Variable Length Markov Chains

Josh Magarick

2012, October 10

Outline

The Markov Assumption

- Suppose we have a sequence of observations X_1^n.
- The standard first order Markov chain makes the assumption that $P(X_t|X_1^{t-1}) = P(X_t|X_{t-1})$.
- More generally, we may have $P(X_t|X_1^{t-1}) = P(X_t|X_{t-k}^{t-1})$ for some fixed k.

The Markov Assumption

- Suppose we have a sequence of observations X_1^n.
- The standard first order Markov chain makes the assumption that $P(X_t|X_1^{t-1}) = P(X_t|X_{t-1})$.
- More generally, we may have $P(X_t|X_1^{t-1}) = P(X_t|X_{t-k}^{t-1})$ for some fixed k.

The Markov Assumption

- Suppose we have a sequence of observations X_1^n.
- The standard first order Markov chain makes the assumption that $P(X_t|X_1^{t-1}) = P(X_t|X_{t-1})$.
- More generally, we may have $P(X_t|X_1^{t-1}) = P(X_t|X_{t-k}^{t-1})$ for some fixed k.

But k cannot be too large

- Typically, when working with k^{th} order Markov chains, we will embed them into a first order chain.
- This means that for an alphabet A, we have effectively $|A|^k$ states.
- In practice, this limits k as we need enough data to estimate our model parameters.
- And do we always believe that k should be the same for all past data?

But k cannot be too large

- Typically, when working with k^{th} order Markov chains, we will embed them into a first order chain.
- This means that for an alphabet A, we have effectively $|A|^k$ states.
- In practice, this limits k as we need enough data to estimate our model parameters.
- And do we always believe that k should be the same for all past data?

But k cannot be too large

- Typically, when working with k^{th} order Markov chains, we will embed them into a first order chain.
- This means that for an alphabet A, we have effectively $|A|^k$ states.
- In practice, this limits k as we need enough data to estimate our model parameters.
- And do we always believe that k should be the same for all past data?

But k cannot be too large

- Typically, when working with k^{th} order Markov chains, we will embed them into a first order chain.
- This means that for an alphabet A, we have effectively $|A|^k$ states.
- In practice, this limits k as we need enough data to estimate our model parameters.
- And do we always believe that k should be the same for all past data?

Motivating examples

- If we are trying to predict the next word or character in a sequence of English words, how far back we want to look may depend on the past data.
- For example, if we see a ‘q’ we probably don’t need to look back any farther. If we see an ‘e’, maybe we do.
- In genetics, if we are trying to predict the next codon, how far back we need to look depends on the preceding codons.

Motivating examples

- If we are trying to predict the next word or character in a sequence of English words, how far back we want to look may depend on the past data.
- For example, if we see a ‘q’ we probably don’t need to look back any farther. If we see an ‘e’, maybe we do.
- In genetics, if we are trying to predict the next codon, how far back we need to look depends on the preceding codons.

Motivating examples

- If we are trying to predict the next word or character in a sequence of English words, how far back we want to look may depend on the past data.
- For example, if we see a ‘q’ we probably don’t need to look back any farther. If we see an ‘e’, maybe we do.
- In genetics, if we are trying to predict the next codon, how far back we need to look depends on the preceding codons.

What is Context?

What if the memory of our Markov chain were a function of the data, and was only long when necessary?

- This is the notion of context.
- In this framework, the probability distribution of the next observation is given by $P(X_t|X_1^{t-1}) = P(X_t|c(X_1^{t-1}))$ where $c(\cdot)$ is our context function.
- The context function is defined by:
  - $c : A^\infty \to \bigcup_{i=0}^{\infty} A^i$
  - $c(x_{-\infty}^0) = x_{-\ell+1}^0$, where $\ell = \min\{p : P(X_1|x_{-\infty}^0) = P(X_1|x_{-p+1}^0)\}$. In other words, the smallest number of states we have to look back so that our estimated probability distribution for the next state is the same as if we could look all the way back.

What is Context?

What if the memory of our Markov chain were a function of the data, and was only long when necessary?

- This is the notion of context.
- In this framework, the probability distribution of the next observation is given by $P(X_t|X_1^{t-1}) = P(X_t|c(X_1^{t-1}))$ where $c(\cdot)$ is our context function.
- The context function is defined by:
  - $c : A^\infty \to \bigcup_{i=0}^{\infty} A^i$
  - $c(x_{-\infty}^0) = x_{-\ell+1}^0$, where $\ell = \min\{p : P(X_1|x_{-\infty}^0) = P(X_1|x_{-p+1}^0)\}$. In other words, the smallest number of states we have to look back so that our estimated probability distribution for the next state is the same as if we could look all the way back.

What is Context?

What if the memory of our Markov chain were a function of the data, and was only long when necessary?

- This is the notion of context.
- In this framework, the probability distribution of the next observation is given by $P(X_t|X_1^{t-1}) = P(X_t|c(X_1^{t-1}))$ where $c(\cdot)$ is our context function.
- The context function is defined by:
  - $c : A^\infty \to \bigcup_{i=0}^{\infty} A^i$
  - $c(x_{-\infty}^0) = x_{-\ell+1}^0$, where $\ell = \min\{p : P(X_1|x_{-\infty}^0) = P(X_1|x_{-p+1}^0)\}$. In other words, the smallest number of states we have to look back so that our estimated probability distribution for the next state is the same as if we could look all the way back.

What is Context?

What if the memory of our Markov chain were a function of the data, and was only long when necessary?

- This is the notion of context.
- In this framework, the probability distribution of the next observation is given by $P(X_t|X_1^{t-1}) = P(X_t|c(X_1^{t-1}))$ where $c(\cdot)$ is our context function.
- The context function is defined by:
  - $c : A^\infty \to \bigcup_{i=0}^{\infty} A^i$
  - $c(x_{-\infty}^0) = x_{-\ell+1}^0$, where $\ell = \min\{p : P(X_1|x_{-\infty}^0) = P(X_1|x_{-p+1}^0)\}$. In other words, the smallest number of states we have to look back so that our estimated probability distribution for the next state is the same as if we could look all the way back.

Representing the Context Function

- A context function can be represented by a tree, where each value in the range is represented by a path down the tree.
- For example, if $A = \{0, 1\}$, our context tree might be:

Figure : A tree representing the contexts $\{01, 11, 000, 100, 10\}$

Representing the Context Function

- A context function can be represented by a tree, where each value in the range is represented by a path down the tree.
- For example, if $A = \{0, 1\}$, our context tree might be:

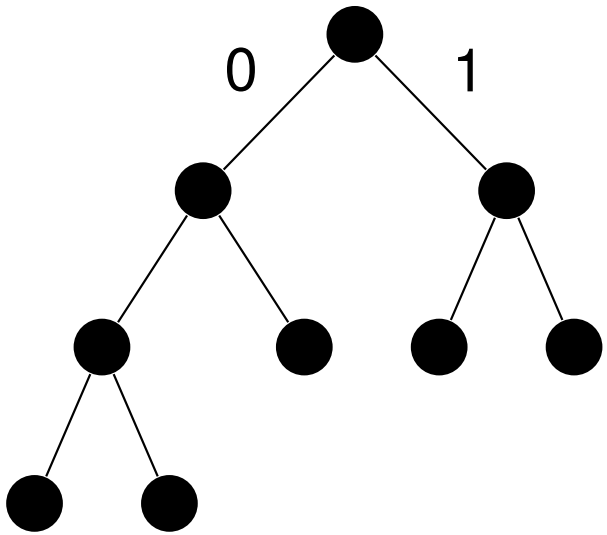

Figure : A tree representing the contexts $\{01, 11, 000, 100, 10\}$

More on Context Trees

- Note that if a context tree has depth k and every internal node has $|A|$ children, it represents the state space of a k^{th} order Markov chain.
- We will often refer to the terminal nodes of a context tree, meaning those nodes with no children.

Using Context to Estimate Probabilities

- Once we have contexts, we want to associate, to each context w a probability distribution $P(\cdot|w)$.
- This is equivalent to associating, with each node of the context tree a distribution over the symbols in A.

Figure : The same tree, but now representing transition probabilities

Using Context to Estimate Probabilities

- Once we have contexts, we want to associate, to each context w a probability distribution $P(\cdot|w)$.
- This is equivalent to associating, with each node of the context tree a distribution over the symbols in A.

Figure : The same tree, but now representing transition probabilities

Using Context to Estimate Probabilities

- Once we have contexts, we want to associate, to each context w a probability distribution $P(\cdot|w)$.
- This is equivalent to associating, with each node of the context tree a distribution over the symbols in A.

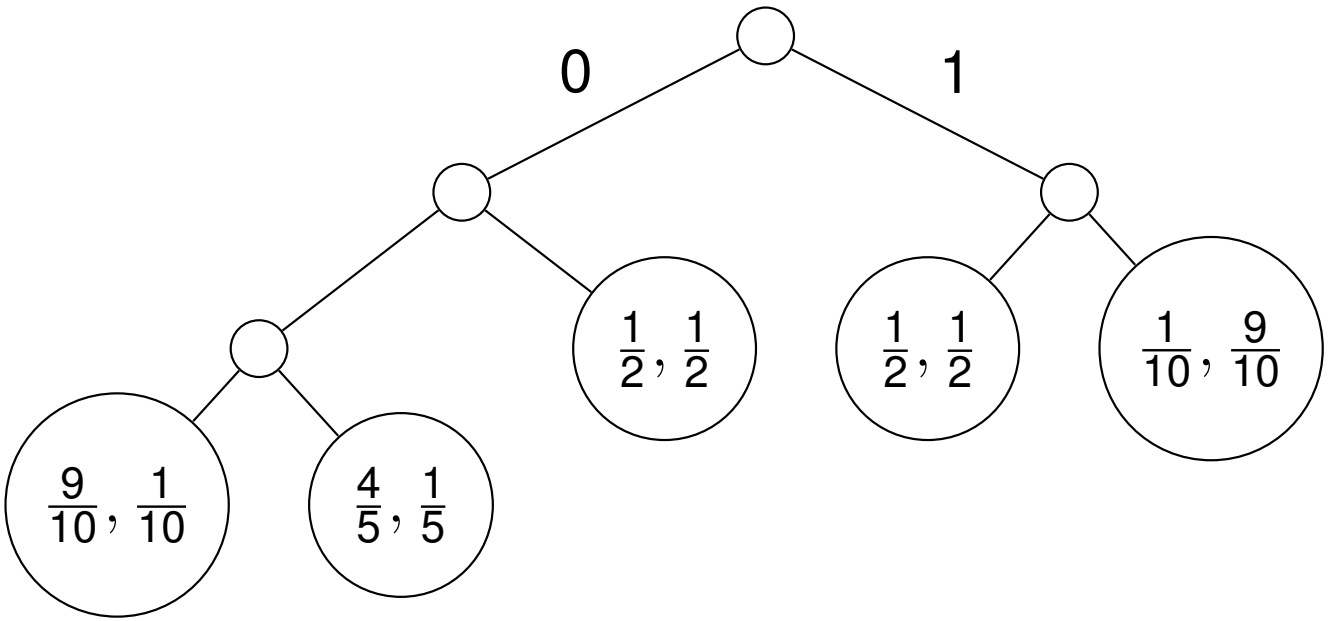

Figure : The same tree, but now representing transition probabilities.

Some notation

- Let $N(w) = \sum_{t=1}^{n-|w|+1} 1_{X_t^{t+|w|-1}=w}$ be the number of occurrences of the subsequence w in our observations. And let $N_{-1}(w)$ be the same except summing up to $n - |w|$.
- This gives estimates of our probability distributions of

$$\hat{P}(w) = N(w)/n, \quad \hat{P}(x|w) = \frac{N(xw)}{N(w)}$$

- We also need a weighted form of relative entropy between a context and that context extended by one symbol.

$$\Delta_{wu} = \sum_{x \in A} \hat{P}(x|wu) \log \left(\frac{\hat{P}(x|wu)}{\hat{P}(x|w)} \right) N(wu)$$

Some notation

- Let $N(w) = \sum_{t=1}^{n-|w|+1} 1_{X_t^{t+|w|-1}=w}$ be the number of occurrences of the subsequence w in our observations. And let $N_{-1}(w)$ be the same except summing up to $n - |w|$.
- This gives estimates of our probability distributions of

$$\hat{P}(w) = N(w)/n, \quad \hat{P}(x|w) = \frac{N(xw)}{N(w)}$$

- We also need a weighted form of relative entropy between a context and that context extended by one symbol.

$$\Delta_{wu} = \sum_{x \in A} \hat{P}(x|wu) \log \left(\frac{\hat{P}(x|wu)}{\hat{P}(x|w)} \right) N(wu)$$

Some notation

- Let $N(w) = \sum_{t=1}^{n-|w|+1} 1_{X_t^{t+|w|-1}=w}$ be the number of occurrences of the subsequence w in our observations. And let $N_{-1}(w)$ be the same except summing up to $n - |w|$.
- This gives estimates of our probability distributions of

$$\hat{P}(w) = N(w)/n, \quad \hat{P}(x|w) = \frac{N(xw)}{N(w)}$$

- We also need a weighted form of relative entropy between a context and that context extended by one symbol.

$$\Delta_{wu} = \sum_{x \in A} \hat{P}(x|wu) \log \left(\frac{\hat{P}(x|wu)}{\hat{P}(x|w)} \right) N(wu)$$

Estimating from data

- Find the largest context tree τ such that for every terminal node $w \in \tau$, $N(w) \geq 2$.
- For each element in the terminal nodes of τ, compute Δ_{wu} and collapse the node wu into w if $\Delta_{wu} < K$. For some cutoff K.
- In practice, this is performed by traversing the tree depth-first and working our way back up.
- What this really amounts to is backward model selection.

Estimating from data

- Find the largest context tree τ such that for every terminal node $w \in \tau$, $N(w) \geq 2$.
- For each element in the terminal nodes of τ, compute Δ_{wu} and collapse the node wu into w if $\Delta_{wu} < K$. For some cutoff K.
- In practice, this is performed by traversing the tree depth-first and working our way back up.
- What this really amounts to is backward model selection.

Estimating from data

- Find the largest context tree τ such that for every terminal node $w \in \tau$, $N(w) \geq 2$.
- For each element in the terminal nodes of τ, compute Δ_{wu} and collapse the node wu into w if $\Delta_{wu} < K$. For some cutoff K.
- In practice, this is performed by traversing the tree depth-first and working our way back up.
- What this really amounts to is backward model selection.

Estimating from data

- Find the largest context tree τ such that for every terminal node $w \in \tau$, $N(w) \geq 2$.
- For each element in the terminal nodes of τ, compute Δ_{wu} and collapse the node wu into w if $\Delta_{wu} < K$. For some cutoff K.
- In practice, this is performed by traversing the tree depth-first and working our way back up.
- What this really amounts to is backward model selection.

Data Time!

- This example will demonstrate the `R` package `VLMC`. Using sleep state data from mice.
- I also have a Python module for which the code will be released shortly.