ALEC MABHIZA CHIRAWU
(亚历克上)

M202161029

MACHINE LEARNING EXAM

ALEC MABHIZA CHIRAWU (亚历克上)
M202161029

Neural Network And Machine Learning Exam.

1) Why should we use random parameter initialization when using back propagation algorithm for parameter learning instead of directly making $w=0$ and $B=0$? [9 marks]

→ The random weights introduce a fixed coupling between the learning dynamics of the forwarded weights.

→ The weights of neural networks must be initialized to small random numbers.

→ The reason is that this is an expectation of the stochastic optimization algorithm used to train the model known as stochastic gradient descent.

→ The main aim is to prevent layer activation outputs from exploding or vanishing during the course of a forward pass through a deep neural network.

2) Ignoring the activation function, the forward calculation and ~~back~~ back propagation of convolution layer in convolution network are transposed. [9 marks]

→ The activation receive the calculated weighted sum of inputs.

→ Then it then adds non-linearly to the network.

→ So if activation function is ignored the architecture becomes linear and will act as a linear regression fundamentally.

~~→ The activation~~

→ The non-linear activations are used to learn the non-linearity in the data.

3) When using the self attention model as a neural layer, analyze it and convolution layer and cycle, see section 8.3. Differences in efficiency and computational complexity of modeling long-distance dependencies. [9 marks]

→ Self attention layer allows the inputs to ~~itself~~ interact with each other and find out who they ~~are~~ should pay more attention to.
→ Therefor the outputs are aggregates of these interactions and attention scores.

a)

Convolutional Network	Bidirectional Recurrent Network
→ Series of hidden layers. → this reduce its time complexity	→ Computes information in strict order → In self attention bidirectional recurrent network the inputs are processed in both forward and backward time order. → this inceases its time efficiency to O(N) and computational complexity.

4) It is proved that for the data set composed of n samples (sample dimension $d > n$), the effective shadow space of principal component analysis does not exceed n-1 dimension. [9 marks]

→ $N-1=1$

→ two points always lie on a line and a line is 1 dimensional.

→ So the exact dimensionality of the space does not matter as long $N > 1$.

→ Points only occupy 1-dimensional ~~sup~~ subspace and the variants is only spread in this subspace.

5) Can ensemble learning avoid over-fitting? [9 marks]

→ Ensemble methods are used to combine based estimators.

→ There are two types of ensemble methods which are averaging ~~and~~ and boosting.

→ Ensemble reduces the risk of overfitting and also increase ~~the~~ the performance.

→ It provide a well generalized model.

6) Calculation questions (4 questions in total, full score 32 points)

1) Suppose there are N samples $x(1), x(2), \cdots, x(N)$ obey normal distribution $N(\mu, \sigma^2)$, where unknown. 1) use maximum likelihood estimation to solve the optimal parameter μ^{ml}; 2) If the parameter μ is a random variable and obeys the normal distribution $N(\mu_0, \sigma_0^2)$, the maximum a posteriori estimation is used to solve the optimal parameter μ^{map}. [8 marks]

1.) $\Rightarrow$ likelihood : $L(v) - \sum_{b=1}^{k} \sum_{h} e^{-\frac{E(v,h)}{Z}}$

$$\Rightarrow L(\mu^{ml}, 2) = \sum_{i=1}^{N} f_i(y_i)$$

$$\Rightarrow \sum_{i=1}^{n} (\mu^{ml})_i \, (x(N_y))$$

2.) $\ln L(y, B) = \ln \sum_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\pi_i)$

for $\div N(\mu_0, \sigma_0^2)$

$$\ln L(\mu_0, \sigma_0^2) = \ln \sum_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \left[\mu_0 \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\sigma_0^2)$$

$$\Rightarrow \frac{\sum_{i=1}^{n} \left(\frac{\mu_0 \pi_i}{1-\pi_i}\right)}{\sum_{i=1}^{n} \left(\ln(1-\sigma_0^2)\right)}$$

$$\Rightarrow 1 - \frac{\left(\frac{\mu_0 \pi_i}{1-\pi_i}\right)}{(1-\sigma_0^2)}$$

6) Calculation questions (4 questions in total, full score 32 points)

1) Suppose there are N samples $x(1), x(2), \cdots, X(N)$ obey normal distribution $N(\mu, \sigma 2)$, where unknown. 1) use maximum likelihood estimation to solve the optimal parameter μ^{ml}; 2) If the parameter μ is a random variable and obeys the normal distribution $N(\mu_0, \sigma_0^2)$, the maximum a posteriori estimation is used to solve the optimal parameter μ^{map}. [8 marks]

1.) $\Rightarrow$ likelihood: $L(V) - \sum_{b=1}^{k} \sum_h e^{-\frac{E(v,h)}{Z}}$

$$\Rightarrow L(\mu^{ml}, 2) = \sum_{i=1}^{N} f_i(y_i)$$

$$\Rightarrow \sum_{i=1}^{n} (\mu^{ml})_i \, (x(N_y)$$

2.) $\ln L(y, B) = \ln \sum_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\pi_i)$

for $\div N(\mu_0, \sigma_0^2)$

$$\ln L(\mu_0, \sigma_0^2) = \ln \sum_{i=1}^{n} f_i(y_i) = \sum_{i=1}^{n} \left[\mu_0 \ln\left(\frac{\pi_i}{1-\pi_i}\right)\right] + \sum_{i=1}^{n} \ln(1-\sigma_0^2)$$

$$\Rightarrow \frac{\sum_{i=1}^{n} \left(\frac{\mu_0 \pi_i}{1-\pi_i}\right)}{\sum_{i=1}^{n} \left(\ln(1-\sigma_0^2)\right)}$$

$$\Rightarrow \frac{1 - \left(\frac{\mu_0 \pi_i}{1-\pi_i}\right)}{(1-\sigma_0^2)}$$

5) Can ensemble learning avoid over-fitting? [9 marks]

→ Ensemble methods are used to combine based estimators.

→ There are two types of ensemble methods which are averaging ~~and~~ and boosting.

→ Ensemble reduces the risk of overfitting and also increase ~~the~~ the performance.

→ It provide a well generalized model.

6) 2) Consider a Bernouli mixed distribution, given a set of training sets $d = \{x(1), X(2), \ldots, X(n)\}$ if the EM algorithm is used for parameter estimation, the parameter update formula of each step is derived. [8marks]

$$d = \{x(1), x(2) \cdots, X(n)\}$$

$$\Rightarrow \sum_{i=1}^{N} r^{(i)} \log(N(x(i) \mid \mu, x(i)) + r^{(i)} \log(x(n))$$

$$\Rightarrow \sum_{i}^{N} (x^{i} - \mu) \left(\frac{r^{(i)}}{x(i)} + \frac{(1 - r^{(i)})}{x(n)} \right)$$

6. 3) In the restricted Boltzmann machine, if the observable varible
[8 marks] serves the Bernoulli distribution and the condition P of the observable variable vi
$p(vi = k|h) = exp(a(k)i + \sum jw(k)ijhj) / (\sum K\ k'=1\ exp(a(k')i + \sum jw(k')ijhj))$, where $k \in [1, K]$ is the value of the observable variable, $W(k)$ and $a(k)$ are parameters, please give the energy function satisfying this conditional distribution.

$$\Rightarrow p(v_i = k|h) = \frac{exp(a(k)_i + \sum_j w(k)_{ij} h_j)}{\sum_{k'=1}^{K} exp(a(k')_i + \sum_j w(k')_{ij} h_j))} \cancel{/(\sum K}$$

$\Rightarrow$ W(k) $\rightarrow$ ~~of~~ observable variable
a(k) $\rightarrow$ parameters

$t = a(k)_i$
$m = \sum_j w(k)_{ij} h_j$
$L = a(k')_i$
$n = \sum_j w(k') ij h_j)$

$$\Rightarrow (v_i = k|h) = \frac{exp(t + m)}{\sum_{k'=1}^{K} exp(L + n)}$$

$$\Rightarrow (V_i = k|h)\left(\sum_{k'=1}^{K} exp(L+n)\right) = exp(t+m)$$

$$\Rightarrow Exp(k,h)\ v_i\ (L+n) = (t+m)$$

$$\Rightarrow Exp(k,h) \Rightarrow v_i + (L+n) = (t+m)$$

$$\Rightarrow v_i - t - m + L + n$$

$$\Rightarrow v_i - a(k)_i - \sum_j w(k)_{ij} h_j + a(k)_i + \sum_j w(k)_{ij} h_j$$

$$\Rightarrow v_i - \cancel{a_i} - \sum_j^3 w_{ij} \cancel{+ a_i} + \sum_j^3 w h_i$$

$$\Rightarrow v_i - \sum_j^3 w_{ij}$$

6 [8 marks] 4) For a discrete random z ~~$p_\theta(z)$~~ with distribution $P_\theta(z)$ and function $f(z)$, how to calculate the expected $L(\theta) = E_{z \sim p_\theta(z)}$ $[f(z)]$ about the distribution parameters θ derivate of

$$\theta^2 = \sum (x-\mu)^2 P(x)$$

$P_\theta(z)$ — random distribution

$$L(\theta) = E_{z} \sim p_\theta(z) [f(z)]$$

$$L(\theta) = P\theta(z), f(z)$$

$$L(\theta^2) = P\theta(z) \, (f(z))$$

$$L(\theta) = \sqrt{P\theta(z) \, f(z)}$$

$$L(\theta) \Rightarrow \sqrt{P_\theta(z)} \, \sqrt{f(z)}$$

using conducting rule:-

$$\underline{L(\theta) = E_{z \sim p} \, \theta(z) \, (f(z))}$$ Proved !!!

7) Proof questions (3 questions in total, full score 23 points)

1) Given a multi classification data set, it is proved that:

1) If the samples of each class in the data set are linearly separable from the samples other than the class, the data set must be linearly separable;

2) If the samples of every two classes in the dataset are linearly separable, the dataset is not ~~necessarily~~ necessarily linearly separable. [8 marks]

1) In this case we are able to draw a line in between the classes.

→ so the data is linearly separable.

→ We can find clusters with cluster purity of 100% using some clustering method like k-means.

2) False: For the sample of two classes all other training are irrelevant, so they can be deleted without changing the position and orientation of the hyperplane.

→ The hyperplane ~~sp~~ separating the two classes might be written as $x = w_0 + w_1 a_1 + w_2 a_2$.

$w_i \Rightarrow$ weights to be learned.

$a_1, a_2 \Rightarrow$ attribute values.

7) 2) The gradient of parameters in LSTM network is deduced and its effect of avoiding gradient disappearance is analyzed [8 marks]

→ LSTM's solve the problem using a unique additive gradient structure that includes direct access to the forget gate's activations, enabling the network to encourage desired behavior from the error gradient using frequent gates update on every time steep of learning process.

7) 3) It is proved that the weight attenuation regularization and 2 regularization have the same effect in the standard random gradient descent, and whether this conclusion is still valid in the momentum method and Adam algorithm is analyzed;

[7 marks]

→ This conclusion is still valid since AdamW is also an aspect on weight decay which is done by L2 regularization.

→ Adam Is a replacement optimization algorithm for stochastic gradient descent for training deep learning models.

→ Adam makes use of AdamW which works the same as L2 regularization.

→ L2 regularization acts like a force that removes a small percentage of weights at each iteration.