National Tsing Hua University

11220IEEM 513600

Deep Learning and Industrial Applications

Homework 2

Name:曾聖閔 Student ID:112034564

Due on 2024.03.21

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.

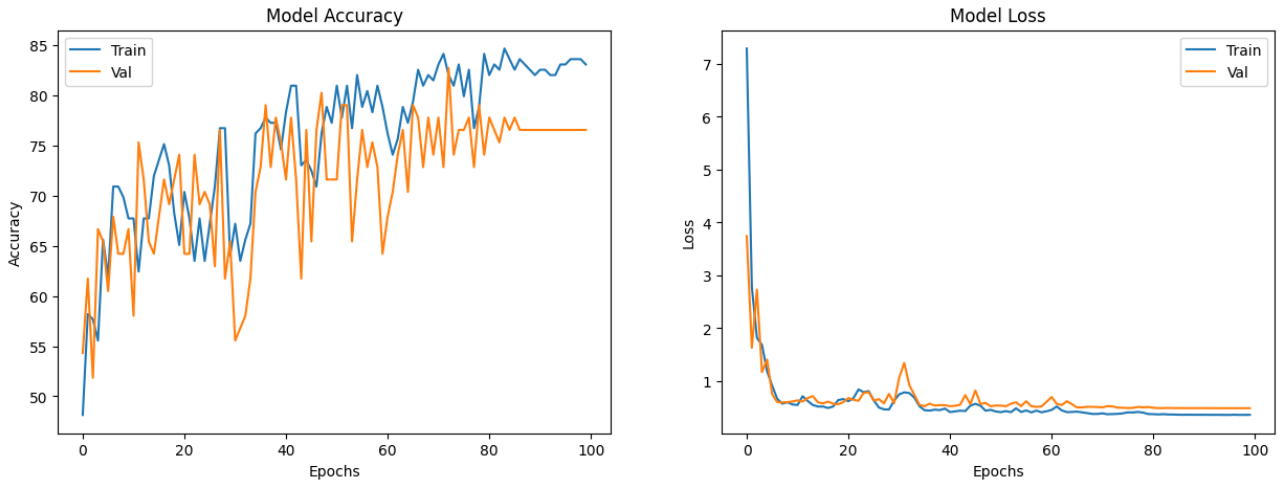
Epoch	Train acc	Train loss	Best Val acc	Best Val loss	Test acc
100	83.5979%	0.3357	81.48%	0.5685	80.64516%
50	80.9524%	0.4529	75.31%	0.5823	67.7419%
500	88.889%	0.1406	82.72%	0.4380	80.64516

Learn rate	Train acc	Train loss	Best Val acc	Best Val loss	Test acc
0.01	89.9471%	0.2610	79.01%	0.544	80.64516%
0.001	85.7143%	0.3554	83.95%	0.4466	74.1935%
0.1	55.0265%	0.6881	53.09%	0.6959	48.3871%

根據微調 epoch、lr，要適度的調整每一個超參數，太大會容易造成 overfitting，太小的話可能導致無法收斂，因此，可使用網格搜尋法縮短調參時間，若無法收斂恐會犧牲搜尋時間過長。

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.)

調整參數會對模型效能影響較大，如 epoch、learn-rate 等參數，部分參數則影響較小。調參主要避免 overfitting 或收斂速度太慢，以圖為例，若是較小的 epoch，訓練時間雖然很快，但容易導致模型還沒有收斂好，acc 會太低，超過 100 之後容易會過度擬合，也使得 acc 更低。因此需選擇適當的超參數，以發揮模型最大效能。



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.)

訓練集和測試集之間準確度的差異通常可稱泛化差異，其中可能因為以下幾點原因。

- 過擬合原因：模型可能對訓練數據過於擬合，捕捉到噪聲和異常值，而不是學習到一般性的模式。因此，在訓練集上表現良好，但在未見過的數據上表現不佳。
- 模型複雜度：如果模型相對於訓練數據的大小過於複雜，它可能會捕捉到噪聲而不是潛在的模式，導致過擬合。
- 數據不匹配：訓練集和測試集之間存在差異，例如分佈的變化或數據質量的差異，可能導致性能差異。
- 訓練不足：如果模型未經過足夠的訓練周期，或者學習率過低，可能尚未收斂到最優解，從而導致泛化能力差。

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.)

   在機器學習模型中選擇相關特徵的方法有很多，以下是一些常見的方法：

   1. Filter Methods：這些方法通過統計測量（如相關係數、卡方檢驗）來對特徵進行評估和排序，然後選擇排名最高的特徵。這些方法不考慮模型，因此計算效率高，但可能會忽略特徵之間的相互作用。

   2. Wrapper Methods：這些方法使用特定的機器學習算法來評估特徵的子集，例如遞歸特徵消除（RFE）或正向選擇。這些方法通常更準確，但也更耗時。

   3. Embedded Methods：這些方法在模型訓練過程中自動進行特徵選擇，例如 Lasso 和 Ridge 回歸中的 L1 正則化，或者決策樹和集成方法中的特徵重要性評估。

   特徵選擇的重要性在於：

   1. 簡化模型：選擇相關特徵有助於簡化模型，減少過度擬合的風險，並提高模型的解釋性。

   2. 降低計算成本：僅使用相關特徵可以減少計算和存儲需求，尤其對於大型數據集來說更加重要。

   3. 提高預測性能：適當地選擇相關特徵可以提高模型的預測性能，因為模型更專注於重要的特徵，從而降低了噪聲的影響。

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 to reference any external sources you consult. (Approximately 150 words, , excluding reference.)

   一個更適合處理表格數據的深度學習模型是 TabNet（Tabular Neural Network）。TabNet 專門設計用於處理結構化數據，如表格數據，具有選擇性注意機制和適應性特徵選擇的關鍵特徵。它通過動態地學習和選擇重要的特徵，能夠更好地捕捉特徵之間的複雜關係，提高了模型的泛化能力。此外，TabNet 還具有過程解釋性，能夠提供對模型預測過程的

解釋，增強了模型的可解釋性。相比於傳統的神經網絡，TabNet 設計了一個輕量級的神經網絡結構，具有較低的計算和存儲需求，能夠更高效地處理大規模的表格數據。

[參考資料]

"TabNet: Attentive Interpretable Tabular Learning", Arik, S., Cao, K., Zhou, T., Vincent, P., "arXiv:1908.07442 [cs, stat]", 2020.

Guyon, I., & Elisseeff, A. (2003). "An introduction to variable and feature selection." Journal of machine learning research, 3(Mar), 1157-1182.

GPT