Computational Genomics – HW 1

Stefan

Collaborator on the assignment: Zaid Ahmad (all Ps), Max Hunter (Problem 1,4,2), Bielmeier

1a) Prove: $\sum_{x_1}\sum_{x_2}\dots\sum_{x_L} P(x_1, x_2, \dots, x_L) = 1$

$$= \sum_{x_1}\sum_{x_2}\dots\sum_{x_L} P(x_1)\cdot\prod_{n=2}^{L} P(x_n \mid x_{n-1})$$

$$= \sum_{x_1}\sum_{x_2}\dots\sum_{x_L} P(x_1)\cdot\prod_{n=2}^{L-1} P(x_n \mid x_{n-1})\cdot P(x_L \mid x_{L-1})$$

$$= \sum_{x_1}\sum_{x_2}\dots\sum_{x_{L-1}} P(x_1)\cdot\prod_{n=2}^{L-1} P(x_n \mid x_{n-1})\cdot \underbrace{\sum_{x_L} P(x_L \mid x_{L-1})}$$

sum of conditional probabilities – here: transitions from state x_{L-1} is $\hat{=} 1$!

$$= \sum_{x_1}\sum_{x_2}\dots\sum_{x_{L-2}} P(x_1)\cdot\prod_{n=2}^{L-2} P(x_n \mid x_{n-1})\cdot \underbrace{\sum_{x_{L-1}} P(x_{L-1} \mid x_{L-2})}_{=1}\cdot 1$$

Do this further until:

$$= \sum_{x_1}\sum_{x_2} P(x_1)\cdot P(x_2 \mid x_1) = \sum_{x_1} P(x_1)\cdot \underbrace{\sum_{x_2} P(x_2 \mid x_1)}_{=1}$$

$$= \sum_{x_1} P(x_1) = 1$$

Sum of all start states $\pi = 1$ (obviously)

$$\Rightarrow \sum_{x_1}\sum_{x_2}\dots\sum_{x_L} P(x_1, x_2, \dots, x_L) = 1$$ is correct under Markov assumption!, Markov chain w/o end state is a valid probability measure over all sequences L.

16) $P(\text{end} \mid x_i) = \delta$ is given. Any state can transition to end with this probability. $\Rightarrow \sum_{x_{i+1}} P(x_{i+1} \mid x_i) = 1 - \delta$

At the same time, here we also assume that transition to end only happens after sequence length of L.

Now: $\sum_{x_1} \sum_{x_2} \dots \sum_{x_L} P(x_1, \dots, x_L, \text{end}) = ?$

$$= \sum_{x_1} \sum_{x_2} \dots \sum_{x_L} P(x_1) \cdot \left(\prod_{n=2}^{L} P(x_n \mid x_{n-1}) \right) \cdot P(\text{end} \mid x_L)$$

$$= \sum_{x_1} \sum_{x_2} \dots \sum_{x_L} P(x_1) \left(\prod_{n=2}^{L-1} P(x_n \mid x_{n-1}) \right) \cdot P(x_L \mid x_{L-1}) \cdot P(\text{end} \mid x_L)$$

$$= \sum_{x_1} \sum_{x_2} \dots \sum_{x_{L-1}} P(x_1) \prod_{n=2}^{L-1} P(x_n \mid x_{n-1}) \cdot \underbrace{\sum_{x_L} P(x_L \mid x_{L-1})}_{=(1-\delta)} \cdot P(\text{end} \mid x_L)$$

summing over x_L would give us $P(\text{end})$, which we do not know. We know that $P(\text{end} \mid x_i) = \delta$, though.

sorry for the weird one's: $1 \hat{=} 1$ (EN) (US) I'm trying

$$= \sum_{x_1} \sum_{x_2} \dots \sum_{x_{L-1}} P(x_1) \prod_{n=2}^{L-1} P(x_n \mid x_{n-1}) \cdot \delta \cdot \underbrace{\sum_{x_L} P(x_L \mid x_{L-1})}_{=(1-\delta)}$$

$$= \delta \sum_{x_1} \sum_{x_2} \dots \sum_{x_{L-1}} P(x_1) \prod_{n=2}^{L-1} P(x_n \mid x_{n-1}) \cdot (1-\delta) \Rightarrow \text{Pull out constant.}$$

$$= \delta (1-\delta) \sum_{x_1} \sum_{x_2} \dots \sum_{x_{L-2}} P(x_1) \prod_{n=2}^{L-2} P(x_n \mid x_{n-1}) \underbrace{\sum_{x_{L-1}} P(x_{L-1} \mid x_{L-2})}_{(1-\delta)}$$

$$= \delta (1-\delta)^2 \sum_{x_1} \sum_{x_2} \dots \sum_{x_{L-2}} P(x_1) \prod_{n=2}^{L-2} P(x_n \mid x_{n-1}), \text{ continue until only } \sum_{x_1} \text{ is left}$$

$$= \delta (1-\delta)^{L-1} \underbrace{\sum_{x_1} P(x_1)}_{=1} = \underline{\underline{\delta (1-\delta)^{L-1}}}$$

$$\Rightarrow \sum_{x_1} \sum_{x_2} \dots \sum_{x_L} P(x_1, \dots, x_L, x_{\text{end}}) = \underline{\underline{\delta (1-\delta)^{L-1}}}$$

Stefan B.
HW 1
Compton

1c) Now - sum of probabilities of all possible sequences
⇒ At any length L ⇒ Sum over all lengths

$$\Rightarrow \sum_{L=1}^{\infty} \sum_{x_1} \sum_{x_2} \cdots \sum_{x_L} P(x_1, \ldots, x_L, \text{end})$$, from before we know

$$\sum_{x_1} \sum_{x_2} \cdots \sum_{x_L} P(x_1, \ldots, x_L, \text{end}) = \delta(1-\delta)^{L-1}$$

$$= \sum_{L=1}^{\infty} \delta(1-\delta)^{L-1} = \delta \sum_{L=1}^{\infty} (1-\delta)^{L-1}$$

$$= \delta \sum_{L=1}^{\infty} \frac{(1-\delta)^L}{1-\delta} = \frac{\delta}{1-\delta} \underbrace{\sum_{L=1}^{\infty} (1-\delta)^L}$$

infinite geometric series with first term $= 1-\delta$ & ratio $1-\delta$ // $|\delta - 1| < 1$? yes!

$$\Rightarrow \sum_{L=1}^{\infty} (1-\delta)^L \text{ converges to } \Rightarrow \frac{1-\delta}{1-(1-\delta)} = \frac{1-\delta}{\delta}$$

$$\Rightarrow \frac{\delta}{1-\delta} \cdot \frac{1-\delta}{\delta} = \underline{\underline{1}}$$

⇒ Markov chain with end state defines valid probability measure.

2b) for $k=1$,
training_accuracy = 0.8586
validation_accuracy = 0.8684

c) see PDF page for plot

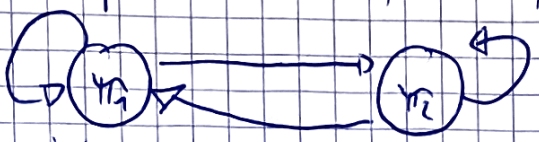
d) I observe a sharp increase for validation accuracy after $k=1$, once we consider 2 or more nucleotides (1)
The growth from $k=2$ over $k=3$ to $k=4$ is existent, but negligible (0.004)(2) At $k=5$, the val_acc begins to decrease again, which is a trend that will probably continue for $k>5$. (3)

Why? (1): with $k=1$, we don't observe a full codon as one transition. With $k>=2$, we do.
(2) The accuracy stays around optimal for $k=2$, $k=3$, $k=4$ as it not overly precise or lax to observe codons.
(3) The accuracy may start to drop as the model considers too big of regions for an accurate reflection of the contained nucleotides & codons (overlaps too big, as well)

e) 2 Problems of $k>20$:
(1) Space issues. My computer ran out of memory for storing huge subsequences in dicts with very small values/observations, e.g. for $k>10$
(2) Accuracy: model may tend to overfit to training data with huge $k>20$, which results in decreasing validation accuracy. Overfitting may be caused by the low count of transitions from specific 20-nucleotide states to another 20-nucleotide-long state.

HW1 - ct'd - Stefan Bielmeier

3a) Example: m (states) = 2, k (possible obs) = 3

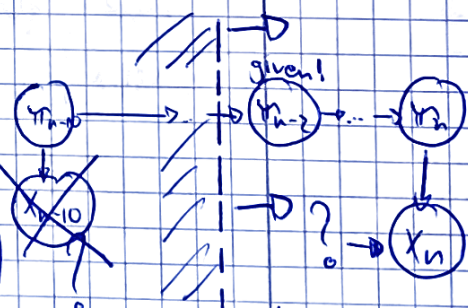


each state has $m = 2$ transition possibilities
there are $m = 2$ states
$\Rightarrow$ # of params for transitions: $m \cdot m$

Observations / emissions?
Each state has $k = 3$ obs possibilities
there are $m = 2$ states
$\Rightarrow$ # of params for emissions: $k \cdot m$
$\Rightarrow$ # of parameters for HMM = $\underline{\underline{m \cdot (m + k)}}$

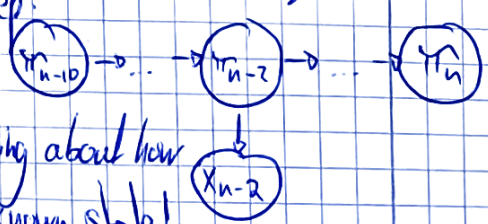
b) False!! $P(x_2 | x_1)$ is not a probability used to calculate the probability of state path & sequence.
only: $P(\pi_2 | \pi_1)$ (transition) & $P(x_1 | \pi_1)$ are used, together with the initial $P(\pi_1 | 0) = P(\pi_1)$ → emission

c) $(x_n \perp\!\!\!\perp x_{n-10}) \mid \pi_{n-2}$
in other words: is this true/equal?
$P(x_n | x_{n-10}, \pi_{n-2}) = P(x_n | x_{n-2})$
$\Rightarrow$ does x_{n-10} add any information for us to calculate x_n, if π_{n-2} is given?? No! ~~Before observations only depend on last state~~



$\Rightarrow$ Equal $\Rightarrow$ True; cond. indep.

d) $(\pi_n | \pi_{n-10}) \mid x_{n-2}$



Given x_{n-2}, can we learn something about how to calc π_n by π_{n-10}? Yes! Last known state!
$\Rightarrow$ Not cond independent $\Rightarrow$ False

5a) Initialize transition_matrix with

	fair	loaded
fair	0.5	0.5
loaded	0.5	0.5

where P(col | row) ≙ transition from row to column.
e.g: P(1|0) ≙ from fair to loaded.
Rows sum up to 1

Init emission_matrix: with

row 0:	fair	1/6	1/6	1/6	1/6	$\frac{1}{6}$	$\frac{1}{6}$	
row 1:	loaded	0.12	0.12	0.12	0.12	0.12	0.4	(assuming 6 is wanted)
column correspond to dice		0	1	2	3	4	5	
		1	2	3	4	5	6	

Results:

(see next page as python terminal output, same format however of the transition & emission matrix)