In The Name of God

Sharif University of Technology

Electrical Engineering Department

Deep Generative Models

Assignment 1

Fall 2024

Instructor: Dr. S. Amini

Due on Mehr 26, 1403 at 23:55

1 Autoregressive Models of Order p

In this question, you will explore the process of Maximum Likelihood Estimation (MLE) for linear Autoregressive (AR) models. Given a time series dataset, your task is to estimate the parameters of an AR model using MLE.

An Autoregressive (AR) model of order p (denoted as AR(p)) is used to describe a time-dependent process and is a classical linear model. The value at any given time is expressed as a linear function of its previous values plus some error term.

This model can be written as:

$$y_t = \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + \epsilon_t$$

where:

- y_t is the value at time t,
- $\phi_1, \phi_2, \ldots, \phi_p$ are the parameters (coefficients) of the model,
- ϵ_t is the error term at time t, assumed to be independently and identically distributed (i.i.d.) with a normal distribution $N(0, \sigma^2)$.

Now, you are supposed to estimate the parameters $\phi_1, \phi_2, \ldots, \phi_p$ and σ^2 that maximize the likelihood function given the observed data $y_1, y_2, \ldots, y_n$.

1.1 Log-Likelihood Function

Formulate the likelihood function for the AR(p) model. Then, derive the log-likelihood function.

1.2 Maximum Likelihood Estimation

Find the parameter values that maximize the log-likelihood function and provide the final estimates for $\phi_1, \phi_2, \ldots, \phi_p$ and σ^2. During your procedure, derive detailed derivation of the likelihood and log-likelihood functions.

2 Autoregressive Models

Given a sequence of random variables $\mathbf{x} = (x_1, x_2, \ldots, x_T)$, where each x_t is a continuous-valued random variable, consider the autoregressive (AR) model where the joint probability distribution of the sequence is factorized as:

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)\cdots p(x_T|x_1, x_2, \ldots, x_{T-1})$$

Now, assume that the conditional distributions $p(x_t|x_1, x_2, \ldots, x_{t-1})$ for all $t \in \{1, \ldots, T\}$ are modeled by a neural network f_θ with parameters θ such that:

$$p(x_t|x_1, x_2, \ldots, x_{t-1}) = \mathcal{N}(f_\theta(x_1, x_2, \ldots, x_{t-1}), \sigma^2)$$

where $\mathcal{N}(\mu, \sigma^2)$ represents a Gaussian distribution with mean μ and fixed variance $\sigma^2 > 0$.

2.1 Maximum Likelihood Estimation

Derive the log-likelihood function for the joint distribution $p(\mathbf{x})$ in terms of θ, and express how you would compute gradients for training the neural network f_θ using backpropagation.

2.2 Predictive Sampling

Describe the process for generating a new sequence $\mathbf{x}' = (x'_1, x'_2, \ldots, x'_T)$ from the model by sampling from the conditional distributions. Explicitly account for how the autoregressive structure influences the sampling process.

2.3 KL Divergence between AR Models

Consider two autoregressive models, one parameterized by f_θ and another by g_ϕ. Write down the expression for the Kullback-Leibler (KL) divergence between the two distributions $p_\theta(\mathbf{x})$ and $q_\phi(\mathbf{x})$. Discuss the computational challenges of directly evaluating this divergence and propose a practical approximation method for high-dimensional sequences.

Hint: You can describe *Monte Carlo* sampling here.

2.4 Stationarity and Long-Term Dependencies

Analyze whether the autoregressive model structure as described can capture long-term dependencies in the sequence. If it cannot, provide a mathematical explanation. Suggest a modification to the autoregressive model (e.g., using a recurrent neural network or Transformer) that might address this limitation and discuss its probabilistic interpretation.

2.5 Bonus

Given the autoregressive factorization, if you observe that the model's predictive performance degrades for longer sequences, hypothesize potential reasons for this degradation. Provide at least two hypotheses grounded in the model's probabilistic structure and optimization process, and suggest corresponding solutions.

3 Real NADE Parameters

In this problem, you will study an extension of the Real NADE model. Recall that, given an autoregressive model

$$p(\boldsymbol{x}) = p(x_1)p(x_2 \mid \boldsymbol{x}_{<2}) \ldots p(x_i \mid \boldsymbol{x}_{<i}) \ldots p(x_n \mid \boldsymbol{x}_{<n}), \tag{1}$$

and Real NADE models the conditional distribution as

$$p(x_1) = \mathcal{N}(x_1 \mid \mu_1, \exp(s_1)) \tag{2}$$

...

$$p(x_i \mid \boldsymbol{x}_{<i}; \boldsymbol{W}, \boldsymbol{c}, \boldsymbol{v}_i, b_i, \boldsymbol{u}_i, d_i) = \mathcal{N}\left(x_i \mid \boldsymbol{v}_i^\top \boldsymbol{h}_i + b_i, \exp(\boldsymbol{u}_i^\top \boldsymbol{h}_i + d_i)\right), \tag{3}$$

where $\boldsymbol{h}_i, \boldsymbol{c}, \boldsymbol{v}_i, \boldsymbol{u}_i \in \mathbb{R}^d$. Now, we would like to make $p(x_i \mid \boldsymbol{x}_{<i})$ follow a mixture of Gaussian

$$p(x_i \mid \boldsymbol{x}_{<i}) = \sum_{c=1}^{C} \pi_i^c \mathcal{N}(\mu_i^c, (\sigma_i^c)^2), \tag{4}$$

where $\sum_{c=1}^{C} \pi_i^c = 1$.

Now the question is: How do you propose to parameterize $\pi_i^c, \mu_i^c, \sigma_i^c, \forall c \in \{1, \ldots, C\}$ as a function of $\boldsymbol{h}_i$? Describe the parameters required and the total number of parameters required for a single $p(x_i \mid \boldsymbol{x}_{<i})$.

4 Monte Carlo Estimation

4.1 A Quick Warm Up

What is the value of this expectation $\mathbb{E}_{\mathsf{x} \sim \mathcal{N}(0,2)}\left[\mathsf{x}^2 + \mathsf{x} + 1\right]$? How do you propose to estimate this quantity with Monte Carlo estimation?

4.2 Variance of K-sample Estimator

What is the variance of the Monte Carlo estimator using K samples?

4.3 Objective Minimization

Now assume we are interested in minimizing the objective

$$F(\theta) = \sum_{n=1}^{N} w_n f(\theta; n) + \lambda R(\theta), \text{where } w_n > 0, \forall n.$$

What is the asymptotic complexity of evaluating $F(\theta)$ and $\nabla F(\theta)$ given θ (assume the function call and gradient evaluation of $f(\theta; n)$ is constant)? How do you propose to use Monte Carlo estimation to acquire an *unbiased* estimation of the objective and its gradient?