National Tsing Hua University
1130IEEM 513600
Deep Learning and Industrial Applications
Homework 2

Name: 巫宛芸 Student ID: 113034601

Due on 2025.03.27

1. (20 pts) Select 2 hyper-parameters of the artificial neural network used in Lab 2 and set 3 different values for each. Perform experiments to compare the effects of varying these hyper-parameters on the loss and accuracy metrics across the training, validation, and test datasets. Present your findings with appropriate tables.

選擇 hyper-parameters : hidden units 與 learning rate。其中，hidden units 分別設定為 64、128、256，而 learning rate 設定為 1e-2、1e-3、1e-4，共進行 9 組比較，結果整理於下表。

Hidden Units	Learning Rate	Train Accuracy	Val Accuracy	Test Accuracy	Train Loss	Val Loss
256	1e-2	87.83%	67.74%	84.00%	0.2986	0.5046
256	1e-3	85.71%	79.01%	77.42%	0.3759	0.4431
256	1e-4	77.78%	65.43%	64.52%	0.4629	0.7355
128	1e-2	87.30%	76.54%	80.65%	0.3027	0.5072
128	1e-3	86.77%	65.43%	61.29%	0.3483	0.5734
128	1e-4	71.96%	77.78%	67.74%	0.5433	0.5393
64	1e-2	86.24%	81.48%	74.19%	0.3421	0.4167
64	1e-3	80.95%	76.54%	70.97%	0.4305	0.5141
64	1e-4	76.19%	66.67%	64.52%	0.5417	0.5552

2. (20 pts) Based on your experiments in Question 1, analyze the outcomes. What differences do you observe with the changes in hyper-parameters? Discuss whether these adjustments contributed to improvements in model performance, you can use plots to support your points. (Approximately 100 words.)

- Learning rate：為 1e-2 時模型收斂速度較快，並且在 hidden=256 時達到最高準確率；1e-3 則表現穩定，1e-4 整體準確率略低。
- Hidden units：數量越多模型表現越穩定，準確率也較高，尤其在 hidden units =256 的表現最佳。

  因此可以觀察到，適中的學習率（1e-3 或 1e-2）搭配較大的 Hidden units，有助於提升模型準確率。

3. (20 pts) In Lab 2, you may have noticed a discrepancy in accuracy between the training and test datasets. What do you think causes this occurrence? Discuss potential reasons for the gap in accuracy. (Approximately 100 words.)

可能有以下幾個原因：

- 資料量偏少：
  訓練資料不夠多，容易讓模型記住訓練資料的細節，而不是學會真正有用的規律。
- 模型太複雜：
  網路層數和參數太多，可能對訓練資料過度學習，在沒看過的測試資料上表現反而變差。
- 資料切分方式可能不均：
  切分訓練與測試資料時，沒特別控制類別比例，導致兩者分布不同。
- 特徵處理不夠細緻：
  像是胸痛類型這種類別資料只簡單轉數字，可能讓模型難以學到有意義的差異。

4. (20 pts) Discuss methodologies for selecting relevant features in a tabular dataset for machine learning models. Highlight the importance of feature selection and how it can impact model performance. You are encouraged to consult external resources to support your arguments. Please cite any sources you refer to. (Approximately 100 words, , excluding reference.)

根據 Chen et al. (2020)，特徵選擇可以幫助降低模型複雜度、訓練時間與過擬合風險，並改善模型的準確率與泛化能力。

常見的方法有：

1. Random Forest - varImp（Feature Importance）

- 利用樹模型中的分裂次數與資訊增益來計算每個特徵的重要程度。
- 優點：簡單直觀，可快速初步篩選關鍵特徵。
- 對效能影響：可移除權重極低的特徵，減少雜訊，提升模型準確率與訓練效率。

2. Boruta

- 基於 Random Forest 的全域特徵選擇法，透過加入隨機雜訊特徵（shadow features）與原始特徵比較，保留顯著性高者。
- 優點：更保守且穩健，不容易錯殺有用特徵。
- 對效能影響：強調「不漏掉重要特徵」，適合高維度或特徵意義不明的資料，有助提升模型穩定性與泛化能力。

3. RFE（Recursive Feature Elimination）

- 從全部特徵開始訓練模型，逐步遞迴移除影響最小的特徵。
- 優點：能搭配任意模型使用，精細控制最終保留的特徵數量。

- 對效能影響：可有效移除冗餘特徵，減少過擬合風險，提升模型解釋性與效能。
-

以上舉的幾種方法可以有效挑出關鍵變數。實驗也顯示使用經過特徵選擇的模型通常比使用全部特徵時表現更佳。

參考文獻：Chen et al., *Selecting critical features for data classification based on machine learning methods*, Journal of Big Data, 2020.
https://journalofbigdata.springeropen.com/articles/10.1186/s40537-020-00327-4

5. (20 pts) While artificial neural networks (ANNs) are versatile, they may not always be the most efficient choice for handling tabular data. Identify and describe an alternative deep learning model that is better suited for tabular datasets. Explain the rationale behind its design specifically for tabular data, including its key features and advantages. Ensure you to reference any external sources you consult. (Approximately 150 words, excluding reference.)

在處理 tabular datasets 時，TabNet 和 FT-Transformer 是比 ANN 更合適的深度學習模型。

- TabNet 是 Google 的模型，利用注意力機制和逐步決策方式，自動挑出每筆資料中最重要的特徵，此模型準確率高，並且具有可解釋性。
- FT-Transformer 則是把每個欄位當作 token，透過 Transformer 的注意力機制來分析特徵之間的關係，適合包含很多類別特徵的資料。

由於這兩個模型都是針對 tabular datasets 的特性所設計，因此比起一般 ANN 更能提升模型效能，也比較不容易過擬合。

參考文獻：

Arik, S. Ö., & Pfister, T. (2021). TabNet: Attentive Interpretable Tabular Learning. Proceedings of the AAAI Conference on Artificial Intelligence, 35(8), 6679-6687.
https://doi.org/10.1609/aaai.v35i8.16826
Gorishniy et al., Revisiting Deep Learning Models for Tabular Data, 2021.
https://openreview.net/pdf?id=i_Q1yrOegLY