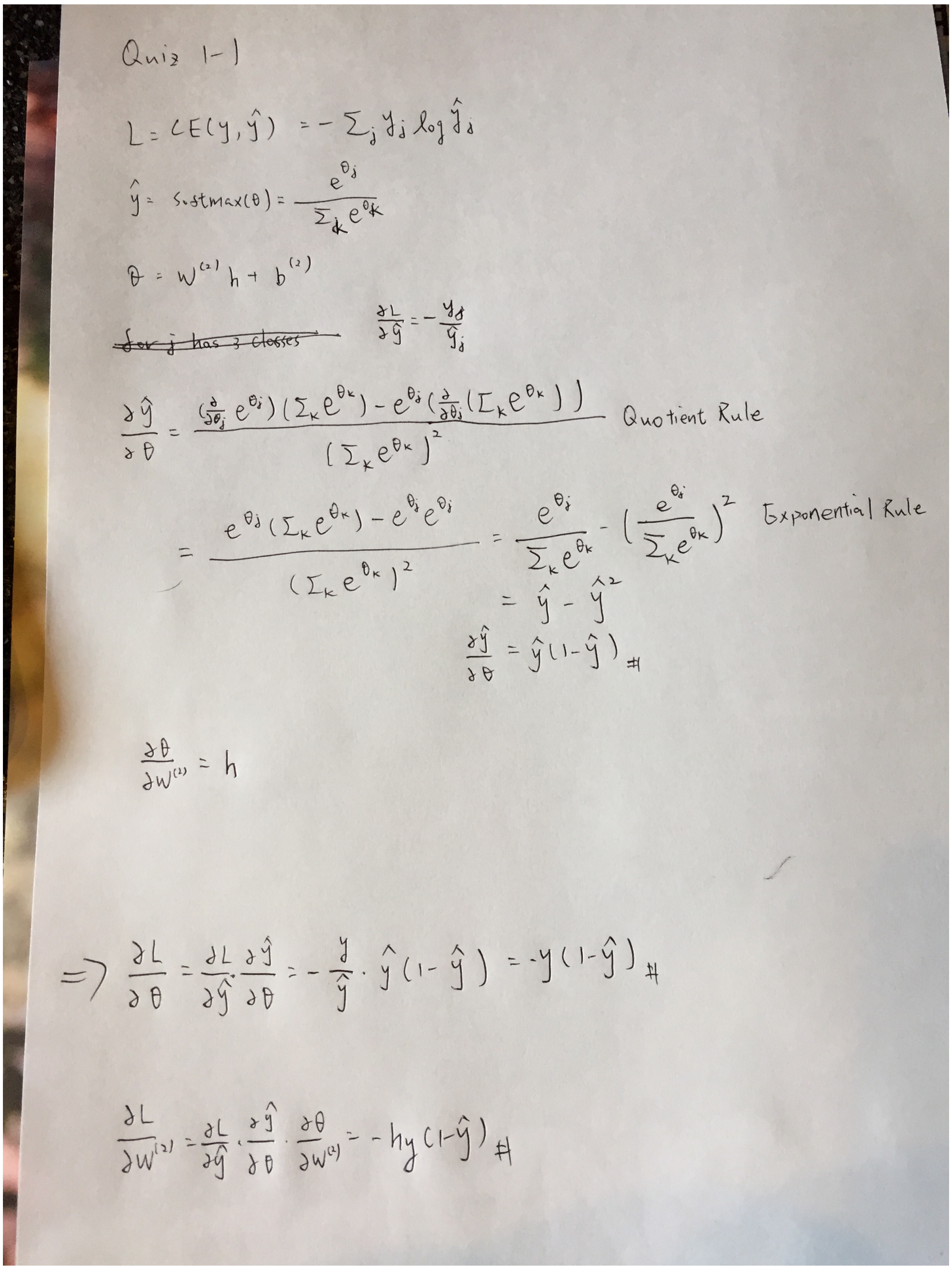
Part 1: Maths and Common Senses

Quiz 1-1:



Quiz 1-1

$$L = CE(y, \hat{y}) = -\sum_j y_j \log \hat{y}_j$$

$$\hat{y} = \text{Softmax}(\theta) = \frac{e^{\theta_j}}{\sum_k e^{\theta_k}}$$

$$\theta = W^{(2)} h + b^{(2)}$$

~~for j has 3 classes~~

$$\frac{\partial L}{\partial \hat{y}} = -\frac{y_j}{\hat{y}_j}$$

$$\frac{\partial \hat{y}}{\partial \theta} = \frac{(\frac{\partial}{\partial \theta_j} e^{\theta_j})(\sum_k e^{\theta_k}) - e^{\theta_j}(\frac{\partial}{\partial \theta_j}(\sum_k e^{\theta_k}))}{(\sum_k e^{\theta_k})^2}$$ Quotient Rule

$$= \frac{e^{\theta_j}(\sum_k e^{\theta_k}) - e^{\theta_j} e^{\theta_j}}{(\sum_k e^{\theta_k})^2} = \frac{e^{\theta_j}}{\sum_k e^{\theta_k}} - \left(\frac{e^{\theta_j}}{\sum_k e^{\theta_k}}\right)^2$$ Exponential Rule

$$= \hat{y} - \hat{y}^2$$

$$\frac{\partial \hat{y}}{\partial \theta} = \hat{y}(1-\hat{y}) \#$$

$$\frac{\partial \theta}{\partial W^{(2)}} = h$$

$$\Rightarrow \frac{\partial L}{\partial \theta} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta} = -\frac{y}{\hat{y}} \cdot \hat{y}(1-\hat{y}) = -y(1-\hat{y}) \#$$

$$\frac{\partial L}{\partial W^{(2)}} = \frac{\partial L}{\partial \hat{y}} \cdot \frac{\partial \hat{y}}{\partial \theta} \cdot \frac{\partial \theta}{\partial W^{(2)}} = -hy(1-\hat{y}) \#$$

Quiz 1-2:

- When will we use F1-score instead of precision(accuracy)?
  當 precision, recall 有落差時，舉例來說當有模型只要全部都猜對的話 precision rate 會很高，但 recall rate 會很低時，這種時候就利用 f1 score 來當權衡的指標。

- Why don’t we use binary classification function as the activation function in neural networks?
  binary function 的 output 僅有 0 與 1，這樣較難以訓練與收斂。

- What is the bias and variance of a machine learning algorithm?
  bias 是指預測值與實際值在數值上的差異；variance 指的是每次的預測值的落差，越高則每次出來的預測值差異越大

- When training a single tree in random forest, we don’t prune the tree, why?
  隨著隨機森林訓練使用 bootstrap aggregation 以及隨機選擇特徵進行分割，每棵樹之間的相關性會很低。這表示雖然單顆樹的 variance 很大，但是整群樹的輸出是合適的（會有 low bias, low variance 特性），因為樹與樹之間不相關。

- What is one-hot encoding? How to prevent overfitting in neural networks? (write down anything you know)
  將 N 種類別轉換為 N 維只有 0, 1 向量。