DGM HW2 Written part

ALT_PDG

2025 年 10 月 30 日

Problem 1

In this problem, you will prove that the Negative Log-Likelihood (NLL) is equivalent to the Cross-Entropy Loss for binary sequence autoregressive models.

Consider a binary sequence $\mathbf{x} = (x_1, x_2, \ldots, x_T)$, where each x_i represents a binary variable, i.e., $x_i \in \{0, 1\}$. The goal of an autoregressive model is to predict the sequence by modeling the conditional probability of each element x_i given all previous elements in the sequence $x_1, x_2, \ldots, x_{i-1}$. An autoregressive model predicts the probability of each binary variable $x_i$as:

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

where $p(x_i \mid x_1, x_2, \ldots, x_{i-1})$ is the probability that the autoregressive model assigns to the event $x_i = 1$, and $1 - p(x_i \mid x_1, x_2, \ldots, x_{i-1})$ is the probability of the event $x_i = 0$.

The Negative Log-Likelihood for the entire sequence x is defined as:

$$\text{NLL}(x) = -\sum_{i=1}^{T} \log p(x_i \mid x_1, x_2, \ldots, x_{i-1}).$$

Here, the NLL measures how well the model predicts the actual observed sequence x. The lower the NLL, the better the model's predictions align with the true sequence. Note that the NLL is a common metric used in the earlier papers on autoregressive models [Van+16].

The Cross-Entropy Loss for the sequence x is defined as:

$$\text{Cross-Entropy Loss} = -\sum_{i=1}^{T} \left(y_i \log p(x_i \mid x_1, x_2, \ldots, x_{i-1}) + (1 - y_i) \log(1 - p(x_i \mid x_1, x_2, \ldots, x_{i-1})) \right)$$

where y_i represents the true label for each x_i (in this binary case, $y_i = x_i$).

Prove that the NLL is equal to the Cross-Entropy Loss for binary sequence autoregressive models. Show all steps clearly.

Proof: By the definition of $p(x_i|x_1, .., x_{i-1})$,

$$\text{NLL}(x) = -\sum_{i=1}^{T} \log p(x_i \mid x_1, x_2, \ldots, x_{i-1}) = -\sum_{i=1}^{T} logP(x_i = 1|x_1, x_2, \ldots, x_{i-1})$$

$$= -\sum_{i=1}^{T} 1 \cdot logP(x_i = 1|x_1, x_2, \ldots, x_{i-1}) + 0 \cdot log(1 - P(x_i = 1|x_1, x_2, \ldots, x_{i-1}))$$

$$= -\sum_{i=1}^{T} x_i logP(x_i = 1|x_1, x_2, \dots, x_{i-1}) + (1 - x_i) \cdot log(1 - P(x_i = 1|x_1, x_2, \dots, x_{i-1}))$$

By the definition that $y_i = x_i$, we know

$$-\sum_{i=1}^{T} y_i logP(x_i = 1|x_1, x_2, \dots, x_{i-1}) + (1 - y_i) \cdot log(1 - P(x_i = 1|x_1, x_2, \dots, x_{i-1}))$$

$$= -\sum_{i=1}^{T} y_i logp(x_i = 1|x_1, x_2, \dots, x_{i-1}) + (1 - y_i) log(1 - p(x_i = 1|x_1, x_2, \dots, x_{i-1})) = Cross - Entropy\ Loss$$

Problem 4

In practice, Maximum Likelihood (ML) estimation (equivalent to minimizing NLL) is often enhanced by incorporating regularization techniques. Regularization biases the ML estimates towards solutions that reflect certain properties, often informed by domain knowledge. One common form of regularization is ℓ_2 regularization, also known as Tikhonov regularization, which is widely used to prevent overfitting by penalizing large parameter values. The ℓ_2 regularization is often applied in the code by using the weight decay parameter in the optimizer, which directly corresponds to the $\lambda\|\theta\|_2^2$ term in the regularization, where θ is the neural network parameter.

It is known that ML estimation with a specific regularizer can be equivalent to Bayesian estimation with a corresponding prior distribution. In this problem, you will explore this equivalence by showing that ML estimation with ℓ_2 regularization is equivalent to Bayesian estimation with a Gaussian prior.

Consider a parameterized model (e.g., an NN) $p_\theta(x; \theta)$ where θ is the network parameter and x is the training data. The regularized maximum likelihood estimate with a penalty $\rho(\cdot)$ is given by:

$$\hat{\theta}_{\text{RML}}(x) = \arg\max_\theta \left[\log p_\theta(x; \theta) - \lambda \rho(\theta)\right]$$

where $\lambda > 0$ is the regularization parameter.

Now, consider the corresponding Bayesian estimation scenario where the parameter θ is treated as a random variable. The Maximum A Posteriori (MAP) estimate is given by:

$$\hat{\theta}_{\text{MAP}}(x) = \arg\max_\theta \left[\log p(\theta \mid x)\right]$$

(a) Using Bayes' theorem, derive the MAP estimator for θ and show that it can be rewritten in a form similar to the regularized ML estimator. Hint: The non-Bayesian case of $p_\theta(x; \theta)$ can be treated as equal to the Bayesian case of $p(x \mid \theta)$, where the former θ is a parameter value and the latter θ is a random variable.

(b) Demonstrate the equivalence between the regularized ML estimator $\hat{\theta}_{\text{RML}}(x)$ with ℓ_2 regularization and the MAP estimator $\hat{\theta}_{\text{MAP}}(x)$ with a Gaussian prior. Specifically, show what the Gaussian prior (its mean and variance) should be in the Bayesian setting to make the two estimators equivalent.

Proof:(a) By Bayes' theorem we know

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = p(x|\theta)\frac{p(\theta)}{p(x)}$$

Hence

$$\hat{\theta}_{\text{MAP}}(x) = \arg\max_\theta \left[\log p(x|\theta) + logp(\theta) - logp(x)\right]$$

Since $p(x)$ is irrelevant with θ, we have

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} \left[\log p(x|\theta) + logp(\theta)\right] = \arg\max_{\theta} \left[\log p_{\theta}(x;\theta) + logp(\theta)\right]$$

(b) Assume the Gaussian prior of $\hat{\theta}_{MAP}(x)$ is μ and $\sigma^2 \in R^N$(we use the diagonal Gaussian distribution), which means

$$p(\theta) = \frac{1}{\sqrt{2\pi}^N \prod_{i=1}^{N} \sigma_i} e^{-\sum_{i=1}^{N} \frac{(\theta_i-\mu_i)^2}{2\sigma_i^2}}$$

$$logp(\theta) = -\sum_{i=1}^{N} \frac{(\theta_i - \mu_i)^2}{2\sigma_i^2} - \sum_{i=1}^{N} log\sigma_i - \frac{N}{2} log2\pi$$

Hence

$$\begin{aligned}\hat{\theta}_{\mathrm{MAP}}(x) &= \arg\max_{\theta} \left[\log p_{\theta}(x;\theta) - \sum_{i=1}^{N} \frac{(\theta_i - \mu_i)^2}{2\sigma_i^2} - \sum_{i=1}^{N} log\sigma_i - \frac{N}{2} log2\pi\right] \\ &= \arg\max_{\theta} \left[\log p_{\theta}(x;\theta) - \sum_{i=1}^{N} \frac{(\theta_i - \mu_i)^2}{2\sigma_i^2} - \sum_{i=1}^{N} log\sigma_i\right]\end{aligned}$$

Now if $\mu_i = 0$ and $\sigma_i = \frac{1}{\sqrt{2\lambda}}$, we have

$$\arg\max_{\theta} \left[\log p_{\theta}(x;\theta) - \lambda\|\theta\|^2 - \sum_{i=1}^{N} log\sigma_i\right]$$

Since this time σ_i(constant defined by λ) are irrelevant with θ, we have

$$\hat{\theta}_{\mathrm{MAP}}(x) = \arg\max_{\theta} \left[\log p_{\theta}(x;\theta) - \lambda\|\theta\|^2\right]$$