Simple Classifier / Logistic Regression

B083040029 邱品諺

本次作業利用Logistic Regression預測房價高低，資料集來自於HousingPrice dataset，此資料集1460筆資料，每筆資料有81個特徵。

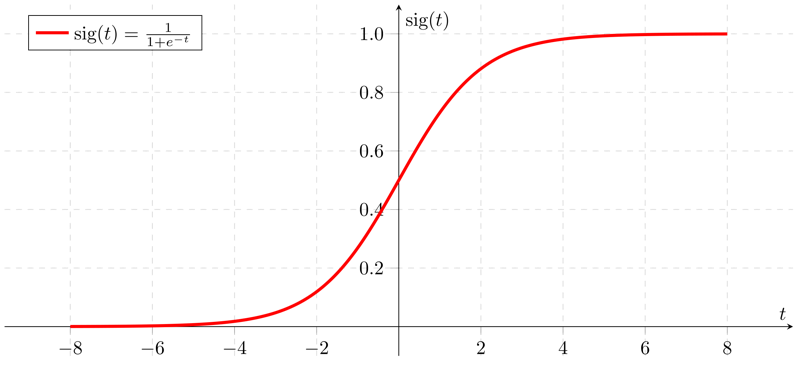
1. Dataloading and Data Preprocessing

載入資料的部分，training, validation, testing資料的比例分別為60%, 20%, 20%。以資料集的GrLivArea欄位為每一筆資料的訓練特徵，以資料集的SalePrice欄位作為每一筆資料的目標預測欄位，並將每筆數值資料進行正規化（Normalization），再以np.stack函式將資料轉換成符合矩陣運算的shape。最後，此部分最後處理完的資料集僅保留了部分資料，因其以ground truth的第30百分位數及第70百分位數為基準，將大於第70百分位數的ground truth設為1、將小於第30百分位數的ground truth設為0，並刪去其他的資料，另外，此問題也變為一個Binary Classification的問題。

1. Set up a Classifier Model

此為Logistic Regression的Model，可以看作Linear Regression再加上一個非線性的Sigmoid Function。

此為Sigmoid Function，輸出的範圍為[0, 1]。以本問題而言，0為低價、1為高價。



1. Loss: Binary Cross Entropy

由於此問題為Binary Classification，因此挑選BCE為Loss Function。Loss Function計算出的數值越低，表示預測值與ground truth愈接近，誤差值越小，準確率越高，反之亦然。另外，由於BCE是一個non-convex function，所以它沒有closed-form解可以直接將解算出來，需要利用numeric methods算出，像是接下來將使用的Gradient Descent。

1. Backpropagation

Backpropagation的主要目的為透過計算Loss Function對Network中各個weight的Gradient，以得知每個參數需要調整的方向。因為如果想要得到某個Function的極值，在微積分中我們可以利用微分=0求得，此處即使用此概念。

而Loss Function對weight的微分藉由Chain Rule其實可表示為Loss Function對Prediction Value的Gradient乘上Prediction Value對weight的Gradient，其中 Prediction Value對weight的Gradient又可利用Chain Rule推為Sigmoid Function對的Gradient乘上對weight的Gradient。

以下為Loss對weight微分經過Chain Rule拆解的結果：

1. Optimizer and Gradient Descent

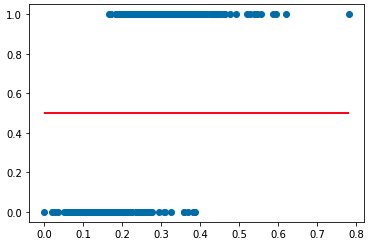
定義好Loss Function後，接著需要Optimizer一步一步的更新Model中的參數，而更新的方法為Gradient Descent，其透過向Gradient的反方向持續向低谷（最小值）前進，直到Converge亦或是達到規定的訓練次數。

Gradient Descent 步驟：

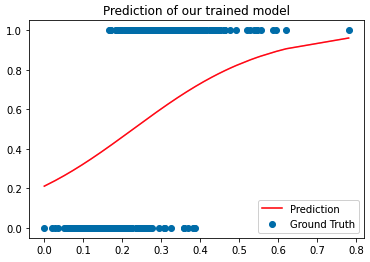
1. 隨機初始化Network中的weight
2. 計算Loss Function
3. 計算Loss Function對各weight的Gradient
4. 依據Gradient更新weight
5. 持續執行2~4直到訓練完畢
6. Training & Solver

訓練中除了要執行上述Gradient Descent每個epoch更新weights的步驟外，訓練之前需要先決定幾個超參數（Hyperparameter），像是Learning rate、epochs (steps)、Batch size等等，而現今多數的訓練方式都一個一個batch下去訓練並計算平均的Gradient以更新Weights，直到訓練至設定的epochs數模型才訓練完畢。以下為本問題不同的訓練次數所訓練出Model結果（藍點為資料點、紅線為預測回歸線）：

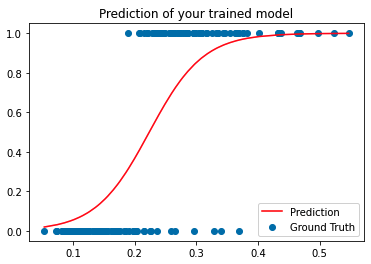
1. 無訓練



1. 400 training steps



1. 25000 training steps



我們可以看到訓練到一定的次數，Model會學得較好、較貼近資料的曲線，Loss也會比較低，也表示準確率會比較高。