CS461 – RECITATION 07

MACHINE LEARNING PRINCIPLES

Daize Dong

2025-10-27

TODAY’S CONTENT

- Markov Model Assignment
- Quiz 03

MARKOV MODELS

1. States: $S = \{s_1, s_2, \dots, s_N\}$
2. Distributions: $\pi, \text{where } \pi(i) = P(z_1 = i)$
3. Transition Matrix: $A, \text{where } A(i,j) = P(z_t = j \mid z_t - 1 = i)$
4. Emission Matrix: $B, \text{where } B(i,j) = P(x_i \mid z_j)$

MARKOV EXAMPLE

Imagine a language only containing the following:

- class N, nouns: {wolf, parrot, …}
- class AND, word {and}
- class V, verbs: {run, fly, …}

MARKOV EXAMPLE

The state transitions and their probabilities look like:

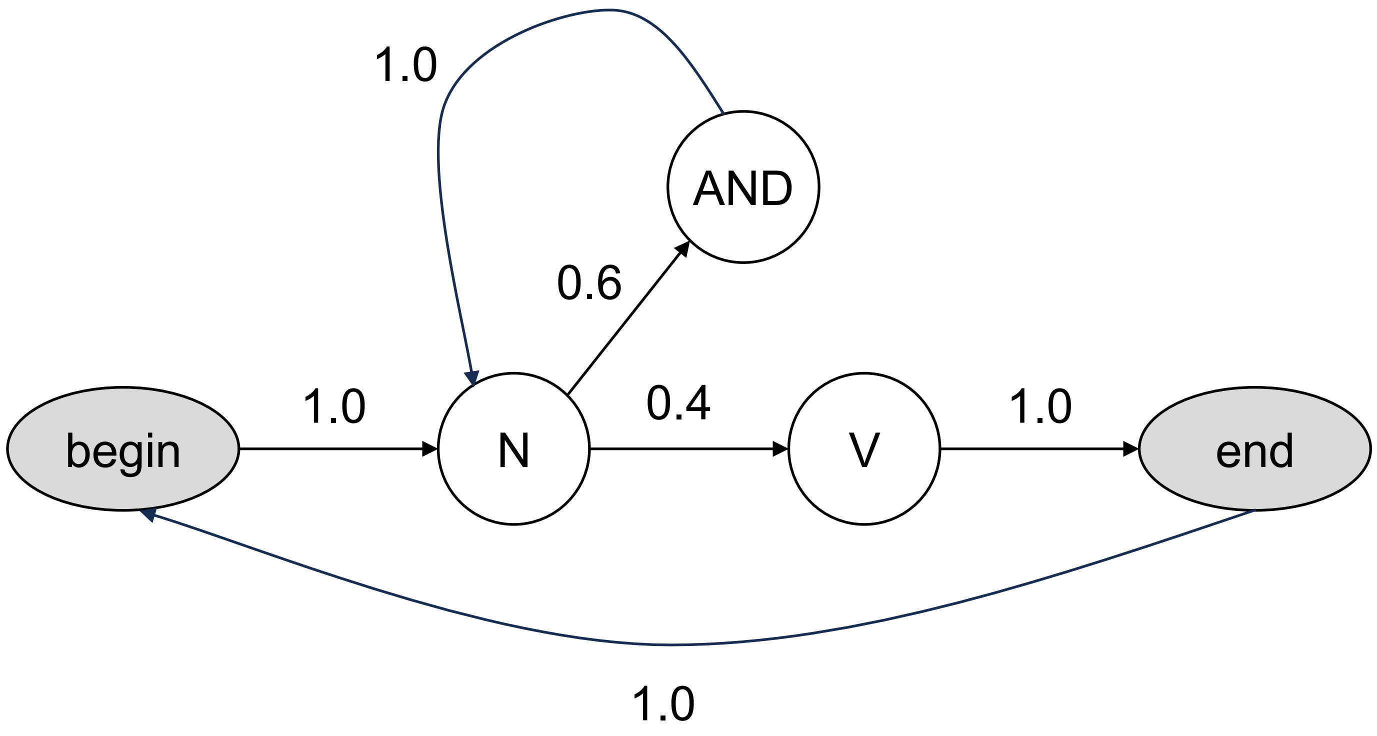
bb: begin

N: {wolf, parrot, …}

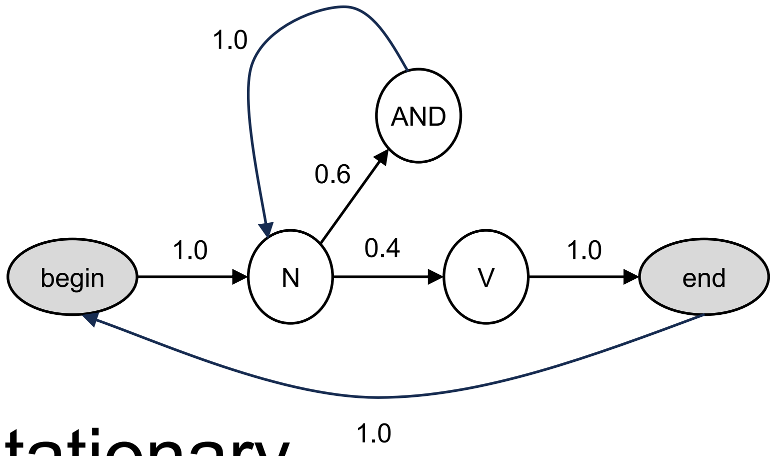
AND: {and}

V: {run, fly, …}

ee: end



MARKOV EXAMPLE



Now solve the following questions:

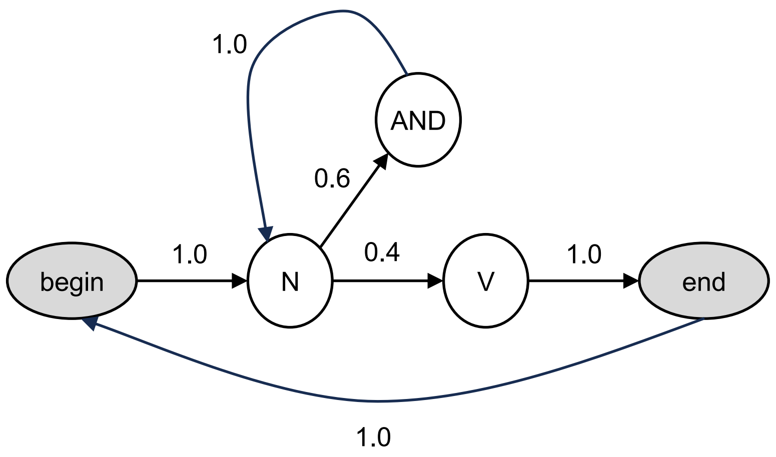
1. Populate a state transition matrix A.
2. Solve this set of equations to determine the stationary distributions, π.
3. Use the fraction of time spent in the **bb** or **ee** states to deduce the average number of words in a sentence.

1. TRANSITION MATRIX

End Points

$$
A = \begin{array}{c} \\ \text{bb} \\ \text{N} \\ \text{AND} \\ \text{V} \\ \text{ee} \end{array}
\begin{array}{c}
\begin{array}{ccccc} \text{bb} & \text{N} & \text{AND} & \text{V} & \text{ee} \end{array} \\
\begin{bmatrix}
0 & 1.0 & 0 & 0 & 0 \\
0 & 0 & 0.6 & 0.4 & 0 \\
0 & 1.0 & 0 & 0 & 0 \\
0 & 0 & 0 & 0 & 1.0 \\
1.0 & 0 & 0 & 0 & 0
\end{bmatrix}
\end{array}
$$

Start Points

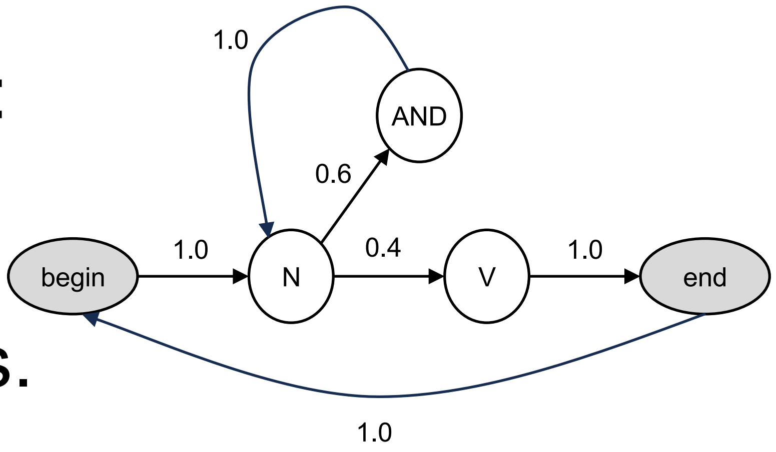


2. STATIONARY DISTRIBUTION

When reached the stationary distribution, we have:

$\pi = \pi A$ and $\sum_{s \in S} \pi(s) = 1$

Here π can be treated as a vector of all state probs.



$$\pi = \pi A \;\rightarrow\; \begin{bmatrix} \pi(bb) \\ \pi(N) \\ \pi(AND) \\ \pi(V) \\ \pi(ee) \end{bmatrix}^T = \begin{bmatrix} \pi(bb) \\ \pi(N) \\ \pi(AND) \\ \pi(V) \\ \pi(ee) \end{bmatrix}^T \begin{bmatrix} 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0.4 & 0 \\ 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \\ 1.0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

2. STATIONARY DISTRIBUTION

$$\pi = \pi A \;\rightarrow \begin{bmatrix} \pi(bb) \\ \pi(N) \\ \pi(AND) \\ \pi(V) \\ \pi(ee) \end{bmatrix}^T = \begin{bmatrix} \pi(bb) \\ \pi(N) \\ \pi(AND) \\ \pi(V) \\ \pi(ee) \end{bmatrix}^T \begin{bmatrix} 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 0.6 & 0.4 & 0 \\ 0 & 1.0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1.0 \\ 1.0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

$$\rightarrow \begin{cases} \pi(bb) = \pi(ee) \\ \pi(N) = \pi(bb) + \pi(AND) \\ \pi(AND) = 0.6\pi(N) \\ \pi(V) = 0.4\pi(N) \\ \pi(ee) = \pi(V) \end{cases} \rightarrow \begin{cases} \pi(bb) = 0.4\pi(N) \\ \pi(N) = \pi(N) \\ \pi(AND) = 0.6\pi(N) \\ \pi(V) = 0.4\pi(N) \\ \pi(ee) = 0.4\pi(N) \end{cases} \quad (1)$$

2. STATIONARY DISTRIBUTION

$$
① \begin{cases}
\pi(bb) = 0.4\pi(N) \\
\pi(N) = \pi(N) \\
\pi(AND) = 0.6\pi(N) \\
\pi(V) = 0.4\pi(N) \\
\pi(ee) = 0.4\pi(N)
\end{cases}
$$

$$\sum_{s \in S} \pi(s) = 1$$

$$\rightarrow \pi(bb) + \pi(N) + \pi(AND) + \pi(V) + \pi(ee) = 1 \quad ②$$

Combine ① and ②, we get:

$$0.4\pi(N) + \pi(N) + 0.6\pi(N) + 0.4\pi(N) + 0.4\pi(N) = 1$$

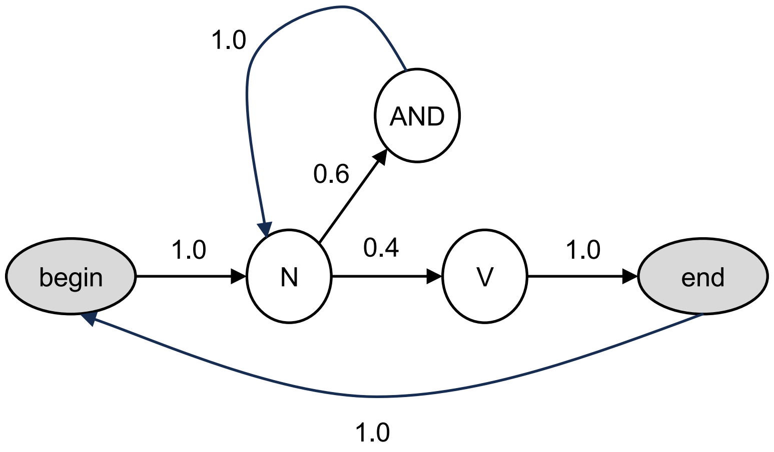
$$2.8\pi(N) = 1$$

$$\pi(N) = \frac{5}{14}$$

2. STATIONARY DISTRIBUTION

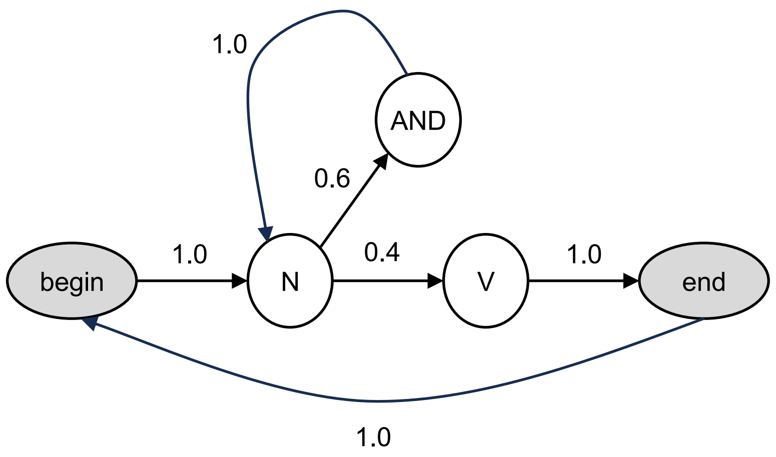
Finally we have:

$$\begin{cases} \pi(bb) = 0.4\pi(N) = \frac{1}{7} \\ \pi(N) = \frac{5}{14} \\ \pi(AND) = 0.6\pi(N) = \frac{3}{14} \\ \pi(V) = 0.4\pi(N) = \frac{1}{7} \\ \pi(ee) = 0.4\pi(N) = \frac{1}{7} \end{cases}$$



3. AVERAGE SENTENCE LENGTH

We just need to find the average number of **bb** OR **ee** states in the sequence (as all sentences must have 1 **bb** and 1 **ee**).



$$\pi(bb) = \pi(ee) = \frac{1}{7}$$

$$\left|\overline{\pi(bb)}\right| = \frac{1}{\pi(bb)} = 7$$

Remember **bb** and **ee** don't count as words:

$$L = \left|\overline{\pi(bb)}\right| - 2 = 5$$

QUIZ 03

QUIZ 03

A	B	C	D
12	5	0	3

1. What is required to create a good ensemble?

(a) Some independence between members of the ensemble.

(b) A data resampling technique to train members of the ensemble.

(c) Strong individual learners in the ensemble.

(d) Weak individual learners in the ensemble.

$\mathrm{Var}\left(\frac{1}{M}\sum_{m=1}^{M} h_m\right) \approx \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2$, where ρ is the correlation factor

Data resampling is not necessary (random forest uses random features)

QUIZ 03

A	B	C	D
1	12	2	5

2. Which of these does not describe an ensemble technique?

(a) For each learner, change the weights of the different samples, pushing the learning to form a different decision boundary. AdaBoost

(b) For each learner, find the principal components of the dataset and drop low variance features.

(c) For each learner, resample observed samples with replacement, changing the data distribution. Bagging

(d) For each learner, select a subset of columns, randomly dropping features.

Random Forest

QUIZ 03

A	B	C	D
0	3	0	17

3. Which of these statements about ensemble techniques and the bias-variance trade-off is false?

(a) Bagging reduces the variance of the training dataset for each ensemble member.

(b) Boosting can ensemble high-bias models to draw complex decision boundaries within datasets that have high variance.

(c) Random forests reduce the variance of the training dataset for each ensemble member.

(d) Adaboost works best with strong learners that already capture the variance of the dataset on their own.

$$\mathrm{Var}\left(\frac{1}{M}\sum_{m=1}^{M} h_m\right) \approx \rho\sigma^2 + \frac{1-\rho}{M}\sigma^2$$

Easy to overfit on strong learners.

QUIZ 03

A	B	C	D
2	2	13	3

4. Which of these statements about (regular, not hidden) Markov models is false?

(a) A time-invariant markov model only looks at the-chain of events within a finite number of steps. N-gram

(b) A 0-gram Markov chain prediction has no dependency upon past tokens.

(c) Training n-gram models with larger n takes more computation, but not more space or data. scales exponentially

(d) As the length of n-grams grow, the training data required to fill the transition matrix rapidly grows.

Assume 3-grams with 5 states, then the number of paths is 5^3=125

QUIZ 03

A	B	C	D
2	0	8	10

5. Which of these statements about a hidden Markov model with 3 hidden states and a vocabulary of 5 observable tokens is true?

(a) The transition matrix, A, is size 3x5.

(b) The emission matrix, B, is size 5x5. 3x5

(c) The parameters of the HMM can be trained with the EM algorithm.

(d) B and C are both true.

Transition matrix: probabilities transiting to the next state.

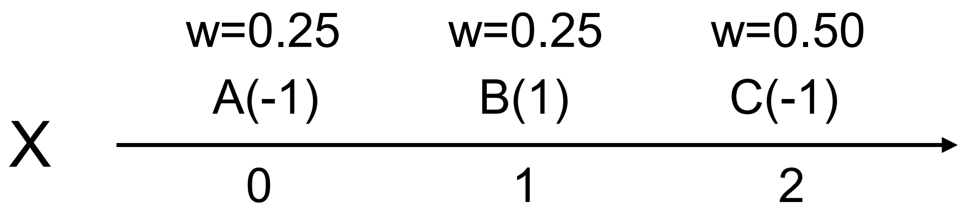
Emission matrix: probabilities of all states during different time steps.

QUIZ 03

6) Consider a set of datapoints with the given x values, classes, and current weights while creating an Adaboost ensemble of decision stumps.

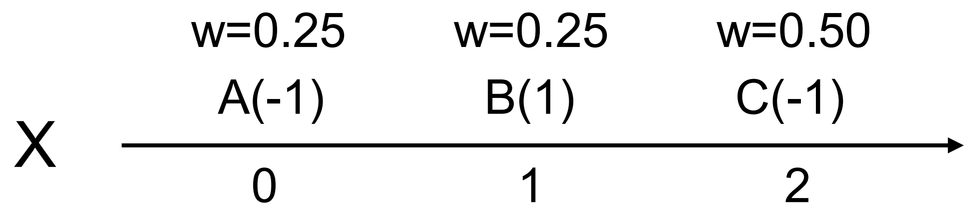
Points	A	B	C
x	0	1	2
class	-1	1	-1
weight	0.25	0.25	0.5

6.1) A decision stump with polarity=1 will classify anything with values < its boundary value as class -1 and any other points as class 1. A stump with polarity=-1 will do the opposite. Using the points in the table, what are **all** of the possible *weighted errors* for decision stumps at each x value and polarity?



QUIZ 03

Points	A	B	C
x	0	1	2
class	-1	1	-1
weight	0.25	0.25	0.5



	For Polarity=1			For Polarity=-1		
X=0	A(1)	B(1)	C(1)	A(-1)	B(-1)	C(-1)
	e=0.25	e=0	e=0.50	e=0	e=0.25	e=0
	Total error=0.75			**Total error=0.25**		
X=1	A(-1)	B(1)	C(1)	A(1)	B(-1)	C(-1)
	e=0	e=0	e=0.50	e=0.25	e=0.25	e=0
	Total error=0.5			**Total error=0.5**		
X=2	A(-1)	B(-1)	C(1)	A(1)	B(1)	C(-1)
	e=0	e=0.25	e=0.50	e=0.25	e=0	e=0
	Total error=0.75			**Total error=0.25**		

6.2) Which of those stumps and polarity will be added into the ensemble? Identify the stump by its x value and polarity.

____

6.3) For simplicity, assign your stump a confidence of $\alpha = 0.7$. It is added into an existing ensemble with two stumps. The first is x=2, polarity=-1, and $\alpha = 0.5$. The second is x=1, polarity=1, and $\alpha = 0.5$. Use this ensemble of three stumps to classify the three points. Voting is done by summing the confidence-weighted votes and taking the sign. Show the sum and class for each point below.

Point A: _______________________________

Point B: _______________________________

Point C: _______________________________

6.2) Which of those stumps and polarity will be added into the ensemble? Identify the stump by its x value and polarity.

Polarity=-1, X=0 OR 2

Just pick up the model with lowest sum error

6.3) For simplicity, assign your stump a confidence of $\alpha = 0.7$. It is added into an existing ensemble with two stumps. The first is x=2, polarity=-1, and $\alpha = 0.5$. The second is x=1, polarity=1, and $\alpha = 0.5$. Use this ensemble of three stumps to classify the three points. Voting is done by summing the confidence-weighted votes and taking the sign. Show the sum and class for each point below.

Point A: ____________________

Point B: ____________________

Point C: ____________________

QUIZ 03

Points	A	B	C
x	0	1	2
class	-1	1	-1
weight	0.25	0.25	0.5

	w=0.25	w=0.25	w=0.50
	A(-1)	B(1)	C(-1)
X	0	1	2

(First Choice)

X=0, Polarity=-1, α=0.7	A(-1) -0.7	B(-1) -0.7	C(-1) -0.7
X=2, Polarity=-1, α=0.5	A(1) 0.5	B(1) 0.5	C(-1) -0.5
X=1, Polarity=1, α=0.5	A(-1) -0.5	B(1) 0.5	C(1) 0.5
SUM	A(-1) -0.7	B(1) 0.3	C(-1) -0.7

QUIZ 03

Points	A	B	C
x	0	1	2
class	-1	1	-1
weight	0.25	0.25	0.5

	w=0.25	w=0.25	w=0.50
	A(-1)	B(1)	C(-1)
X	0	1	2

(Another situation)

X=2, Polarity=-1, α=0.7	A(1) 0.7	B(1) 0.7	C(-1) -0.7
X=2, Polarity=-1, α=0.5	A(1) 0.5	B(1) 0.5	C(-1) -0.5
X=1, Polarity=1, α=0.5	A(-1) -0.5	B(1) 0.5	C(1) 0.5
SUM	A(1) 0.7	B(1) 1.7	C(-1) -0.7

6.2) Which of those stumps and polarity will be added into the ensemble? Identify the stump by its x value and polarity.

Polarity=-1, X=0 OR 2

Just pick up the model with lowest sum error

6.3) For simplicity, assign your stump a confidence of $\alpha = 0.7$. It is added into an existing ensemble with two stumps. The first is x=2, polarity=-1, and $\alpha = 0.5$. The second is x=1, polarity=1, and $\alpha = 0.5$. Use this ensemble of three stumps to classify the three points. Voting is done by summing the confidence-weighted votes and taking the sign. Show the sum and class for each point below.

Point A: -1, s=-0.7 OR 1,s=0.7

Point B: 1, s=0.3 OR 1,s=1.7

Both are correct.

Point C: -1, s=-0.7 OR -1,s=-0.7

Q&A