National Tsing Hua University

11220IEEM 513600

Deep Learning and Industrial Applications

Homework 2

Name: Student ID:

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.**

我嘗試修改 Epoch(50, 100, 150)和 Learning rate(0.001, 0.01, 0.02)兩個超參數。比較各自的組合 Training 後，對於 Test 的 Acc 結果如下表：

LR \ Epoch	50	100	150
0.001	67.74	64.52	61.29
0.01	77.42	74.19	70.97
0.02	70.97	67.74	67.74

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.)**

根據上表數據表格可以得出以下發現(參考自網路資料&Chatgpt 總結)：

- **Learning Rate 的選擇**：在所有組合中，LR 為 0.01 的組合在不同的時期下表現最為穩定，並且通常能夠獲得較高的測試準確度。
- **Epoch 數量的影響**：在其他條件相同的情況下，隨著 Epoch 數量的增加，模型的性能並不總是穩定提升。過多的 Epoch 可能導致模型過擬合或收斂速度過慢，而過少的時期則可能導致欠擬合。

- **調整超參數的重要性**：選擇合適的學習率和時期數量組合能夠顯著影響模型的性能和收斂速度。未來可以使用超參數組合選擇的 package 來求解適當的超參數組合(e.g. Optuna 或使用 FastAi)

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.)

我認為可能是在資料的前處理不足夠導致的，由於神經網路是一以機器邏輯的預測，因此通常如果前處理未更仔細的話，容易因為一些小幅度不同就飄掉，最後導致訓練結果無法適用於 test file。

4. 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.)

特徵選擇除了可以使用相關係數直接選擇(Pearson or Spearman)，還可以直接透過正則化(Regularization)處理，其中正則化包含有 L1(Lasso regression)和 L2(Ridge regression)兩種方式，透過給予懲罰項給與某些參數限制。

5. 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.)

對比傳統機器學習方法（如支持向量機、回歸分析或 XGBoost 等）和神經網路，可以發現它們的訓練方式確實存在差異。傳統機器學習方法更多地側重於從人的角度出發，通過事先定義的特徵和規則來訓練模型，因此在具有一定規律或結構化信息的表格型數據上表現良好。這是因為在這類數據中，人們可以更容易地識別出重要的特徵和模式，從而進行適當的特徵工程，並選擇合適的模型進行訓練。

相比之下，神經網路則更傾向於從數據中自動學習特徵和模式，更多地體現了從機器的角度出發。在表格型數據中，尤其是在特徵較少或者特徵之間相關性不強的情況下，神經網路可能會面臨更多的挑戰，因為其需要大量的數據來發現隱藏在數據中的複雜模式，並且需要花費更多的時間和計算資源來訓練，也容易誤取到不重要的特徵。

因此，在處理具有一定規律或結構化信息的表格型數據時，首先嘗試傳統機器學習方法是比較適當的。這些方法通常能夠更容易地抓住數據中的特徵和模式，從而在實踐中取得更好的結果。如果傳統機器學習方法無法滿足需求，或者數據具有複雜的非線性關係，並且擁有足夠的數據量和計算資源，那麼再考慮嘗試神經網路等深度學習方法。