Machine Learning Report #2

B03902015 資工三 簡瑋德

1. 請說明你實作的generative model，其訓練方式和準確率為何？

- Feature的部分，只考慮「age」這一項，並視為discreate的數值
- 目標函數定為$\prod_{(y,x)\in D}\mathbb{P}(x,y)$，其中$\mathbb{P}(x,y)=\mathbb{P}(x)\mathbb{P}(y|x)$
- 為使目標函數最大化，可得$\mathbb{P}(y=1|x=t)=\frac{c_{t1}}{c_t}$, where c_{t1} is the number of training data with $(x,y)=(t,1)$ and c_t is the number of training data with $x=t$
- Accuracy on validation set: about 65.5%

2. 請說明你實作的discriminative model，其訓練方式和準確率為何？

- Feature的部份使用全部106項資料，實作logistic regression
- 對資料做feature scaling，將feature的各項數值縮放到$[0,1]$
- $\mathbb{P}(y=1|x)=\sigma(W\mathbf{x}+b)$, where $W\in\mathbb{R}^{2\times 106}$ and $b\in\mathbb{R}^2$
- 損失函數定為$-\sum(y\mathbb{P}(y=1|x)+(1-y)\mathbb{P}(y=0|x))$，即corss entrophy
- 以gradient descent更新參數W,b
- Batch size設為10000，跑1500個epoch
- Accuracy on validation set: about 84.3%

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。

- Feature scaling可大致分為「Rescaling」或「Standardization」，第一種是我原本的方式，第二種則是本題要求的方法。
- Rescaling $\Rightarrow x'=\frac{x-\min(x)}{\max(x)-\min(x)}$
- Standardization $\Rightarrow x'=\frac{x-\bar{x}}{\sigma}$, where σ is the standard deviation
- 實作比較的結果是兩者差不多，validation set的準確率都落在84%初

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。

- 為實作L2-norm正規化，我替損失函數加上$\lambda h(w)$這一項，使曲線圓滑一些
- 套用後，在training set上的loss明顯提高，且準確率無法提升到85%以上
- Accuracy on validation set: about 84.2%
- 我認為模型表現不好原因可能是這些：需要更多的epoch來降低loss、原本的模型複雜度本來就不高，因此正規化效果有限

5. 請討論你認為哪個attribute對結果影響最大

- 就我的觀察，「Age」可能是一項很重要的指標
- Generative model中，單考慮「Age」就能有65%以上的準確率
- 有嘗試加入「Age的平方」當作新的feature，在validation set上的表現也有些微的提升