Latent Autoregressive Models and Language Modeling

Introduction to Latent Autoregressive Models

Since h_t is never directly observed, we refer to it as a latent variable. This leads to what is known as a latent autoregressive model. In statistics, dynamics that remain unchanged over time are referred to as *stationary*.

What is Language Modeling?

Language modeling involves predicting the likelihood of a sequence of words or tokens, such as a sentence.

Why is Language Modeling an Autoregressive Problem?

At first glance, language modeling may not seem like an autoregressive problem. However, it can be treated as one by leveraging the *chain rule of probability*.

Key Concepts in Language Modeling

1. **Joint Probability:**
   - In language modeling, the goal is to calculate the probability of an entire sequence of tokens:
   $$P(x_1, x_2, \ldots, x_n)$$
   - Here, $x_1, x_2, \ldots, x_n$ represent the tokens (e.g., words) in the sequence, and n is the total number of tokens.
2. **Chain Rule of Probability:**
   - The chain rule allows us to decompose the joint probability into a product of conditional probabilities:
   $$P(x_1, x_2, \ldots, x_n) = P(x_1) \cdot P(x_2 \mid x_1) \cdot P(x_3 \mid x_1, x_2) \cdot \ldots \cdot P(x_n \mid x_1, x_2, \ldots, x_{n-1})$$
   - Simplified:
     - Predict $P(x_1)$ (the probability of the first token).
     - Then predict $P(x_2 \mid x_1)$ (the probability of the second token given the first).
     - Continue predicting each token conditioned on the preceding tokens.
3. **Autoregressive Prediction:**

- An autoregressive model predicts each token sequentially, using the preceding tokens as input.
- The chain rule decomposition aligns perfectly with this step-by-step prediction process:

$$P(x_i \mid x_1, x_2, \ldots, x_{i-1})$$

Why is this Important?

This approach transforms language modeling into a sequence of conditional predictions, which is the foundation of autoregressive models such as GPT. For example:

- When generating a sentence, the model predicts the next word one at a time, conditioned on the words already generated.

Summary

Language modeling, though seemingly complex, can be simplified using the chain rule of probability. This reduces the task of predicting the probability of an entire sequence into smaller, manageable steps. This process aligns with the functioning of autoregressive models, which generate tokens sequentially while considering previous tokens.