. Invece di calcolare l'entropia di S_n, che dipende da n, calcoliamo la PER-WORD-ENTROPY (o entropy rate):

$$H'(S_n) = -\frac{1}{n} \sum_{S_n} P[S_n] \log P[S_n]$$

Che è una somma di tutte le frasi di lunghezza n.
Definiamo ora l'entropia di un LM $P(s)$:

$$H(P) \triangleq \lim_{n \to \infty} H'(S_n) = -\lim_{n \to \infty} \frac{1}{n} \sum_{S_n} P[S_n] \log P[S_n]$$

Però noi non conosciamo $P(s)$ ma solo la sua approssimazione $Q(s)$. Per cui definiamo la **cross entropy** $H(P,Q)$:

$$H(P,Q) = -\lim_{M\to\infty} \frac{1}{M} \sum_{S_M} P[S_M] \log Q[S_M]$$

Ancora abbiamo nella formula $P(s)$ che non siamo in grado di calcolare.

TEOREMA SHANNON-McMILLAN-BRSIMAN

Se un LM è **stazionario** e **ergodico** allora si può calcolare la cross-entropy:

$$H(P,Q) = -\lim_{M\to\infty} \sum_{S_M} \frac{1}{M} \log Q[S_M]$$

Con S_M **singola** sequenza randomica di parole.

Stationary LM

Un LM si dice stazionario sse:

$$P[w_{t+1} \dots w_{t+M}] = P[w_1 \dots w_M]$$

Per ogni sequenza di parole e per ogni parametro di shift t Il che vuol dire che la distribuzione di probabilità per la parola a tempo t non cambia a tempo $t+1$. Modelli come il bigram sono stazionari. I linguaggi naturali non lo sono perché la probabilità della prossima parola può dipendere da eventi (o parole) arbitrariamente distanti nel tempo. Noi, infatti, stiamo **approssimando**.

Ergodic LM

Un LM è definito ergodico se l'expectation di una parola $\delta[w_i]$ su $P(S_M)$ può essere calcolata come la media temporale della singola lunga sequenza casuale $w_1 w_2 ..$

In realtà la parola w_i potrebbe non apparire mai in una lunga sequenza, portando $\delta[w_i] = 0$. Questa proprietà vuole che $\delta[w_i] > 0$.

NEGATIVE LOG LIKELIHOOD

Quindi, senza conoscere P, e data una singola lunga sequenza $w_1 \dots w_T$ generata tramite $P(s)$, si approssima:

$$H(P,Q) \approx -\frac{1}{T} \log Q[w_1 \dots w_T]$$

Detta negative log likelihood (NLL) di $Q(S_M)$

PERPLEXITY e CROSS ENTROPY

La lunga sequenza $w_1 \dots w_T$ è detta evaluation corpus e se possiamo calcolare la perplexity $PP(w_1 \dots w_T)$.

Usando un trucco matematico possiamo dire:

$$PP(w_1 \dots w_T) = \exp\left(\log PP(w_1 \dots w_T)\right)$$

$$= \exp\left(\log Q[w_1 \dots w_T]^{-1/T}\right) = \exp\left(-\frac{1}{T} \log Q[w_1 \dots w_T]\right)$$

$$= \exp\left(H(P,Q)\right)$$

Quindi perplexity e cross-entropy sono strettamente correlate, ma non uguali! Da notare che crescono e decrescono insieme.

MINIMIZZIAMO LA LOSS FUNCTION

Il problema si riduce nel risolvere:

$$\arg\min_Q \left(\mathcal{L}(P, Q) \right)$$

$$\downarrow$$

$$\arg\min_Q \left(D_{KL}(P \| Q) \right)$$

$$\downarrow$$

$$\arg\min_Q \left(H(P, Q) - H(P) \right)$$

$$\downarrow$$ perché $H(P)$ non dipende da Q

$$\arg\min_Q \left(H(P, Q) \right)$$

$$\downarrow$$

$$\arg\min_Q \left(-\frac{1}{T} \log Q[w_1 \dots w_T] \right)$$

$$\downarrow$$

$$\arg\max_Q \left(+\frac{1}{T} \log Q[w_1 \dots w_T] \right)$$