A2

Yuhan Hu

2021/2/21

1

a

784, since in the dataset we are using, each image has $28 \times 28 = 784$ pixels, intercept omitted.

b

$p(T|X, w_1, ..., w_k) = \Pi_{i=1}^{N}[\Pi_{k=1}^{K} p(C_k|x_n)^{t_{nk}}] = \Pi_{i=1}^{N}[\Pi_{k=1}^{N} K y_{nk}^{t_{nk}}] = \Pi_{k=1}^{K} y_{nk}^{\sum_N t_{nk}}$

where $y_{nk} = p(C_k|x_n) = \frac{\exp(w_k^T x_n)}{\sum_j \exp(w_j^T X_n)}$

Take negative logrithm

$-\ln E(w_1, ..., w_k) = -\ln p(T|X, w_1, ..., w_k) = \sum_{n=1}^{N} \ln \Pi_{k=1}^{K} y_{nk}^{t_{nk}} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \ln y_{nk}^{t_{nk}} = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}$

$E = -\sum_{n=1}^{N} \sum_{k=1}^{K} t_{nk} \ln y_{nk}$ is the cross-entropy function

Taking gradient by chain rule, regard $w_i x_n = a_i$

$\frac{\partial E}{\partial y_{nk}} = -\sum_{n=1}^{N} \sum_{k=1}^{K} \frac{t_{nk}}{y_{nk}}$

$\frac{\partial y_{nk}}{\partial a_j} = \frac{\partial}{\partial a_i}(\frac{\exp(a_k)}{\sum_j \exp(a_j)}) = y_{ni}(1 - y_{ni})$ if $k = i$ or $-y_{nk} y_{ni}$ if $k \neq i$

$\frac{\partial E}{\partial y} \frac{\partial y}{\partial a_i} = \sum_{n=1}^{N} \sum_{k=1}^{K} -[\frac{t_{ni}}{y_{ni}} \times y_{ni}(1 - y_{ni}) + \frac{t_{nk}}{y_{nk}} \times (-y_{nk} y_{ni})]$

the inner part can be expanded and converted to $-[t_{ni} - t_{ni} y_{ni}^{(k=i)} - t_{nk} y_{ni}^{(k \neq i)}] = y_{ni} \sum_{i=1}^{K} t_{ni} - t_{ni} = y_{nk} - t_{nk}$
where $\sum_{i=1}^{K} t_{ni} = 1$ since in $[t_{n1}...t_{nk}]$ there is exactly one $t_{ni} = 1$ and all other $t_{nk} = 0$

then $\frac{\partial E}{\partial y} \frac{\partial y}{\partial a_i} = \sum_{n=1}^{N} \sum_{i=1}^{K} (y_{ni} - t_{ni})$

$\frac{\partial a_i}{\partial w_j} = x_n$ when $j = i$, otherwise 0

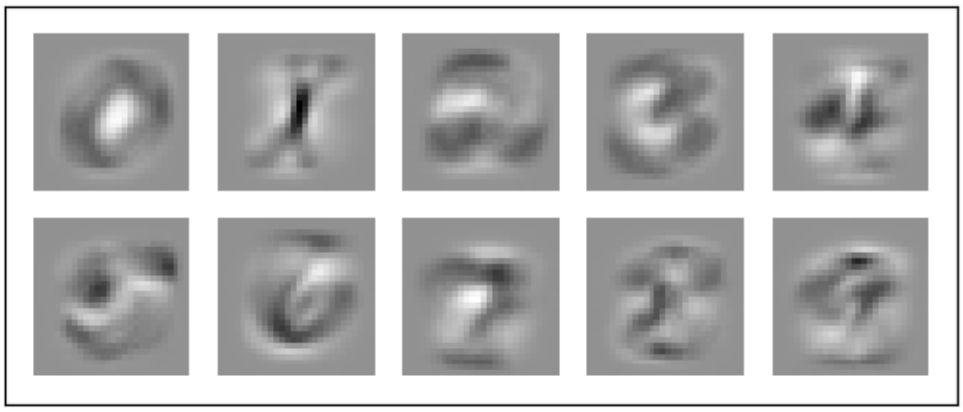
$\nabla E_{w_j}(W) = \frac{\partial E}{\partial y_{nk}} \frac{\partial y_{nk}}{\partial a_i} \frac{\partial a_i}{\partial w_j} = \sum_{n=1}^{N} \sum_{i=j}^{K} (y_{ni} - t_{ni}) x_n$ where there only exist one $i = j$

so $\nabla E_{w_k}(W) = \sum_{n=1}^{N} (y_{nk} - t_{nk}) x_n$ here trun i back to k to keep the answer consistent with handout, does not affect tge equation

c

See attached code at the end of document.

d



2

a

$p(t,X| , ,)= {n=1}^{N} K{k=1}[_k N(X_n| _n,)]^{t_k} $

$= \Pi_{n=1}^{N}[\pi_k N(X_n|\mu_n, \Sigma)]^{\sum_{k=1}^{K} t_k}$

add log at the front, we have $\ln(\sum_{n=1}^{N}[\pi_k N(X_n|\mu_n, \Sigma)]^{\sum_{k=1}^{K} t_k})$

since for every n, only one $t_k = 1$ and $\forall j \neq k, t_k = 0$

$\widehat{p(C_k)} = \widehat{\pi_k} = \frac{N_k}{N}$

$\mu_k = \frac{1}{N_k} \sum_{n \in C_k} x_{nk}$

b

See attached code.

c

Accuracy are given in part D, maximum possible number of training images were used.

d

1000 sample

For Q1, training accuracy is 0.796, test accuracy is 0.686.

For Q2, training accuracy is 0.991, test accuracy is 0.61.

10000 sample

For Q1, training accuracy is 0.8679, test accuracy is 0.834.

For Q2, training accuracy is 0.8678, test accuracy is 0.786.

30000 sample

For Q1, training accuracy is 0.885, test accuracy is 0.871.

For Q2, training accuracy is 0.8441, test accuracy is 0.799.

maximum number of sample

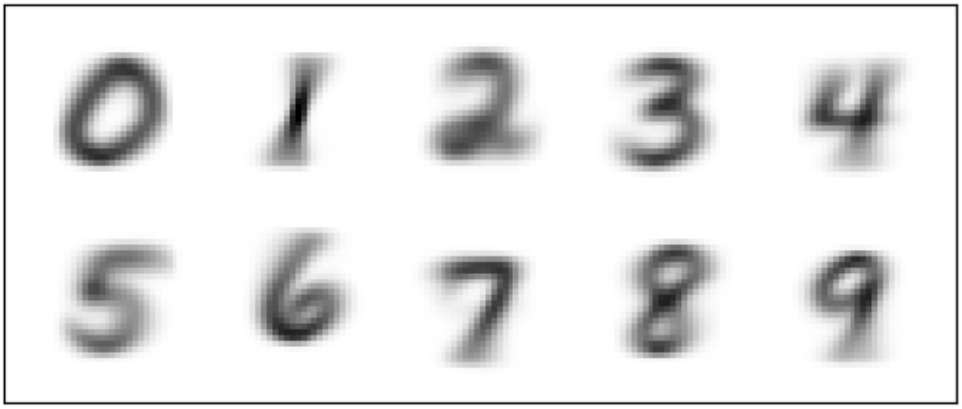
For Q1, training accuracy is 0.894, test accuracy is 0.881.

For Q2, training accuracy is 0.841, test accuracy is 0.803.

Performance of logistic regression is better than this generative model. Difference between training accuracy and test accuracy of both model are getting smaller as the training data size gets larger. Test accuracy of both models increase as sample size increase. Training error of this generative model decreases as sample size increase while training error of logistic regression increase as traing data size increases.

e

Digit 0 to 10



10 images from digit 3