1-1

$$\frac{\partial}{\partial \theta_i} CE(y,\hat{y}) = \frac{\partial}{\partial \hat{y}_i} CE(y,\hat{y}) \frac{\partial \hat{y}_i}{\partial \theta_i} = \frac{\partial}{\partial \hat{y}_i}(-y_i \log \hat{y}_i) \frac{\partial}{\partial \theta_i}\left(\frac{e^{\theta_i}}{\sum_{j=1}^{c} e^{\theta_j}}\right)$$

$$= -\frac{y_i}{\hat{y}_i}\left(\frac{e^{\theta_i}}{\sum_{j=1}^{c} e^{\theta_j}} - \frac{e^{\theta_i} e^{\theta_i}}{\left(\sum_{j=1}^{c} e^{\theta_j}\right)^2}\right)$$

$$= -\frac{y_i}{\hat{y}_i}\left(\hat{y}_i - \hat{y}_i^2\right) = -y_i(1-\hat{y}_i)$$

$$\therefore \nabla_\theta CE(y,\hat{y}) = -\sum_{i=1}^{c} y_i(1-\hat{y}_i)$$

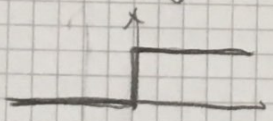
$$\frac{\partial}{\partial w_{ij}} CE(y,\hat{y}) = \frac{\partial}{\partial \theta_i} CE(y,\hat{y}) \frac{\partial \theta_i}{\partial w_{ij}} = -y_i(1-\hat{y}_i) \frac{\partial}{\partial w_{ij}} \sum_{j=0}^{d}(w_{ij} h_j + b_j)$$

$$= -y_i(1-\hat{y}_i) h_j$$

1-2

1. $P = \frac{tp}{tp+fp}$, $R = \frac{tp}{tp+fn}$. 如果只用 precision 的話 model 只需保証預測為 true 的都是對的，不用管還有多少沒找到的 true，recall 則反之。所以要用結合兩者的 f_1 score.

2. binary classification 的圖長這樣，微分之後只有在 0 那點是無限大，其他都是 0，沒辦法做 gradient descent.



3. model 在預測不同的資料集時，如果 variance 大，則結果不穩定，如果 bias 大，則預測的都偏向某個方向。類似於信度與效度的概念

4. random forest 是 bagging 的一種算法，經過多個 decision tree 投票才得到結果，比較沒有 overfit 的問題

5. 把類別型的資料轉換成多了 Binary，如類別有 (1, 2, 3)，則分別換成 (1, 0, 0) (0, 1, 0) (0, 0, 1)

6. Regularize. Dropout, Batch Normalization