Machine Learning Project 2

50K Predictor

Team Members

- B03902089 資工三 林良翰

Questions

- 我在X_{train}裡的每一筆資料都是個row vector，因此μ、Σ、w都是以row為單位，轉置之後才是column vector

1. 請說明你實作的generative model，其訓練方式和準確率為何？
   - 讀入訓練資料X_{train}以及每個資料的標記Y_{train}，然後將資料X_{train}中每一筆$y=0$和$y=1$分別作平均μ_1、μ_2和共變數(Covariance)矩陣Σ_1、Σ_2，然後再用Σ_1、Σ_2以及兩個種類的數量比例p_1、p_2算出共用的共變數矩陣Σ
   - 計算出共變數矩陣Σ的行列式值$|\Sigma|$和反矩陣Σ^{-1}
   - 計算出常態分佈(Guassian Distribution)函數中自然對數以外的數字(減少之後迴圈計算量)：$C=\left((2\pi)^D\ |\Sigma|\right)^{-\frac{1}{2}}$。
   - 讀入測試資料X_{test}，將每一筆測試資料帶入常態分佈函數
     $f_{\mu_1,\Sigma_1}(x)=\left((2\pi)^D\ |\Sigma|\right)^{-\frac{1}{2}}exp\left(-\frac{1}{2}(x-\mu_1)^T(\Sigma_1)^{-1}(x-\mu_1)\right)$
     $f_{\mu_2,\Sigma_2}(x)=\left((2\pi)^D\ |\Sigma|\right)^{-\frac{1}{2}}exp\left(-\frac{1}{2}(x-\mu_2)^T(\Sigma_2)^{-1}(x-\mu_2)\right)$
     計算出每一筆資料他在第一類的機率：$P_1=\frac{p_1 f_{\mu_1,\Sigma_1}(x)}{p_1 f_{\mu_1,\Sigma_1}(x)+p_2 f_{\mu_2,\Sigma_2}(x)}$
     如果$P_1\geq 0.5$歸類在第一類，$P_1<0.5$則歸類在第二類
   - 輸出結果，上傳Kaggle後得到84.103%的準確率
2. 請說明你實作的discriminative model，其訓練方式和準確率為何？
   - 讀入訓練資料$X_{train}=\begin{bmatrix}x_1\\ \vdots\\ x_n\end{bmatrix}$以及每個資料的標記$Y_{train}=\begin{bmatrix}y_1\\ \vdots\\ y_n\end{bmatrix}$
   - 對第0, 1, 3, 4, 5個特徵做標準化(Normalization)
     $X=\frac{X-\mu}{\sigma}$，μ是平均值(mean)，σ是標準差(Standard Deviation)
   - 初始化$w=0.001\times[random\,(-1\sim 1)]$、$b=random\,(-1\sim 1)$，使用隨機種子$random.seed(100)$
   - 訓練迴圈內：
     - 對每一筆資料X_i算出$z=x_i\cdot w+b$

        - 將w帶入Sigmoid函數$\sigma\left(z\right)=\frac{1}{1+e^{-z}}$算出機率$p$
        - 計算誤差函數對w和b的梯度(Gradient)$\frac{\partial L}{\partial w}=\left(p-y_i\right)x_i$、$\frac{\partial L}{\partial b}=\left(p-y_i\right)$
    - 加總每一筆資料的梯度後，乘上學習比率(Learning Rate)，算出來的值拿去減現在的w和b
      $w_{new}=w_{old}-\eta\sum_{i=1}^{n}\left(p-y_i\right)x_i$
      $b_{new}=b_{old}-\eta\sum_{i=1}^{n}\left(p-y_i\right)$
    - 重複迴圈動作，直到訓練資料內的準確率上升的幅度夠小才停止
    - 讀入測試資料X_{test}，對每一筆測試資料帶入Sigmoid函數算出它屬於第一類的機率，大於0.5就歸類為第一類，反之則第二類。
    - 輸出結果，上傳Kaggle後得到85.43%的準確率
3. 請實作輸入特徵標準化(feature normalization)，並討論其對於你的模型準確率的影響。
    - 使用Logistic Regression，使用所有特徵，100 iteration，並上傳Kaggle測試結果
    - 沒有用特徵標準化的話，Learning Rate=10^{-9}，準確率最高達到76.032%
    - 使用標準化之後，Learning Rate=10^{-4}，準確率最高達到85.43%
    - 如果資料沒有標準化的話，很容易因為每一種特中的分佈都不一樣，影響到訓練的準確度，例如第二個特徵和其他特徵比起來，數值大上非常多，因此在計算Sigmoid的時候很容易被這個特徵所影響，造成訓練上的限制。
4. 請實作logistic regression的正規化(regularization)，並討論其對於你的模型準確率的影響。
    - 使用所有特徵、標準化(Normalization)第0, 1, 3, 4, 5個特徵
    - Learning Rate=10^{-4}，$\lambda=0.01$，Iteration=100
    - 測試過後的結果發現，正規化出來的結果會稍微好一些，準確率為84.496%
5. 請討論你認為哪個attribute對結果影響最大？
    - 使用Probalistic Generative Model分析
    - 要把這106種特徵做排列組合去看每個特徵對訓練成果的影響是無法做到的，其計算量是目前電腦做不到的，所以我自己用一種比較簡單而且計算量不高的方法去分析每個特徵對訓練成果的影響幅度。
    - $i=1\sim105$，使用第$0\sim i$個特徵作為訓練特徵，在逐一加入新的特徵的過程中，算出他準確率變化的幅度$\left(Acc_i-Acc_{i-1}\right)\times i$，當作是該特徵對訓練成果的影響大小。
    - $i=105\sim0$，使用第$i\sim0$個特徵作為訓練特徵，在逐一加入新的特徵的過程中，算出他準確率變化的幅度$(Acc_i-Acc_{i+1})\times(106-i)$，當作是該特徵對訓練成果的影響大小。
    - 最後綜合上述兩次的結果，我發現第$0, 3, 4, 5, 10, 24, 27, 29, 33, 41, 47, 54$有較大的影響力，然而在第$6$之後的特徵都只有零跟一的變化，因此第0, 3, 4, 5個特徵是影響最大的幾個。
    - 如果要說影響最大的特徵，我認為應該是第0個特徵，也就是年齡，因為當我拿掉年齡之後，準確率會直接降到80%左右。