6.864 Problem Set #2

Dongyoung Kim

October 12, 2015

Notation

Throughout this pset, I will use the notation $x^{(1:n)}$ to denote a list of $\{x_1, x_2, ..., x_n\}$

Question 1

1.1 dimensionality: $\theta = N^2 + (N-1)|\Sigma|$

1.2

$$\begin{aligned} P &= p(y_1|'START')p('the'|y_1=1)p(y_2=2|y_1=1)p(\text{"}dog\text{"}|y_2=2)p(y_3=1|y_2=2) \\ &\times p(\text{"}the\text{"}|y_3=1)p('END'|y_3=1) \\ &= 0 \end{aligned}$$

The last equality comes from the fact that $p(y_3 = 1|y_2 = 2) = 0$

Question 2

2.1 $|\mathcal{T}|^n$

2.2

$$\begin{aligned} &\pi(0,*) = 1 \\ &\pi(0,v) = 0 \quad \text{for all other values of } v \end{aligned}$$

2.3 We will prove using induction.

Let us consider recursive formula,

$$\pi(k,v) = \max_{u \in \mathcal{T}} \pi(k-1,u) a_{u,v} b_v(x_k) \tag{1}$$

From 2.2, we know,

$$\pi(0, v) = \begin{cases} 1 & \text{if } v = * \\ 0 & \text{otherwise} \end{cases}$$

- Base case for $k = 1$:

  Given by base case,

$$\begin{aligned} \pi(1, v) &= \max_{u \in \mathcal{T}} \pi(0, u) a_{u,v} b_v(x_1) \\ &= \pi(0, *) a_{*,v} b_v(x_1) \\ &= 1 \cdot a_{*,v} \cdot b_v(x_1) \\ &= r(y_1 = v) \end{aligned}$$

  Since y_1 is fixed, (1) satisfies trivially

- Prove inductive step k assuming step $k-1$ is true :

  Assume $\pi(k-1, u) = \max_{y^{(1:k-1)} \in S(k-1,u)} r(y^{(1:k-1)}) \quad \forall u$

  Then,

$$\begin{aligned} \max_{y^{(1:k)} \in S(k,v)} r(y^{(1:k)}) &= \max_{y^{(1:k)} \in S(k,v)} r(y^{(1:k-1)}) a_{y_{k-1}, y_k} b_{y_k}(x_k) \\ &= \max_{y^{(1:k-1)}} r(y^{(1:k-1)}) a_{y_{k-1}, v} b_v(x_k) \\ &= \max_{u \in \mathcal{T}, y^{(1:k-1)} \in S(k-1,v)} r(y^{(1:k-1)}) a_{u_v} b_v x_k \\ &= \max_{u \in \mathcal{T}} \pi(k-1, u) a_{u,v} b_v(x_k) \end{aligned}$$

  Which is precisely our recursion formula in (1)

2.4 We are essentially filling out a table of size $n|\mathcal{T}|$, and each step takes $O(|\mathcal{T}|)$ cost. Total cost is therefore $O(n|\mathcal{T}|^2)$

Question 3

3.1 Consider

$$p(y|x^i) = \frac{p(x^i, y)}{\sum_{y'} p(x^i, y')}$$

Where $p(x^i, y)$ is the likelihood of the tag sequence y associated with a output sequence x^i

The fractional count $\overline{count}$ is simply
$\overline{count}(u \to v) = \sum_i \sum_y p(x^i, y) count(x^i, y, u \to v)$

3.2 $a_{u,v} = \dfrac{\overline{count}(u \to v)}{\sum_{v'} \overline{count}(u \to v')}$

3.3

$$\begin{aligned}\sum_p \alpha_p(j)\beta_p(j) &= \sum_p p(x^{(1:j-1)}, y_j = p|\theta) p(x^{(j:n)}|y_j = p, \theta) \\ &= \sum_p p(x^{(1:n)}, y_j = p|\theta) \\ &= p(x^{(1:n)}|\theta)\end{aligned}$$

3.4

$$p(y_i = p|x^{(1:n)}, \theta) = \frac{p(y_i = p, x^{(1:n)}|\theta)}{p(x^{(1:n)}|\theta)}$$

We have $p(x^{(1:n)}|\theta) = \sum_p \alpha_p(j)\beta_p(j)$, so we only have to prove the numerator is indeed as given, which can be verified by applying chain rule

$$\begin{aligned}\alpha_p(i)\beta_p(i) &= p(x^{(1:j-1)}, y_j = p|\theta) p(x^{(j:n)}|y_j = p, \theta) \\ &= p(y_j = p, x^{(1:n)}|\theta)\end{aligned}$$

Question 4

4.1 Among unigram, bigram, and trigram, the model with the highest likelihood on the test data will dominate over others. Since trigram has the most overfitting effect, $\lambda_3 = 1$ and $\lambda_2 = \lambda_1 = 0$

4.2

(a) Define

$$p(\lambda, w_t, w_{t-1}, w_{t-2}) = \begin{cases} \lambda_1 p_{ML}(w_t) & \text{if } \lambda \text{ is an instance of } \lambda_1 \\ \lambda_2 p_{ML}(w_t|w_{t-1}) & \text{if } \lambda \text{ is an instance of } \lambda_2 \\ \lambda_3 p_{ML}(w_t|w_{t-1}, w_{t-2}) & \text{if } \lambda \text{ is an instance of } \lambda_3 \end{cases}$$

Then,

$$\hat{n}(\lambda) = \sum_t \frac{p(\lambda, w_t, w_{t-1}, w_{t-2})}{\sum_{\lambda'} p(\lambda', w_t, w_{t-1}, w_{t-2})}$$

(b)

$$\lambda_y^{(k+1)} = \frac{\hat{n}(\lambda_y^{(k)})}{\sum_{y'} \hat{n}(\lambda_{y'}^{(k)})}$$

4.3 $\lambda_2^{(t)}$ will be zero for all iteration t. That is because fractional count $\hat{n}(\lambda_2)$ is always zero. The EM will find a suboptimal solution over λ_1 and λ_3 where λ_2 is fixed as zero.

Question 5

5.1

$$\hat{n}_t(z) = \sum_{i=1}^{N_t} p(z|w_i, t)$$

$$\hat{n}(w, z) = \sum_{t \in [1,n]} \sum_{i=1}^{N_t} \delta(w_i - w) p(z|w_i, t)$$

5.2

$$\theta_{z|t} = \frac{\hat{n}_t(z)}{\sum_{z'} \hat{n}_t(z)}$$

$$\theta_{w|z} = \frac{\hat{n}(w, z)}{\sum_{w'} \hat{n}(w', z)}$$

5.2

(a) Done

(b) We can look at the goal function of the EM Algorithm (the likelihood) and optimize the number of topics. Also, if the word distributions for each topic $\theta_{w|z}$ is far away from each other (KL divergence could be a good criteria), it would be a good sign that each topic represents different spectra of words.

(c) Politics: President, administration, democrats

War1: Military, special, forces,

War2: War, forces, Afghanistan, American

Economy 1: Economy, company, industry

Economy 2: Business, season, power

Journalism 1: News, stories, New, York, Times, day

Journalism 2: Journal, news, paint, Washington, editorial

NYT: New, York, Times, people, Tuesday

Finance: Business, year, city, bank

Random: clown, fright, crouch, cornerbacks

The topics are not unique, i.e., what seems like a similar topic to the human mind could be classified under multiple different labels. Furthermore, many words are seen with high probability across multiple topic labels.

(d) EM algorithm may converge at a sub-optimal local maxima depending on the starting values of the parameter. To work around this issue, randomizing the starting state is the key. If we assign a uniform value to the parameters, the algorithm will converge at the same point for any given test runs.