1. MSE vs Cross-Entropy in Classification

問題：

In class, we discussed that MSE is often used for regression, while classification uses losses like Cross Entropy. Why?

Will we learn classification-specific loss functions? How does MSE perform regarding numerical stability?

回答與解釋：

MSE 適合 Regression 的原因：

MSE 等價於假設資料符合 Gaussian noise 的最大似然估計，因此用於連續值預測非常合理。

Cross-Entropy 適合 Classification 的原因：

分類問題中，標籤服從 Bernoulli 或 Categorical 分布，其最大似然等價於 Cross-Entropy。使用 logits + BCEWithLogitsLoss 時梯度更穩定、收斂更快。

MSE 用在分類的問題：

搭配 sigmoid/softmax 時產生多一項飽和導數 $\sigma'(z)$，當 z 很大或很小時梯度趨近 0，學習變得極慢，且對錯誤分類不敏感，數值表現差於 CE。

之後會遇到的分類 Loss：

weighted CE、focal loss、hinge loss (SVM)，這些都會在進階 ML 或 DL 課程看到。

相關文獻：

https://papers.nips.cc/paper/2016/hash/33f2d1c4fdf7d5a2b8b0bb2745078e1c-Abstract.html

https://arxiv.org/abs/1708.02002

https://arxiv.org/abs/1901.07884

2. Sigmoid 與 Vanishing Gradient 問題

當初提問問題：

Why is the sigmoid activation prone to vanishing gradients? Is there any alternatives or solutions?

回答與解釋：

Sigmoid 容易造成梯度消失的原因：

Sigmoid 的導數 $\sigma'(z) = \sigma(z)(1-\sigma(z))$。在 $z \to \pm\infty$ 時，$\sigma(z) \to 0$ or 1，因此導數趨近 0。深度網路使用鏈式法則會讓梯度逐層相乘，造成梯度快速衰減。

替代方法：

Activation：ReLU、Leaky ReLU、ELU、GELU、Swish（皆較不容易飽和）。

訓練技巧：BatchNorm、Residual connections、He/Xavier 初始化。

相關文獻：

-https://arxiv.org/abs/1502.01852

-https://arxiv.org/abs/1603.00391

-https://arxiv.org/abs/1706.02515

3. GDA 為何假設 Gaussian？若資料不是 Gaussian 有何後果？如何改善？

當時提問問題：

Why can each category be assumed to be Gaussian? If not Gaussian, what happens? How to improve?

回答與解釋：

假設 Gaussian 的原因：

1. 數學 tractable：高斯分布的對數似然、MLE、邊際分布都有 closed form。
2. 參數少、樣本效率高。
3. 在許多自然資料裡，特徵常由多個因素累積，近似 Gaussian（CLT）。

若資料不是 Gaussian 的後果：

1. Decision boundary 錯誤（誤設模型）
2. 機率校準差。
3. 類別呈多峰分布時，LDA/QDA 會表現不佳。

改進方式：

1. 使用 **Gaussian Mixture Models**（多峰）。
2. 放鬆協方差矩陣（QDA）。
3. 非參數密度估計。
4. 改用 logistic regression 或 neural nets 等 discriminative 模型。

相關文獻：

- https://www.cs.cmu.edu/~tom/mlbook.html
- https://arxiv.org/abs/1805.07451
- https://www.stat.cmu.edu/~larry/all-of-statistics/

4. 為何 MLE 難用於非正規化模型？Score Matching 為何可行？

當時提問問題：

Why is MLE hard for unnormalized models while score matching works?

回答與解釋：

非正規化模型的問題：

1. $p(x) = \tilde{p}(x) / Z(\theta)$，其中 $Z(\theta)=\int \tilde{p}(x)dx$ 通常不可積分。
2. MLE 必須計算 $\partial \log Z / \partial \theta$，其中包含對模型分布的期望，是 computationally intractable。

Score Matching 的突破：

使用 $s\theta(x) = \nabla x \log p\theta(x)$，其中 $\nabla x \log p\theta(x) = \nabla x \log \tilde{p}\theta(x) - \nabla x \log Z(\theta)$，因為 $Z(\theta)$ 與 x 無關 → $\nabla x \log Z(\theta)=0$，因此 score matching 完全不用 $Z(\theta)$。因此能訓練 energy-based models 或 diffusion models 中的 score network。

相關文獻：

Hyvärinen (2005) Score Matching: https://www.cs.helsinki.fi/u/ahyvarin/papers/NCR-Hyvarinen2005.pdf

https://arxiv.org/abs/1907.05600

https://arxiv.org/abs/2011.13456

5. Noise-scale Weight λ(σ) 如何設計？其效果是什麼？

當初提問問題：

How should λ(σ) be designed in noise-conditioned score matching? What are the impacts?

回答與解釋：

為什麼需要 λ(σ)：

Score 在噪聲 σ 下大小約為 $O(1/\sigma)$，若不加權，某些噪聲層級的 loss 會壓過其他層級。λ(σ) 能平衡不同 σ 的貢獻。

常見 λ(σ) 設計：

1. $\lambda(\sigma) \propto \sigma^2$（NCSN 中最常見，能平衡 loss 尺度）。
2. λ(σ) = 常數（訓練較穩定但細節較差）。
3. $\lambda(\sigma) \propto 1/\sigma^2$（強化細節，但訓練不穩）。

影響：

1. 偏向大 σ → global structure 好但細節差。
2. 偏向小 σ → 細節好但取樣不穩。

3. σ^2 → 在穩定與細節間取得最佳平衡。

相關文獻：

- https://arxiv.org/abs/1907.05600

- https://arxiv.org/abs/2011.13456

- https://arxiv.org/abs/2011.13456