Sequence Models

Instructor: John Thickstun

Discussion Board: Available on Ed!

Zoom Link: Available on Canvas

Instructor Contact: thickstn@cs.washington.edu

Course Webpage: https://courses.cs.washington.edu/courses/cse599i/20au/

A Small Clarification

- A formal course in probability will carefully distinguish between
    1. Random variables, e.g. $X : \Omega \to \mathbb{R}^d$.
    2. Samples, e.g. $x_1, \ldots, x_n \in \mathbb{R}^d$.
- I will regularly blur the distinction between these objects.
- If it helps, I will never talk about the formal construction Ω in this course.

The Sequential Setting

- Vector sequences $\mathbf{x} \in \mathbb{R}^{T \times d}$ where $x_t \in \mathbb{R}^d$ is the value at time t.
- Given sequences $\mathbf{x}^1, \ldots, \mathbf{x}^n \sim p$, estimate p.
- Ignore temporal structure? Treat the sequences as vectors $\mathbf{x} \in \mathbb{R}^{Td}$?
  - What if T varies for each sequence
  - Can temporal structure help us learn?
  - What if we only have n=1 very long sequence? Can we still learn?

Autoregressive Modeling

- Factor the joint distribution into conditionals:

$$p(\mathbf{x}) = \prod_{t=1}^{T} p(x_t | x_{<t}).$$

- Learn the conditional distributions $\hat{p}(x_t | x_{<t})$.

- Iteratively sample $\hat{x}_t \sim \hat{p}(\cdot | \hat{x}_{<t})$ to construct $\hat{\mathbf{x}} \sim \hat{p}$.

Linear Autoregressive Models

- Scalar linear autoregressive model of order 1: $\mathbf{x} \in \mathbb{R}^{T \times 1}$.
- Randomness is driven by white noise $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$.
- Model is defined by linear dynamics, with parameters $\mu, \rho \in \mathbb{R}$:

$$x_t = \rho(x_{t-1} - \mu) + \mu + \varepsilon_t,$$
$$x_1 = \mu + \varepsilon_0.$$

- Markov model: $p(x_t|x_{<t}) = p(x_t|x_{t-1}) = \mathcal{N}(\rho(x_{t-1} - \mu) + \mu, \sigma^2)$.

Stationary Moments

Definition. *A time series* $\boldsymbol{x}$ *has* ***stationary p'th moment*** *if* $\mathbb{E}[x_t^p]$ *is constant, for all* $t \in \mathbb{N}$.

- Scalar AR(1) model:

$$x_t = \rho(x_{t-1} - \mu) + \mu + \varepsilon_t,$$
$$x_1 = \mu + \varepsilon_0.$$

- The 1'st moment (mean) of an AR(1) model is stationary: $\mathbb{E}[x_t] = \mu$.

Estimating the Mean

- Suppose we want to estimate the mean of process given a sample $\mathbf{x} \in \mathbb{R}^{T \times 1}$.
- Assume that $\mathbf{x}$ has stationary mean $\mathbb{E}[\mathbf{x}] = \mathbb{E}[x_t] = \mu$.
- Use the sample mean as an estimator? $\hat{\mu} = \frac{1}{T} \sum_{t=1}^{T} x_t$.
- Is it consistent? I.e. does $\lim_{T \to \infty} \hat{\mu} = \mu$?

Estimating the Mean

- Is the sample mean consistent? I.e. does $\lim_{T \to \infty} \hat{\mu} = \mu$?

- Chebyshev's inequality: $P(|\hat{\mu} - \mu| \geq \delta) \leq \frac{\mathrm{Var}(\hat{\mu})}{\delta^2}$.

- But what about the variance?

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{T^2} \left(\sum_{t=1}^{T} \mathrm{Var}(x_t) + 2 \sum_{1 \leq s < t \leq T} \mathrm{Cov}(x_s, x_t) \right).$$

Wide-Sense Stationarity

Definition. *A time series X is **wide-sense stationary** if it has stationary mean and*

$$\mathrm{Cov}(x_t, x_{t-k}) = \mathrm{Cov}(x_s, x_{s-k}),\ \forall s, t, k \in \mathbb{N}.$$

*In this case, we define the **autocovariance** function $\gamma : \mathcal{N} \to \mathbb{R}$ by*

$$\gamma(k) = \mathrm{Cov}(x_{k+1}, x_1).$$

$$\mathrm{Var}(\hat{\mu}) = \frac{1}{T^2}\left(T\gamma(0) + 2\sum_{k=1}^{T}(T-k)\gamma(k)\right) = \frac{\gamma(0)}{T} + \frac{2}{T}\sum_{k=1}^{T}\left(1 - \frac{k}{T}\right)\gamma(k) \leq \frac{\gamma(0)}{T} + \frac{2}{T}\sum_{k=1}^{T}\gamma(k).$$

Estimating the Mean

- Putting it all together: $P(|\hat{\mu} - \mu| \geq \delta) \leq \frac{\mathrm{Var}(\hat{\mu})}{\delta^2} \leq \frac{\gamma(0)}{T} + \frac{2}{T}\sum_{k=1}^{T}\gamma(k).$

- If $\sum_{k=1}^{\infty}\gamma(k) < \infty$ then $\mathrm{Var}(\hat{\mu}) = O(1/T)$.

- The faster the autocovariance decays, the easier it is to learn.

Estimating the Mean: AR(1)

- Recall the scalar AR(1) model:

$$x_t = \rho(x_{t-1} - \mu) + \mu + \varepsilon_t,$$
$$x_1 = \mu + \varepsilon_0.$$

- For an AR(1) process $\mathbf{x} \in \mathbb{R}^{T \times 1}$: $\mathrm{Var}[x_t] = \rho^2 \mathrm{Var}[x_{t-1}] + \sigma^2$.

- So $\mathbf{x}$ has stationary variance iff $|\rho| < 1$, in which case $\mathrm{Var}[\mathbf{x}] = \dfrac{\sigma^2}{1 - \rho^2}$.

- By induction, $\gamma(k) = \dfrac{\rho^k \sigma^2}{1 - \rho^2}$ and it follows that $\displaystyle\sum_{k=1}^{\infty} \gamma(k) < \infty$.

5-Minute Break

Vector Autoregression

- We can generalize scalar autoregression to vectors:

$$\mathbf{x} \in \mathbb{R}^{T \times d}, \quad \mu \in \mathbb{R}^d, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma^2 I).$$

- We can also generalize to condition on more history.

- Vector autoregressive model with order p:

$$x_t = \sum_{k=1}^{p} \rho_k (x_{t-k} - \mu) + \mu + \varepsilon_t,$$

$$x_1 = \mu + \varepsilon_0.$$

Parameter Estimation

- Vector AR(p) model:

$$x_t = \sum_{k=1}^{p} \rho_k (x_{t-1} - \mu) + \mu + \varepsilon_t,$$

$$x_1 = \mu + \varepsilon_0.$$

- How to estimate the regression coefficients $\rho \in \mathbb{R}^p$?
- Maximize the conditional likelihood: $p(x_t | x_{t-p}, \ldots, x_{t-1})$.
- This is just a least-squares problem!

Discrete Sequence Modeling

- What about discrete sequences $\mathbf{w} \in \mathcal{V}^T$?
- Tokens $w_t \in \mathcal{V}$ for some finite vocabulary $\mathcal{V}$.
- Tabulate the probability of each sequence? (Large or countable space)
- Tabulate the probability of conditional distributions $p(w_t|w_{<t})$?

n-gram Modeling

- Tabulate the probability of conditional distributions $p(w_t|w_{<t})$?
- Just $|\mathcal{V}|$ items in our table for each conditional, but $O(|\mathcal{V}^T|)$ conditionals.
- Truncate the history: model $p(w_t|w_{t-n+1}, \ldots, w_{t-1})$.
- Only $O(|\mathcal{V}^{n-1}|)$ conditionals to learn: much more reasonable.
- Compare to AR(p)!

n-gram Assumptions

- In n-gram modeling, we make two (false) assumptions about the data.

Definition. *A process is* ***strictly stationary*** *if its joint distribution is invariant to time shifts:*

$$p(w_t, \ldots, w_{t+n}) = p(w_s, \ldots, w_{s+n}), \text{ for all } s, t, n \in \mathbb{N}.$$

- Strict stationarity is equivalent to n'th moment stationarity for all moments n.

Definition. *A process is* ***n'th order Markov*** *if it has limited-horizon temporal dependencies:*

$$p(w_t|w_{<t}) = p(w_t|w_{t-n+1}, \ldots, w_{t-1}), \text{ for all } t \in \mathbb{N}.$$

- n'th order Markov is equivalent to the order p assumption in an AR(p) model.

Preview of Upcoming Topics

- Unify the discrete and continuous sequence modeling perspectives.
- Replace linear AR(p) parameterization with neural parameterizations using a paradigm called the Neural Autoregressive Distribution Estimation (NADE).
- Dig into the details of how to parameterize a NADE.
- Homework 1 goes out later today (on the website).