ML Week7 Assignment

WANG TZU YI

October 21, 2025

Score Matching and its Role in Score-based Generative Models

1. Concept of Score Matching

Score matching, first proposed by Hyvärinen (2005), is a method for training probabilistic models without directly computing the intractable normalization constant of the probability density function (PDF).

Given a data distribution $p_{\text{data}}(x)$ and a model distribution $p_\theta(x)$, the goal is to make them as close as possible. Instead of maximizing the likelihood $\log p_\theta(x)$, score matching minimizes the difference between their *score functions*, defined as the gradient of the log-density with respect to the data:

$$\nabla_x \log p_\theta(x) \approx \nabla_x \log p_{\text{data}}(x).$$

The score function indicates the direction in which the density increases most rapidly in the data space.

The objective of score matching can be expressed as:

$$\mathcal{L}_{\text{SM}}(\theta) = \mathbb{E}_{p_{\text{data}}} \left[\frac{1}{2} \|\nabla_x \log p_\theta(x) - \nabla_x \log p_{\text{data}}(x)\|^2 \right].$$

Minimizing this loss makes the model's score field match that of the true data distribution. If two score functions are identical, their probability densities differ only by a constant factor and thus represent the same distribution.

2. Denoising Score Matching (DSM)

Directly estimating $\nabla_x \log p_{\text{data}}(x)$ is intractable because $p_{\text{data}}(x)$ is unknown. Denoising Score Matching (Vincent, 2011) solves this problem by perturbing the data with Gaussian noise:

$$x = x_0 + \epsilon, \quad \epsilon \sim \mathcal{N}(0, \sigma^2 I).$$

It then learns to predict the score of the conditional distribution $p(x|x_0)$:

$$\mathcal{L}_{\text{DSM}}(\theta) = \mathbb{E}_{p_{\text{data}}(x_0)p(x|x_0)} \left[\| s_\theta(x, \sigma) - \nabla_x \log p(x|x_0) \|^2 \right],$$

where

$$\nabla_x \log p(x|x_0) = -\frac{x - x_0}{\sigma^2}.$$

Thus, DSM trains a neural network $s_\theta(x, \sigma)$ to estimate the gradient field (score) of noisy data.

3. From DSM to Diffusion Models

Song and Ermon (2019, 2020) generalized DSM into a continuous-time framework. They interpreted the data corruption process as a *diffusion process*:

$$dx = g(t)\, dw_t, \quad x(0) = x_0,$$

where w_t is a Wiener process and $g(t)$ controls the noise level. As time increases, x_t gradually becomes pure Gaussian noise.

The reverse process, which transforms noise back into data, is governed by the following stochastic differential equation (SDE):

$$dx = [f(x, t) - g(t)^2 \nabla_x \log p_t(x)]\, dt + g(t)\, d\bar{w}_t,$$

where the unknown term $\nabla_x \log p_t(x)$ is replaced by the trained score network $s_\theta(x, t)$. By simulating this reverse SDE (or its deterministic ODE form), we can generate new samples starting from random Gaussian noise.

4. Summary

In summary, score matching provides a way to learn the geometric structure of a data distribution by matching its gradient field. Denoising Score Matching extends this idea to noisy data, making it practically trainable. Diffusion models further interpret the corruption and denoising processes as continuous stochastic dynamics, allowing high-quality generation by integrating the learned score field backward in time.

$$\boxed{\text{Score Matching } \to \text{DSM (Noisy Data)} \to \text{Diffusion Process for Generation.}}$$