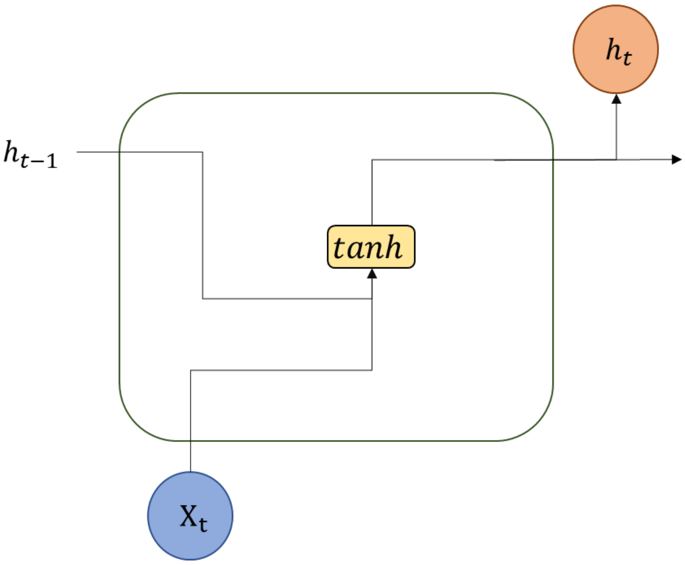
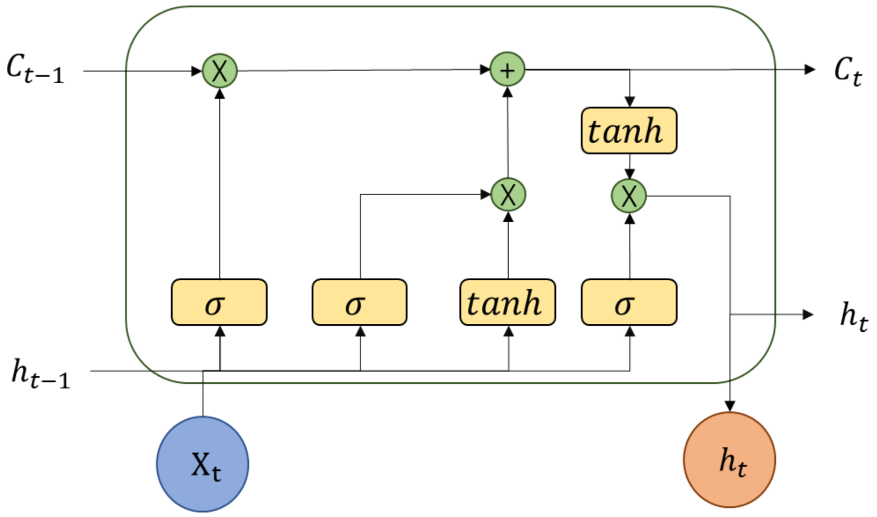
學號：R10921019　　系級：電機所碩二　　姓名：鄭皓方

1. (1.5%) 請說明RNN、GRU、LSTM等模型之間的異同。另外，如果你Kaggle上最佳的預測結果並不是使用上述三種模型產生的話，請額外說明你使用的model為何，以及簡介其背後的原理/機制。

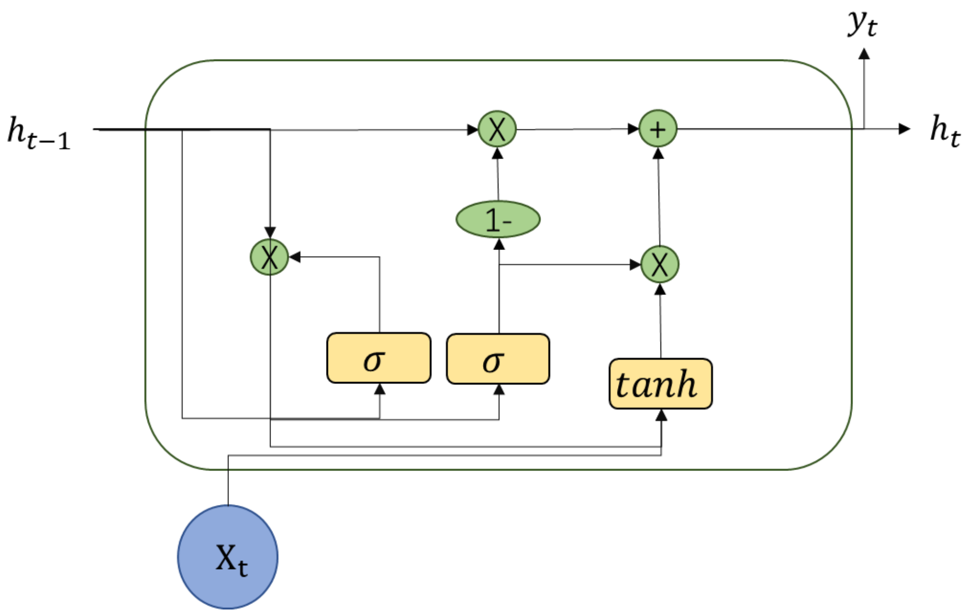
如上課所說，RNN是由一個輸入層、隱藏層、及一個輸出層。輸入的通常是一個sequence。在隱藏層中，每個input都會進入一個kernal，如下圖。這個模型會參考前面的資料和現在的輸入乘上彼此的參數後經過一個tanh函數後輸出。然後這一個kernal的狀態會被傳到下一個。因此對於最後一個單位而言，相距比較遠的資料彼此之間的影響就會比較小。



LSTM為RNN的一種特化版本。對於每一個kernal中有四個控制閘門，和RNN一樣有一個tanh閘門進行重複學習，ht保留上一個kernal的狀態。比較特別的是上面有一條儲存重要信息的輸送帶Ct，第一個閘門會決定要保留多少前面的重要信息(Ct-1)，此閘門是用sigmoid 函數構成；第二個sigmoid閘門是決定要不要把重複學習的結果加入到重要信息帶；第三個tanh的閘門則是進行重複學習並且和第二個的結果進行內積。最後一個閘們是要決定要不要保留前面的kernal狀態，並且和重複學習的結果進行內積產生當前的kernal狀態。LSTM的重要信息帶可以保存距離較遠的資料，對於較長的input sequence會比RNN有更好的效果。



GRU 則是LSTM的一種變形，如下圖這個演算法，他把第二個和第三個sugmoid的閘門整合在一起，也就是說同時調控了重要訊息帶和kernal狀態來進行狀態運測，這樣可以減少要training的參數，可以縮短需要訓練的時間。同時可以達到記憶重要信息的效果。



2. (0.5%) 請解釋為何RNN模型會發生gradient vanishing以及gradient exploding，以及這兩個現象對training可能會有什麼不良影響。

Gradient vanish: 在訓練RNN模型時，通常使用的是sigmoid函數當作activate function，也會利用backpropagation 來計算graidient。利用chain rule，可以寫出Loss function L對參數w的偏微分為:

$$\frac{\partial L}{\partial W} \propto \sum_{i=k+1}^{T} \prod_{i=k+1}^{y} \{\frac{\partial h_i}{\partial h_{i-1}}\} \frac{\partial h_k}{\partial W}$$

由上述公式可以得知，各個kernal之間的比例對於計算gradient有很大的影響。再加上微分的結果是exponential的函數，增加或減少非常劇烈，比例遞增的話就會發生exploding gradient，減少的話就會發生vanishing gradient ，如下。

1. $Vanishing\ Gradient\ \left\|\frac{\partial h_i}{\partial h_{i-1}}\right\| < 1$
2. $Exploding\ Gradient\ \left\|\frac{\partial h_i}{\partial h_{i-1}}\right\| > 1$

3. (1%) 相較於Sample Code來說，你做了哪些修改或嘗試(如模型架構、資料前處理、後處理等)？請描述你做的嘗試以及其理由，如果你認為你的做法帶來的進步與第一題的回答有關的話也請詳述之。(請注意若你的解釋太過不合理，則不論你在leader board上分數多高，這題都無法拿到滿分)

資料前處理:

1. 我把句子全部都轉成小寫、因為我不希望因為大小寫的不同讓生成的vector差很多。

2. 把標點符號分開，標點符號會跟最後一個單字被視為同一個字，因此我在前面加一隔空格，讓split的時候能被分開。

3. 以及把有@符號的字串換成he，這樣可以有效的減少單字量，從14變成9萬多。

4. 把n’t的縮寫換成not而且跟單字分開，我覺得not對語意表達很重要，如果跟單字黏在一起容易被誤認。

模型架構:

我使用LSTM 架構，因為我發現input的句子都很長，用長記憶的單元比較容易記住一些重要的單字。舉例來說，如我有人在發文的開頭就罵了一個髒話，整篇文章經過RNN後，開頭的髒話對正負面的判斷就是微乎其微。用LSTM或GRU才能有效保留重要的訊息。至於這兩個之間的選擇我是看leaderboard的分數來做判斷，最後LSTM稍稍勝出。如下表:

Model	
RNN	0.7433
LSTM	0.8038
GRU	0.7909

此外，我有把model結構加深，提高hidden dimension 的深度，讓model的能力更強。

4. (0.5%) 請簡述你leader board上表現最好的實驗結果中使用的embedding為何？如何產生？

我跟助教的方式一樣，只是我有把csv檔改成no label的，這樣可以更多的單字庫，同時也有用上說過的前處理方法減少誤認的單字，總量從14萬多變成8萬多。我有把word2vec的iteration調成10，我希望他能多迭帶幾次，讓相似的單字能被分到類似的vector，但是如果太多iter會跑太久。後面的部分沒有做太多變動，對於每個句子，就把看過的單字賦予原本指定的vector，沒看過的就給unk的vector，最後回傳tensor給後面的model做訓練

5. Play with your models!

(1) (0.5%) 在本題中，s1、s2互為彼此的valid permutation，若且唯若s1、s2兩句子的單字種類、數量相同、排列順序不同且各自皆為有意義並且合乎文法的句子。例如，A student is a banana 是 A banana is a student的valid permutation。請找出一組互為彼此的 valid permutation且使你的model產生相反的prediction的s1、s2。(s1、s2須具備合乎邏輯且有實際生活意義的語意)

我選擇的句子是一句英文俚語

Better late than never. 晚做總比不做好

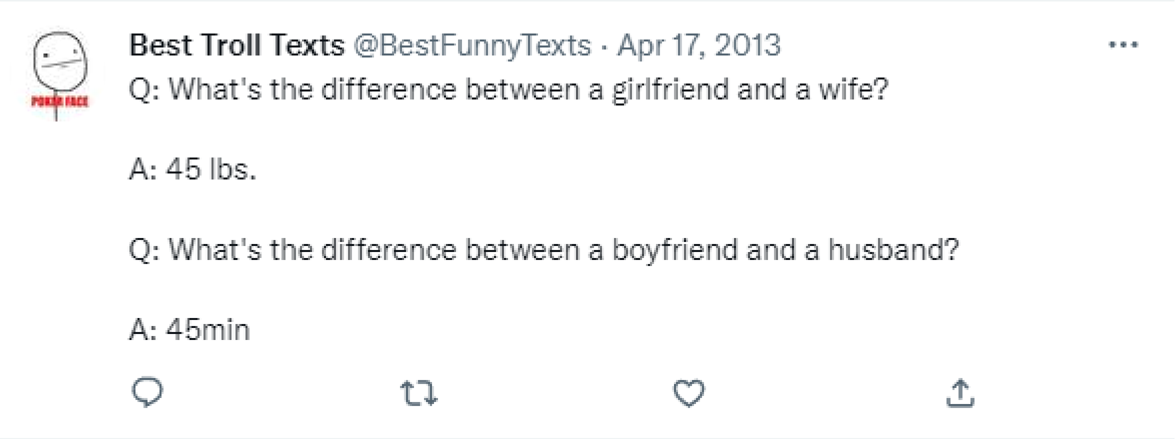
model return true，soft pridict 在0.58左右，非常接近負面，我試著調換一下順序，變成厭世的版本。

Better never than late. 既然都晚了，乾脆不要做了。

model return false，soft pridict 在0.49左右，驚險過關。可見我的model可以對於不同句子的排列方式有不同的解讀，而且看起來還算合理。

(2) (1%，bonus) 請從網路上(如FB、IG、Twitter)找出一則能夠讓model預測錯誤的「反串」酸留言，並將截圖附於report上 (即找到一則真實世界存在的留言，使得人類知道這留言應是negative，但model outputs positive)

我找到的酸留言如下，共兩個，第一個是在嘲諷已婚男女不維持自己的身材和性能力(非本人立場)，model 給出了0.88的高分，代表他覺得這是一個正面的推文，第二則是有一個大叔的抗憂鬱藥被偷了，他祝福偷他要的人可以快樂，這則推文model給出0.22的負面判斷，然而把stolen改成take的話，就會上升到0.77，我推測負面單字對整體的影響很大。



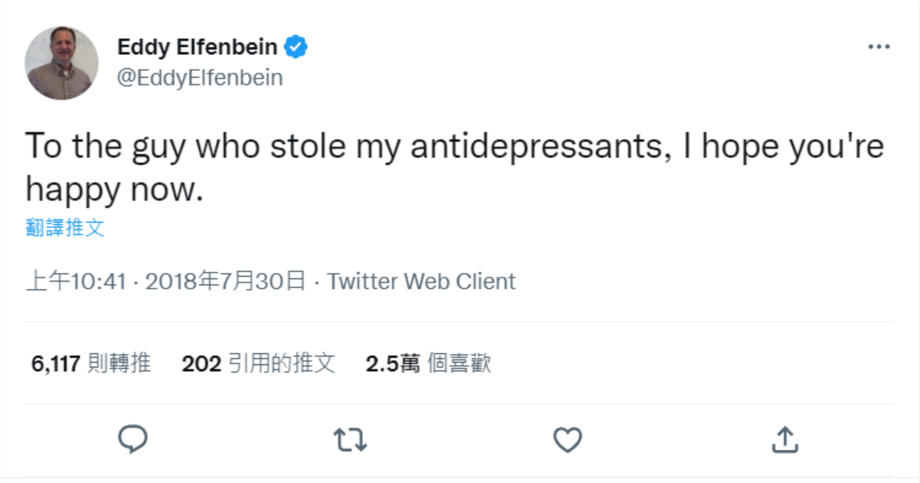
Best Troll Texts @BestFunnyTexts · Apr 17, 2013

Q: What's the difference between a girlfriend and a wife?

A: 45 lbs.

Q: What's the difference between a boyfriend and a husband?

A: 45min



Eddy Elfenbein
@EddyElfenbein

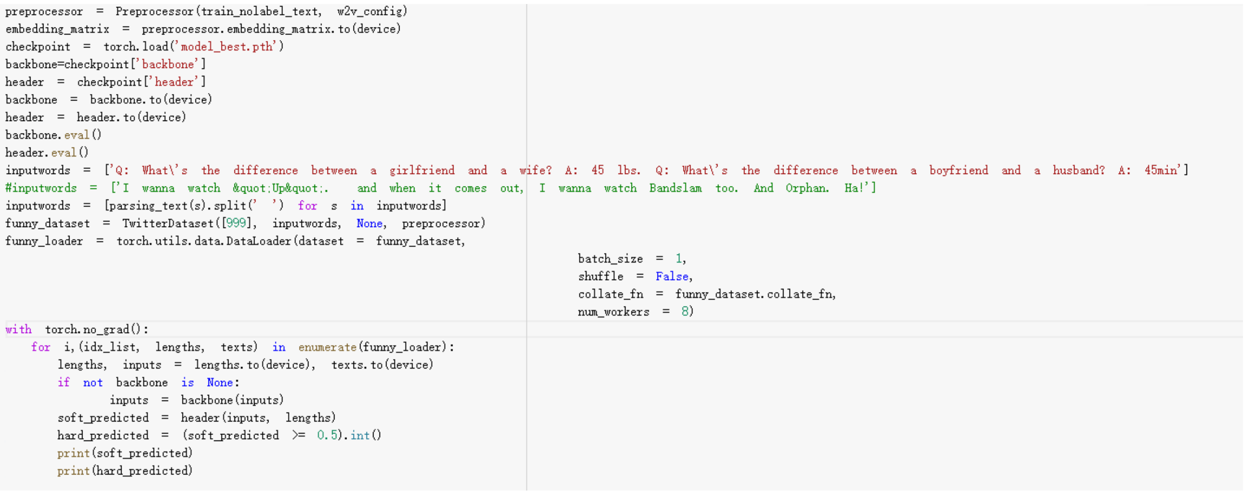
To the guy who stole my antidepressants, I hope you're happy now.

翻譯推文

上午10:41 · 2018年7月30日 · Twitter Web Client

6,117 則轉推 202 引用的推文 2.5萬 個喜歡

最後附上我這題的程式碼



```
preprocessor = Preprocessor(train_nolabel_text, w2v_config)
embedding_matrix = preprocessor.embedding_matrix.to(device)
checkpoint = torch.load('model_best.pth')
backbone=checkpoint['backbone']
header = checkpoint['header']
backbone = backbone.to(device)
header = header.to(device)
backbone.eval()
header.eval()
inputwords = ['Q: What\'s the difference between a girlfriend and a wife? A: 45 lbs. Q: What\'s the difference between a boyfriend and a husband? A: 45min']
#inputwords = ['I wanna watch &quot;Up&quot;.  and when it comes out, I wanna watch Bandslam too. And Orphan. Ha!']
inputwords = [parsing_text(s).split(' ') for s in inputwords]
funny_dataset = TwitterDataset([999], inputwords, None, preprocessor)
funny_loader = torch.utils.data.DataLoader(dataset = funny_dataset,
                                                                        batch_size = 1,
                                                                        shuffle = False,
                                                                        collate_fn = funny_dataset.collate_fn,
                                                                        num_workers = 8)
with torch.no_grad():
    for i,(idx_list, lengths, texts) in enumerate(funny_loader):
        lengths, inputs = lengths.to(device), texts.to(device)
        if not backbone is None:
                inputs = backbone(inputs)
        soft_predicted = header(inputs, lengths)
        hard_predicted = (soft_predicted >= 0.5).int()
        print(soft_predicted)
        print(hard_predicted)
```

6. (4%)Math problem:

https://hackmd.io/@1H2AB7kCSAS3NPw2FffsGg/H1ucY0pNo

1. when $t = 1$

$z = w \cdot x^1 + b = 3 + 0 = 3$

$z^i = w_i \cdot x^1 + b = 100 - 10 = 90 \qquad f(z^i) = \frac{1}{1+e^{-90}} = 1$

$z^f = w_f \cdot x^1 + b = -100 + 110 = 10 \qquad f(z^f) = \frac{1}{1+e^{-10}} = 1$

$z^o = w_o \cdot x^1 + b = 0 - 10 = -10 \qquad f(z^o) = \frac{1}{1+e^{10}} = 0$

$c_1' = f(z^i)\, g(z) + c_0 f(z^f) = 1 \cdot 3 + 0 = 3$

$y_1' = f(z^o)\, h(c') = c' f(z^o) = 0$

when $t=2$

$z = w \cdot x^2 + b = -2$

$z^i = w^i \cdot x^2 + b = 90 \qquad f(z^i) = \frac{1}{1+e^{-90}} = 1$

$z^f = w^f \cdot x^2 + b = 10 \qquad f(z^f) = \frac{1}{1+e^{-10}} = 1$

$z^o = w^o \cdot x^2 + b = 90 \qquad f(z^o) = \frac{1}{1+e^{-90}} = 1$

$c_2' = f(z^i)\, g(z) + c_1' f(z^f) = -2 + 3 \cdot 1 = 1$

$y_2' = f(z_o)\, h(c') = c_2' \cdot f(z^o) = 1$

when $t=3$

$z = 4 \quad z_i = 190 \quad z_f = -90 \quad z_o = 90$

$f(z_i) = 1 \quad f(z^f) = 0 \quad f(z_o) = 1$

$c_3' = f(z_i)\, g(z) + c_2 f(z_f) = 4$

$y_3' = f(z_o)\, h(c_3') = 4$

when $t=4$

$z=0$ $z_i=90$ $z_f=10$ $z_0=90$

$f(z_i)=1$ $f(z_f)=1$ $f(z_0)=1$

$C_4'=1\cdot 0+4\cdot 1=4$

$y_4'=1\cdot 4=4$

t	1	2	3	4	
z^i	90	90	190	90	
$f(z^i)$	1	1	1	1	
zf	10	10	-90	10	
$f(zf)$	1	1	0	1	
z^0	-10	90	90	90	
$f(z^0)$	0	1	1	1	
z	3	-2	4	0	
C'	3	1	4	4	
y	0	1	4	4	

2. Derieving the loss function of cross-entropy

$$L(y, \hat{y}) = -\frac{1}{N}\sum_{i=1}^{N} y_i \cdot \log(\hat{y_i}) + (1-y_i)\log(1-\hat{y_i})$$

$$\frac{\partial L}{\partial w_0} = \sum_{i=1}^{N}\frac{\partial L}{\partial \hat{y}_i}\frac{\partial \hat{y}_i}{\partial w_0} = -\frac{1}{N}\sum_{i=1}^{N}\left\{ \frac{y_i}{\hat{y_i}} - \frac{(1-y_i)}{(1-\hat{y_i})} \right] \frac{\partial \hat{y_i}}{\partial w_0}$$

$$y_i = \frac{1}{1+\exp(-w_0 h_2)} = \frac{1}{1+\exp(-w_0 \tanh(w_i x_2 + w_h h_1)}$$

$$= \frac{1}{1+\exp(-w_0 \tanh(w_i x_2 + W_h \tanh(w_i x_1 + w_h h_0)}$$

$$\frac{\partial y_i}{\partial w_0} = \frac{-w_0 h_2^2}{(1+\exp(-w_0 h_2))^2}$$

$$\frac{\partial L}{\partial w_0} = -\frac{1}{N}\sum_{i=1}^{N}\left\{ \frac{y_i}{\hat{y_i}} - \frac{(1-y_i)}{(1-\hat{y_i})} \right] \frac{-w_0 h_2^2}{(1+\exp(-w_0 h_2))^2}$$

$$\frac{\partial L}{\partial w_h} = -\sum_{i=1}^{N}\frac{\partial L}{\partial \hat{y_i}}\frac{\partial \hat{y_i}}{\partial w_h} = -\frac{1}{N}\sum_{i=1}^{N}\left\{ \frac{y_i}{\hat{y_i}} - \frac{(1-y_i)}{(1-\hat{y_i})} \right] \boxed{\frac{\partial \hat{y_i}}{\partial w_h}}$$

$$\Rightarrow \frac{\partial \hat{y_i}}{\partial h_2}\frac{\partial h_2}{\partial w_h} = \frac{-w_0 h_2^2}{(1+\exp(-w_0 h_2))^2}\frac{\partial h_2}{\partial w_h}$$

$$= \frac{-w_o h_2^2}{(1+\exp(-w_o h_2))^2} \operatorname{sech}^2(w_i x_2 + w_h h_1) \left[\overset{h_1}{\boxed{\tanh(w_i x_1 + w_h h_0)}} + w_h \sec^2(w_i x_1 + w_h h_0) \cdot h_0 \right]$$

where $h_0 \Rightarrow 0$

$$= \frac{-w_o h_2^2}{(1+\exp(-w_o h_2))^2} \operatorname{sech}^2(w_i x_2 + w_h h_1) \cdot h_1$$

$$\frac{\partial L}{\partial w_h} = -\frac{1}{N} \sum_{i=1}^{N} \left\{ \frac{y_i}{\hat{y}_i} - \frac{(1-y_i)}{(1-\hat{y}_i)} \right] \frac{-w_o h_2^2}{(1+\exp(-w_o h_2))^2} \operatorname{sech}^2(w_i x_2 + w_h h_1) \cdot h_1$$

#

$$\frac{\partial L}{\partial w_i} = \frac{\partial L}{\partial \hat{y}_i} \frac{\partial \hat{y}_i}{\partial h_2} \frac{\partial h_2}{\partial w_i}$$

$$\frac{\partial h_2}{\partial w_i} = \frac{\partial \{ \tanh(w_i x_2 + w_h \tanh(w_i x_1 + w_h h_0)) \}}{\partial w_i}$$

$$= \operatorname{sech}^2(w_i x_2 + w_h h_1)(x_2 + w_h \operatorname{sech}^2(w_i x_1) \cdot x_1)$$

$$\frac{\partial L}{\partial w_i} = -\frac{1}{N} \sum_{i=1}^{N} \left[\frac{y_i}{\hat{y}_i} - \frac{(1-y_i)}{(1-\hat{y}_i)} \right] \frac{-w_o h_2^2}{(1+\exp(-w_o h_2))^2}$$

$$\cdot \operatorname{sech}^2(w_i x_2 + w_h h_1)(x_2 + w_h \operatorname{sech}^2(w_i x_1) \cdot x_1)$$

#

3、

construct exponential risk function

$$\hat{R}_s(g_{T+1}) = L(g^1_{T+1}, g^2_{T+1} \cdots g^k_{T+1}) = \sum_{i=1}^{m} \exp\left(\frac{1}{K-1}\sum_{k\neq y_i} g^k_{T+1} - g^{\hat{y}_i}_{T+1}\right)$$

$$f_t = \underset{f\in F}{\operatorname{argmin}} \frac{\partial}{\partial\alpha} \hat{R}_s(g_t + \alpha f)\Big|_{\alpha=0}$$

$$= \underset{f\in F}{\operatorname{argmin}} \frac{\partial}{\partial\alpha} \sum_{i=1}^{m} \exp\left(\frac{1}{K-1}\sum g^k_t + \alpha f^k - (g^{\hat{y}_i}_t + \alpha f^{\hat{y}_i}\right)\Big|_{\alpha=0}$$

$$= \underset{f\in F}{\operatorname{argmin}} \sum_{i=1}^{m} \exp\left(\frac{1}{K-1}\sum g^k_t + \alpha f^k - g^{\hat{y}_i}_t - \alpha f^{\hat{y}_i}\right)\left(\frac{1}{K-1}\sum \alpha f^k - \alpha f^{\hat{y}_i}\right)\Big|_{\alpha=0}$$

$$= \underset{f\in F}{\operatorname{argmin}} \sum_{i=1}^{m} \exp\left(\frac{1}{K-1}\sum_{k\neq y_i} g^k_t - g^{\hat{y}_i}_t\right)\left(\frac{1}{K-1}\sum_{k\neq y_i} f^k - f^{\hat{y}_i}\right)$$

therefore $$Z_t = \sum_{i=1}^{m} \exp\left(\frac{1}{K-1}\sum_{k\neq y_i} g^k_t - g^{y_i}_t\right)$$

$$D_t(i) = \frac{\exp\left(\frac{1}{K-1}\sum_{k\neq y_i} g^k_t - g^{y_i}_t\right)}{Z(t)}$$

$$f_t = \underset{f\in F}{\operatorname{argmin}} \sum_{i=1}^{m} Z_t D_t(i) \frac{1}{K-1} \sum_{k\neq\hat{y}_i} f^k - f^{\hat{y}_i}$$

$$= \underset{f\in F}{\operatorname{argmin}}\ Z_t \underset{i\sim D_t}{\mathbb{E}} \left[\frac{1}{K-1}\sum_{k\neq\hat{y}_i} f^k - f^{\hat{y}_i}\right]$$

$$= \underset{f\in F}{\operatorname{argmin}}\ Z_t \left[\underset{i\sim D_t}{\mathbb{E}}\left(\frac{1}{K-1}\sum_{k\neq y_i} f^k\right) - \underset{i\sim D_t}{\mathbb{E}}\left(f^{\hat{y}_i}\right)\right]$$

$$= \underset{f\in F}{\operatorname{argmin}}\ Z_t \frac{1}{K-1} f^{f(x_i)}\, \mathbb{P}\{f(x_i)\neq\hat{y}_i] - f^{y_i}\left(1 - P(f(x_i)\neq y_i\right)$$

$$= \underset{f\in F}{\operatorname{argmin}} Z_t\left(\frac{1}{K-1} f^{f(x_i)} + f^{\hat{y}_i}\right) \mathbb{P}\{f(x_i)\neq\hat{y}_i] - f^{\hat{y}_i}$$

$$= \underset{f\in F}{\operatorname{argmin}}\ Z_t \underset{i\sim D_t}{\mathbb{P}}\{f(x_i)\neq\hat{y}_i]$$

$$\alpha_t = \underset{\alpha \in R}{\operatorname{argmin}} \; \hat{R}_S (g_t + \alpha f_t)$$

$$= \underset{\alpha \in R}{\operatorname{argmin}} \; \exp\Big(\frac{1}{K-1} \sum_{k \neq y_i} g_t^k - g_t^{\hat{y}_i}\Big) \exp\Big(\frac{1}{K-1} \sum_{k \neq \hat{y}_i} \alpha f_t^k - \alpha f_t^{\hat{y}_i}\Big)$$

$$= \underset{\alpha \in R}{\operatorname{argmin}} \; Z_t \mathop{\mathbb{E}}_{i \sim D_t} \Big[\exp\Big(\frac{1}{K-1} \sum_{k \neq y_i} \alpha f_t^k - \alpha f_t^{\hat{y}_i}\Big) \Big]$$

$$\epsilon_t = \mathbb{P}\big[f_t(x_i) \neq \hat{y}_i \big]$$

$$= \underset{\alpha \in R}{\operatorname{argmin}} \; Z_t \Big[\varepsilon_t e^{\frac{1}{\alpha-1} \alpha f_t^{f(x_i)}} + (1-\varepsilon_t) e^{-\alpha f_t^{f(x_i)}} \Big]$$

$$\Rightarrow \frac{\partial}{\partial \alpha} \Big[\varepsilon_t e^{\frac{1}{\alpha-1} \alpha f_t^{f(x_i)}} + (1-\varepsilon_t) e^{-\alpha f(x_i)} \Big]$$

$$= \frac{1}{K-1} f_t^{f(x_i)} \varepsilon_t e^{\frac{1}{K-1} \alpha f_t^{f(x_i)}} - f_t^{f(x_i)} (1-\varepsilon_A) e^{-\alpha f_t^{f(x_i)}}$$

$$= 0$$

$$\frac{1}{K-1} \varepsilon_t e^{\frac{1}{K-1} \alpha f_t^{f(x_i)}} = (1-\varepsilon_A) e^{-\alpha f_t^{f(x_i)}}$$

$$\frac{1-\varepsilon_t}{\varepsilon_t} (K-1) = e^{\alpha (\frac{1}{K-1} + 1) f_t^{f(x_i)}}$$

take log $\quad \alpha = \dfrac{K-1}{K} \dfrac{1}{f_t^{f(x_i)}} \log\Big(\dfrac{1-\varepsilon_t}{\varepsilon_t} (K-1) \Big)$

4、(1) $\mathbb{P}\left[z_i=k \mid x_i;\theta^{(t)}\right] = \dfrac{p(x_i, z_i=k;\ \theta^{(t)})}{\sum_{j=1}^{K} p(x_i, z_i=j;\theta^{(t)})}$

$$= \frac{P\{x_i=k;\theta^{(t)}]\, P\{x_i \mid z_i=k;\theta^{(t)}]}{\sum_{j=1}^{K} P\{x_i=i;\theta^{(t)}]\, P\{x_i \mid z_i=i;\theta^{(t)}]} = \frac{\pi_k^{(t)} N(x_i;\mu_k^{(t)},\Sigma_k^{(t)})}{\sum_{j=1}^{K}\pi_j^{(t)} N(x_i;\mu_j^{(t)},\Sigma_j^{(t)})}$$

$$= \delta_{ik}^{(t)}$$

where $N(x_i;\mu_j^{(t)},\Sigma_j^{(t)}) = \dfrac{1}{\sqrt{(2\pi)^m|\Sigma_j|}} \exp\left(-\frac{1}{2}(x_i-\mu_j)^T\Sigma_j^{-1}(x_i-\mu_j)\right)$

$$\log P(x_i, z_i=k,\theta) = \log \pi_k \frac{1}{\sqrt{(2\pi)^m|\Sigma_k|}} \exp\left(-\frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right)$$

$$= \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m|\Sigma_k|}}\right) - \frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)$$

$$Q(\theta|\theta^{(t)}) = \mathbb{E}_{z\sim P(\cdot|x;\theta^{(t)})}\left[\sum_{i=1}^{N}\log P(x_i, z_i;\theta)\right]$$

$$= \sum_{i=1}^{N}\mathbb{E}_{z_i\sim P(\cdot|x;\theta^{t})}\left\{\log P(x_i, z_i;\theta)\right]$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K}\mathbb{P}\left\{z_i=k \mid x_i;\theta^{t}\right] \log P(x_i, z_i;\theta)$$

$$= \sum_{i=1}^{N}\sum_{k=1}^{K}\delta_{ik}^{(t)}\left\{\log\left(\frac{\pi_k}{\sqrt{(2\pi)^m|\Sigma_k|}}\right) - \frac{1}{2}(x_i-\mu_k)^T\Sigma_k^{-1}(x_i-\mu_k)\right\}$$

(2) $\theta^{(t+1)} = \underset{\theta \in \Theta}{\operatorname{argmax}}\ Q(\theta | \theta^t)$

$$\nabla_{\mu_k} Q(\theta | \theta^t) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} \cdot \frac{1}{2} \nabla_{\mu_k} \boxed{\left((x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) \right)}$$

$$\Rightarrow (x_i - \mu_k - \Delta\mu_k)^T \Sigma_k^{-1} (x_i - \mu_k - \Delta\mu_k) = \{ (x_i - \mu_k)^T - \Delta\mu_k^T] \Sigma_k^{-1} \{ (x_i - \mu_k) - \Delta\mu_k]$$

$$= (x_i - \mu_k)^T \Sigma_k^{-1} (x_i - \mu_k) - \Delta\mu_k^T (2 \Sigma_k^{-1} (x_i - \mu_k)) + \Delta\mu_k^T \Sigma_k^{-1} \Delta\mu_k$$

$$\nabla_{\mu_k} Q(\theta | \theta^t) = \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} \cdot \frac{1}{2} \cdot 2 \Sigma_k^{-1} (x_i - \mu_k) = 0$$

$$\Rightarrow \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} (x_i - \mu_k^{(t+1)}) = 0 \quad \Rightarrow \sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} x_i = \sum_{k=1}^{K} \delta_{ik}^{(t)} \mu_k^{(t+1)}$$

$$\Rightarrow \mu_k^{(t+1)} = \frac{\sum_{i=1}^{N} \sum_{k=1}^{K} \delta_{ik}^{(t)} x_i}{\sum_{k=1}^{K} \delta_{ik}^{(t)}}$$

for Σ_k : First rewrite

$$Q(\theta | \theta^{(t)}) = \sum_{n=1}^{N} \sum_{k=1}^{K} \delta_{nk}^{(t)} \left\{ \log\left(\frac{\pi_k}{\sqrt{(2\pi)^m}} \right) + \frac{1}{2} \left(\log|\Sigma_k^{-1}| - \operatorname{Trace}\left(\Sigma_k^{-1} (x_n - \mu_k)(x_n - \mu_k)^T \right) \right) \right\}$$

let $\Sigma_k^{-1} = [a_{ij}^k]$, then

$$\frac{\partial}{\partial a_{ij}^k} \log Q(\theta | \theta^{(t)}) = \frac{1}{2} \sum_{n=1}^{N} \delta_{nk}^{(t)} \left\{ \overset{= e_j^T \Sigma_k e_i}{\boxed{\frac{\partial \log|\Sigma_k^{-1}|}{\partial a_{ij}^k}}} - e_j^T (x_n - \mu_k)(x_n - \mu_k)^T e_i \right\}$$

$$= \sum_{n=1}^{N} \delta_{nk}^{(t)} \frac{1}{2} \left\{ e_j^T [\Sigma_k - (x_n - \mu_k)(x_i - \mu_k)^T] e_i \right\} = 0$$

$$\Rightarrow \sum_{n=1}^{N} \delta_{nk}^{(t)} (x_i - \mu_k)(x_i - \mu_k)^T = \sum_{n=1}^{N} \delta_{mk}^{(t)} \Sigma_k^{(t+1)}$$

$$\therefore \Sigma_k^{(t+1)} = \frac{\sum_{n=1}^{N} \sum_{k=1}^{N} \delta_{nk}^{(t)} (x_i - \mu_k)(x_i - \mu_k)^T}{\sum_{n=1}^{N} \delta_{nk}^{(t)}}$$

Pareial derivative

constraint $\sum_{k=1}^{K} \pi_k = 1$

$$\nabla_{\pi_k} \left(Q(\theta|\theta^t) - \lambda \sum_{k=1}^{K} \pi_k \right) = \sum_{i=1}^{N} \frac{\delta_{ik}^{(t)}}{\pi_k} - \lambda = 0$$

$$\Rightarrow \pi_k^{(t+1)} = \sum_{i=1}^{N} \frac{\delta_{ik}^{(t)}}{\lambda}, \quad \sum_{k=1}^{K} \pi_k^{(t+1)} = \sum_{k=1}^{K} \sum_{i=1}^{N} \frac{\delta_{ik}^{(t)}}{\lambda} = 1$$

therefore, $\lambda = \sum_{k=1}^{N} \sum_{i=1}^{N} \delta_{ik}^{(t)} = N$

$$\pi_k^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \delta_{ik}^{(t)}$$