COMS W4701: Artificial Intelligence, Summer 2022

Homework 4 Solutions

Problem 1: Robot Localization (30 points)

(a) The following table is the transition matrix T:

	A	B	C	D	E	F
A	0	$\frac{1}{2}$	$\frac{1}{2}$	0	0	0
B	$\frac{1}{2}$	0	0	$\frac{1}{4}$	0	0
C	$\frac{1}{2}$	0	0	$\frac{1}{4}$	0	0
D	0	$\frac{1}{2}$	$\frac{1}{2}$	0	1	1
E	0	0	0	$\frac{1}{4}$	0	0
F	0	0	0	$\frac{1}{4}$	0	0

To solve for the stationary distribution π, we can find its eigenvector whose eigenvalue is $\lambda = 1$, which is $[\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{3}, \frac{1}{12}, \frac{1}{12}]^T$

(b) The observation matrix for # is a diagonal matrix $O_{\#} = diag(0, \frac{1}{4}, \frac{1}{4}, 0, \frac{1}{2}, \frac{1}{2})$.

We have $X_0 = [0, 0, 0, 1, 0, 0]^T$. The belief distribution is $\Pr(X_1|e_1)$ $= \Pr(e_1|X_1)\Pr(X_1|X_0)Pr(X_0) = O_{\#}@T@X_0$ (Note: @ is matrix multiplication). After normalization, $\Pr(X_1|e_1) = [0, \frac{1}{6}, \frac{1}{6}, 0, \frac{1}{3}, \frac{1}{3}]^T$

(c) The observation matrix for empty is a diagonal matrix $O_{empty} = diag(\frac{1}{2}, \frac{1}{2}, \frac{1}{2}, 1, \frac{1}{4}, \frac{1}{4})$. Following the above logic, we can compute $\Pr(X_2|e_1, e_2) = O_{empty}@T@\Pr(X_1|e_1) = [\frac{1}{11}, 0, 0, \frac{10}{11}, 0, 0]^T$.

(d) By observation, we know that the (X_1, X_2) sequences with nonzero probability are BA, BD, CA, CD, ED, and FD. We compute $\Pr(X1, X2|e_1, e_2) = \Pr(X_1)\Pr(e_1|X_1)\Pr(X_2|X_1)\Pr(e_2|X_2)$ for each of these sequences and normalize to obtain the nonzero joint probabilities:

$\Pr(B, A|e_1, e_2) = \frac{1}{22}$, $\Pr(B, D|e_1, e_2) = \frac{1}{11}$, $\Pr(C, A|e_1, e_2) = \frac{1}{22}$, $\Pr(C, D|e_1, e_2) = \frac{1}{11}$, $\Pr(E, D|e_1, e_2) = \frac{4}{11}$, $\Pr(F, D|e_1, e_2) = \frac{4}{11}$. DED and DFD are the most likely sequences.

(e) We initialize $\beta_2 = [1, 1, 1, 1, 1, 1]^T$, then $b = \Pr(e_2|X_1) = T^T@O_{empty}@\beta_2 = [\frac{1}{2}, \frac{3}{4}, \frac{3}{4}, \frac{3}{8}, 1, 1]^T$. This is the probability that we observe "empty" in the second time step given that we are in each of the respective states in the first time step.

(f) $f * b = [0, \frac{1}{6}, \frac{1}{6}, 0, \frac{1}{3}, \frac{1}{3}]^T * [\frac{1}{2}, \frac{3}{4}, \frac{3}{4}, \frac{3}{8}, 1, 1]^T$, which normalizes to $[0, \frac{3}{22}, \frac{3}{22}, 0, \frac{4}{11}, \frac{4}{11}]^T$. We can get the same from (d) by marginalizing out the second state from the joint distribution. $\Pr(B|e_1, e_2) = \frac{1}{22} + \frac{1}{11} = \frac{3}{22}$, $\Pr(C|e_1, e_2) = \frac{1}{22} + \frac{1}{11} = \frac{3}{22}$, $\Pr(E|e_1, e_2) = \frac{4}{11}$, $\Pr(F|e_1, e_2) = \frac{4}{11}$, and the rest are zero.

Problem 2: Descendants of Effects (20 points)

(a)

$$
\begin{aligned}
Pr(D_0, D_1, \ldots, D_n) &= \sum_{A,B} Pr(A, B, D_0, D_1, \ldots, D_n) \\
&= \sum_{A,B} Pr(D_1, \ldots, D_n|D_0) Pr(D_0|A, B) Pr(A) Pr(B) \\
&= \sum_{A,B} Pr(D_0|A, B) Pr(A) Pr(B) \prod_{i=1}^{n} Pr(D_i|D_{i-1}) \\
&= Pr(D_0) \prod_{i=1}^{n} Pr(D_i|D_{i-1})
\end{aligned}
$$

The size of the full joint distribution should be 2^{n+1} as there are $n+1$ random binary variables and there are a total of 2 nonzero entries: $Pr(+d_0, +d_1, \ldots, +d_n)$ and $Pr(-d_0, -d_1, \ldots, -d_n)$.

(b) $Pr(+d_n|D_0) = 1$ if $D_0 = +d_0$ else 0. To get the same result from the joint distribution in part (a),

$$
\begin{aligned}
Pr(+d_n|D_0) &= \frac{\sum_{D_1, D_2, \ldots, D_{n-1}} Pr(D_0, D_1, \ldots, +d_n)}{Pr(D_0)} \\
&= \frac{\sum_{D_1, D_2, \ldots, D_{n-1}} Pr(D_0, D_1, \ldots, +d_n)}{\sum_{A,B} Pr(D_0|A, B) Pr(A) Pr(B)} \\
&= \frac{\sum_{D_1, D_2, \ldots, D_{n-1}} Pr(+d_n|D_{n-1}) \prod_{i=1}^{n-1} Pr(D_i|D_{i-1}) \sum_{A,B} Pr(D_0|A, B) Pr(A) Pr(B)}{\sum_{A,B} Pr(D_0|A, B) Pr(A) Pr(B)} \\
&= \sum_{D_1, D_2, \ldots, D_{n-1}} Pr(+d_n|D_{n-1}) \prod_{i=1}^{n-1} Pr(D_i|D_{i-1})
\end{aligned}
$$

(c)

$$
\begin{aligned}
Pr(A, B, D_0, +d_n) &= Pr(+d_n|D_0, A, B) Pr(D_0|A, B) Pr(A) Pr(B) \\
&= Pr(+d_n|D_0) Pr(D_0|A, B) Pr(A) Pr(B)
\end{aligned}
$$

(d)

$$
\begin{aligned}
Pr(A, B| + d_n) &= \frac{Pr(A, B, d_k)}{Pr(d_k)} \\
&= \frac{\sum_{D_0, \ldots, D_{n-1}} \prod_{i=1}^{n} Pr(D_i|D_{i-1})\ Pr(D_0|A, B) Pr(A, B)}{Pr(+d_n)}
\end{aligned}
$$

A	B	D_0	Pr($D_0 \mid A, B$)	$Pr(A, B \mid D_0)$
$+a$	$+b$	$+d_0$	1.0	$1.0(0.5)^2 = 0.25$
$+a$	$+b$	$-d_0$	0.0	0
$+a$	$-b$	$+d_0$	0.5	$0.5(0.5)^2 = 0.125$
$+a$	$-b$	$-d_0$	0.5	$0.5(0.5)^2 = 0.125$
$-a$	$+b$	$+d_0$	0.5	$0.5(0.5)^2 = 0.125$
$-a$	$+b$	$-d_0$	0.5	$0.5(0.5)^2 = 0.125$
$-a$	$-b$	$+d_0$	0.0	0
$-a$	$-b$	$-d_0$	1.0	$1.0(0.5)^2 = 0.25$

A	B	$Pr(A, B \mid D_0 = +d_0)$
$+a$	$+b$	0.5
$+a$	$-b$	0.25
$-a$	$+b$	0.25
$-a$	$-b$	0.0

A	$Pr(A \mid D_0 = +d_0)$
$+a$	0.75
$-a$	0.25

B	$Pr(B \mid D_0 = +d_0)$
$+b$	0.75
$-b$	0.25

We note that $Pr(A = +a, B+ = b \mid D_k = +d_k) = 0.25$ and $Pr(A = +a \mid D_k = +d_k) \times Pr(B+ = b \mid D_k = +d_k) = 0.5625$ which concludes that A and B are not independent conditioned on $+d_k$.

Problem 3: Part-of-Speech Tagging

3.3: Model Evaluation (8 points)

(a) Train accuracy is around 95.27% and the test accuracy is around 88.32%. The reason why the test accuracy is lower is that there are unseen words and unseen transitions in the test set.

(b) Each word might have a different POS; probability-based model only outputs the most likely sequence, which may not be true.

3.5: Inference Comparison (6 points)

In general, to see this result we need to have a sentence containing a word that can take on at least two different parts of speech. Since the difference between the forward and forward-backward algorithms is that the latter takes into account the context of the entire sentence while the former only uses the part of the sentence before the word, one approach would be to construct a sentence in which the second part of the sentence is long and/or gives some new context.