1.

1.

$$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}$$

$$\boxed{P_n = \theta(-y_n(Az_n+B)) = \frac{1}{1+e^{y_n(Az_n+B)}}}$$

原式 $F(A,B) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1+e^{\underbrace{-y_n(Az_n+B)}_{s}}\right)$

$$= -\frac{1}{N}\sum_{n=1}^{N} \ln\left(\frac{1}{1+e^{-s}}\right) = -\frac{1}{N}\sum_{n=1}^{N} \ln\left(\frac{e^s}{1+e^s}\right) = -\frac{1}{N}\sum_{n=1}^{N} \ln(1-P_n)$$

$$\frac{\partial P_n}{\partial A} = -P_n(1-P_n)(y_n z_n) \qquad \frac{\partial P_n}{\partial B} = -P_n(1-P_n)(y_n)$$

$$\frac{\partial F}{\partial A} = -\frac{1}{N}\sum_{n=1}^{N} \frac{1}{1-P_n}(-1)\frac{\partial P_n}{\partial A} \qquad \frac{\partial F}{\partial B} = -\frac{1}{N}\sum_{n=1}^{N} \frac{1}{1-P_n}(-1)\frac{\partial P_n}{\partial B}$$

$$= -\frac{1}{N}\sum_{n=1}^{N} \frac{1}{\cancel{1-P_n}}(-1)(-P_n)\cancel{(1-P_n)}(y_n z_n) \qquad = -\frac{1}{N}\sum_{n=1}^{N} \frac{1}{\cancel{1-P_n}}(-1)(-P_n)\cancel{(1-P_n)}(y_n)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} P_n y_n z_n \qquad = -\frac{1}{N}\sum_{n=1}^{N} P_n y_n$$

$$\nabla F(A,B) = \left[-\frac{1}{N}\sum_{n=1}^{N} P_n y_n z_n,\quad -\frac{1}{N}\sum_{n=1}^{N} P_n y_n\right] \#$$

2.

$$H(F) = \begin{bmatrix} \dfrac{\partial^2 F(A,B)}{\partial A^2} & \dfrac{\partial^2 F(A,B)}{\partial A \partial B} \\ \dfrac{\partial^2 F(A,B)}{\partial A \partial B} & \dfrac{\partial^2 F(A,B)}{\partial B^2} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} z_n^2 P_n(1-P_n) & \frac{1}{N}\sum_{n=1}^{N} z_n P_n(1-P_n) \\ \frac{1}{N}\sum_{n=1}^{N} z_n P_n(1-P_n) & \frac{1}{N}\sum_{n=1}^{N} P_n(1-P_n) \end{bmatrix}$$

$$\frac{\partial^2 F(A,B)}{\partial A^2} = \frac{\partial}{\partial A}\left(-\frac{1}{N}\sum_{n=1}^{N} P_n y_n z_n\right)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} y_n z_n \frac{\partial P_n}{\partial A} = -\frac{1}{N}\sum_{n=1}^{N} y_n z_n(-P_n)(1-P_n)(y_n z_n)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} \overset{=1}{y_n^2} z_n^2 (-P_n)(1-P_n) = \frac{1}{N}\sum_{n=1}^{N} z_n^2 P_n(1-P_n)$$

$$\frac{\partial^2 F(A,B)}{\partial B^2} = \frac{\partial}{\partial B}\left(-\frac{1}{N}\sum_{n=1}^{N} P_n y_n\right)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} y_n \frac{\partial P_n}{\partial B} = -\frac{1}{N}\sum_{n=1}^{N} y_n(-P_n)(1-P_n)y_n = \frac{1}{N}\sum_{n=1}^{N} P_n(1-P_n)$$

$$\frac{\partial^2 F(A,B)}{\partial A \partial B} = \frac{\partial}{\partial B}\left(-\frac{1}{N}\sum_{n=1}^{N} P_n y_n z_n\right)$$

$$= -\frac{1}{N}\sum_{n=1}^{N} y_n z_n \frac{\partial P_n}{\partial B} = -\frac{1}{N}\sum_{n=1}^{N} y_n z_n(-P_n)(1-P_n)y_n = \frac{1}{N}\sum_{n=1}^{N} z_n (P_n)(1-P_n)$$

3.

3.

We can prove that the matrix $H(F)$ is positive semi-definite by finding all of its eigenvalue ≥ 0.

Let the eigenvalue $= \lambda$

$$H(F)B = \lambda B = (\lambda I)B$$

$$(H(F) - \lambda I)B = 0$$

$$H(F) - \lambda I = \begin{bmatrix} \frac{1}{N}\sum_{n=1}^{N} z_n^2 P_n(1-P_n) - \lambda & \frac{1}{N}\sum_{n=1}^{N} z_n P_n(1-P_n) \\ \frac{1}{N}\sum_{n=1}^{N} z_n P_n(1-P_n) & \frac{1}{N}\sum_{n=1}^{N} P_n(1-P_n) - \lambda \end{bmatrix}$$

Let $\frac{1}{N}\sum_{n=1}^{N} P_n(1-P_n) = \alpha$

$$\det(H(F) - \lambda I)$$

$$= (z_n^2\alpha - \lambda)(\alpha - \lambda) - (z_n\alpha)(z_n\alpha)$$

$$= \cancel{z_n^2\alpha^2} - \cancel{z_n^2\alpha^2} + \lambda^2 - \lambda(\alpha + z_n^2\alpha)$$

$$= \lambda(\lambda - \alpha - z_n^2\alpha) = 0$$

$$\lambda = 0 \ (\geq 0) \Rightarrow \text{ok}.$$

$$\lambda = \alpha + z_n^2\alpha = \alpha(1 + z_n^2) = \frac{1}{N}\sum_{n=1}^{N} \underbrace{P_n}_{\text{between 0 and 1}} \overbrace{(1-P_n)}^{\text{between 0 and 1}} \underbrace{(1+z_n^2)}_{\text{positive}} (\geq 0) \Rightarrow \text{ok}$$

With the eigenvalues 0 and $\alpha + z_n^2\alpha \geq 0$, we can proof that $H(F)$ is positive semi-definite. ※

4.

4、假設 $\text{sign}(0) = +1$

$$(W_0, W_1, W_2, \ldots\ldots, W_d) = (d-1, +1, +1, \ldots\ldots, +1) ※$$

$$gA(x) = \text{sign}\left(\sum_{i=1}^{d} x_i + d - 1\right) ※$$

Only when all the x_n is -1, $gA(x) = \text{sign}(-d+1-1) = -1$, which means that if any $x_n = +1$, $gA(x)$ will be 1.

5.

5.

$$\frac{\partial e_n}{\partial w_{ij}^{(l)}} = \delta_j^{(l)} \left(x_i^{(l-1)} \right)$$

$$\delta_j^{(l)} = \frac{\partial e_n}{\partial s_j^{(l)}} = \sum_k \left(\delta_k^{(l+1)}\right)\left(w_{jk}^{(l+1)}\right)\left(\tanh'\left(s_j^{(l)}\right)\right)$$

$x_i^{(l)} = 0$ for $l \geq 1$

$\delta_j^{(l)} = 0$ for $l < L$, $\delta_1^{(L)} \neq 0$

if $y_n \neq 0$, $x_i^{(L-1)} = 0$, $i \neq 0$ and $x_0^{(L-1)} = 1$

All the gradient components must be zero except for the gradient components with respect to $W_{01}^{(L)}$ #

6.

6.

$$\underbrace{d^{(0)}}_{\text{input}} + \underbrace{1}_{\text{constant}} = 12 \qquad \sum_{n=1}^{N} x_n = 48$$

max: $12 \cdot x_1 + x_1 \cdot x_2 + \dots + x_n \cdot 1$

while $N=1$, $12 \times (48-1) + 48 \times 1 = 612$

while $N=2$, $12 \times (x_1 - 1) + (x_1)(48 - x_1 - 1) + (48 - x_1)$

$= 12x_1 - 12 + 47x_1 - x_1^2 + 48 - x_1$

[redacted]

$= -x_1^2 + 58x_1 + 36 = -(x_1 - 29)^2 + 877$

NNet $= 11 - 28 - 18 - 1$

$(11+1) \quad (28+1) \quad (18+1)$

max: $12 \times 28 + 29 \times 18 + 19 \times 1 = 336 + 522 + 19 = 877$ #

7.

7.

$$err_n(w) = \|\underset{d\times1}{x_n} - \underset{d\times1}{w}\,\underset{1\times d}{w^T}\,\underset{d\times1}{x_n}\|^2$$

$$= \underset{1\times d}{(x_n - ww^Tx_n)^T}\ \underset{d\times1}{(x_n - ww^Tx_n)}$$

$$= x_n^Tx_n - 2(x_n^Tw)(w^Tx_n) + x_n^Tww^Tww^Tx_n$$

$$= x_n^Tx_n \underbrace{- 2(w^Tx_n)^2}_{(-2w^Tx_n)(w^Tx_n)} + (w^Tx_n)^2(w^Tw)$$

$$\nabla_w err(w) = -2x_n(w^Tx_n) - 2w^Tx_n(x_n) + \frac{\partial}{\partial w}(w^Tx_n)^2(w^Tw)$$

$$= -4w^Tx_nx_n + (w^Tx_n)^2\cdot 2w + (w^Tw)\cdot\frac{\partial}{\partial w}(w^Tx_n)^2$$

$$= -4w^Tx_nx_n + 2(w^Tx_n)^2w + (w^Tw)(w^Tx_nx_n + w^Tx_nx_n)$$

$$= \underbrace{-4w^Tx_nx_n}_{d\times1} + \underbrace{2(w^Tx_n)^2w}_{d\times1} + \underbrace{2(w^Tw)(w^Tx_n)x_n}_{d\times1}$$ #

8.

8.

$$E_{in}(w) = \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^T(x_n+\epsilon_n)\|^2$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(x_n^Tx_n - 2x_n^Tww^T(x_n+\epsilon_n) + (x_n+\epsilon_n)^Tww^Tww^T(x_n+\epsilon_n)\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(x_n^Tx_n - 2x_n^Tww^Tx_n - 2x_n^Tww^T\epsilon_n + x_n^Tww^Tww^Tx_n + \epsilon_n^Tww^Tww^T\epsilon_n + 2\epsilon_n^Tww^Tww^Tx_n\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\left(x_n^Tx_n - 2x_n^Tww^Tx_n + x_n^Tww^Tww^Tx_n\right) + \frac{1}{N}\sum_{n=1}^{N}\left(-2x_n^Tww^T\epsilon_n + \epsilon_n^Tww^Tww^T\epsilon_n + 2\epsilon_n^Tww^Tww^Tx_n\right)$$

$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \frac{1}{N}\sum_{n=1}^{N}\left(-2x_n^Tww^T\epsilon_n + \epsilon_n^Tww^Tww^T\epsilon_n + 2\epsilon_n^Tww^Tww^Tx_n\right)$$

We know that $E[\epsilon_n\epsilon_n^T] = I_n$

and the ϵ_n's first order's Expectation $= 0$

So, $$E[E_{in}(w)] = \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \frac{1}{N}\sum_{n=1}^{N}E[\epsilon_n^Tww^Tww^T\epsilon_n]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \frac{1}{N}\sum_{n=1}^{N}E[w^Tw\,\epsilon_n^T\epsilon_n\,w^Tw]$$

$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + \frac{1}{N}\sum_{n=1}^{N}w^Tw\,\underbrace{E[\epsilon_n^T\epsilon_n]}_{=I_n}\,w^Tw$$

$$= \frac{1}{N}\sum_{n=1}^{N}\|x_n - ww^Tx_n\|^2 + (w^Tw)^2$$

$$\Omega(w) = (w^Tw)^2$$ #

9.

9.

error function $\sum_{i=1}^{d} (g_i(x) - x_i)^2$

$$= \sum_{i=1}^{d} (x_i w_{ij}^{(1)} w_{ji}^{(2)} - x_i)^2$$

$$= \sum_{i=1}^{d} (x_i u_{ij} u_{ij}^T - x_i)^2$$

10.

10.

$$\frac{\partial E_{in}(u)}{\partial u_{ij}}$$

$$= \sum_{i=1}^{d} \frac{\partial}{\partial u_{ij}} (x_i u_{ij} u_{ij}^T - x_i)^2$$

$$= \sum_{i=1}^{d} 2 (x_i u_{ij} u_{ij}^T - x_i)(2 x_i u_{ij})$$

$$= \sum_{i=1}^{d} 2 (x_i u_{ij} u_{ij}^T - x_i) x_i (2 u_{ij})$$

$$\frac{\partial E_{in}(w)}{\partial w_{ij}^{(1)}} + \frac{\partial E_{in}(w)}{\partial w_{ji}^{(2)}}$$

$$= \sum_{i=1}^{d} 2 (x_i w_{ij}^{(1)} w_{ji}^{(2)} - x_i)(x_i w_{ji}^{(2)})$$

$$+ \sum_{i=1}^{d} 2 (x_i w_{ij}^{(1)} w_{ji}^{(2)} - x_i)(x_i w_{ij}^{(1)})$$

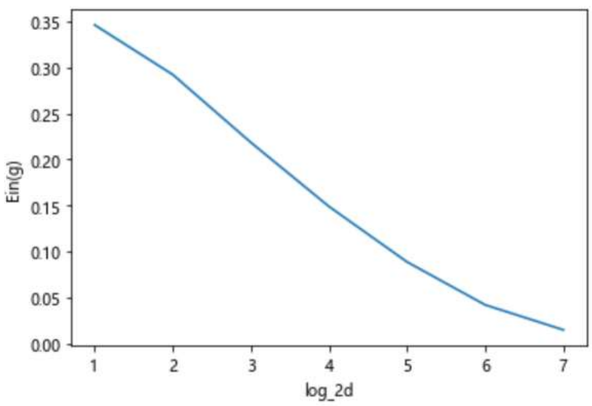
$$= \sum_{i=1}^{d} 2 (x_i w_{ij}^{(1)} w_{ji}^{(2)} - x_i) x_i (w_{ij}^{(1)} + w_{ji}^{(2)})$$

We know that $u_{ij} = w_{ij}^{(1)}$, and $u_{ij} = w_{ji}^{(2)}$,

So we can prove that $\frac{\partial E_{in}(u)}{\partial u_{ij}} = \frac{\partial E_{in}(w)}{\partial w_{ij}^{(1)}} + \frac{\partial E_{in}(w)}{\partial w_{ji}^{(2)}}$

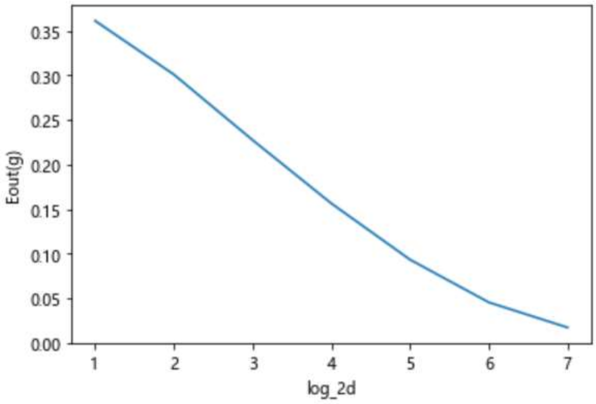
11.

Out[93]: Text(0, 0.5, 'Ein(g)')



As d goes bigger, the size of hidden layer gets bigger, the Ein(g) gets smaller.

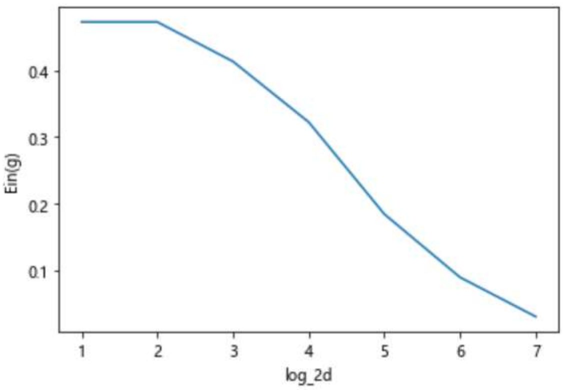
12.



Same with 11, as d goes bigger, the size of hidden layer gets bigger, the Ein(g) gets smaller.
However, Eout's error rate is bigger than Ein in every d, because the model is fit by training data but not testing data.

13.

Out[16]: Text(0, 0.5, 'Ein(g)')

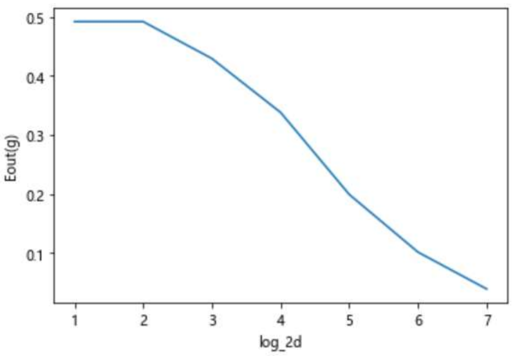


13.

As d goes bigger, the size of hidden layer goes bigger, Ein(g) gets smaller, but still bigger than autoencoder in 11.

14.

Out[17]: Text(0, 0.5, 'Eout(g)')



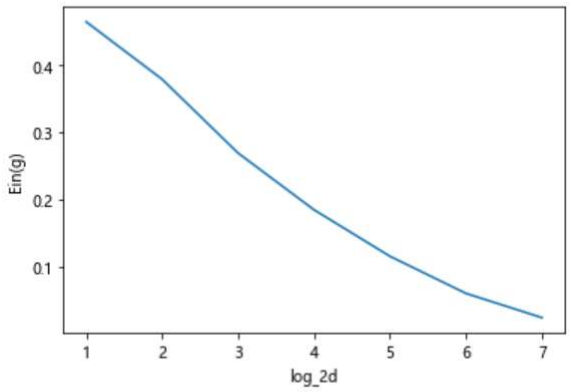
14.

Same with 13, as d goes bigger, the size of hidden layer goes bigger, Ein(g) gets smaller, but still bigger than autoencoder in 12.
Also, the error rate that the test data performs are worse than 13.

15.

```
[0.46454766, 0.3790338, 0.26911533, 0.18459733, 0.11539804, 0.060169443, 0.02405729]
```

Out[12]: `Text(0, 0.5, 'Ein(g)')`

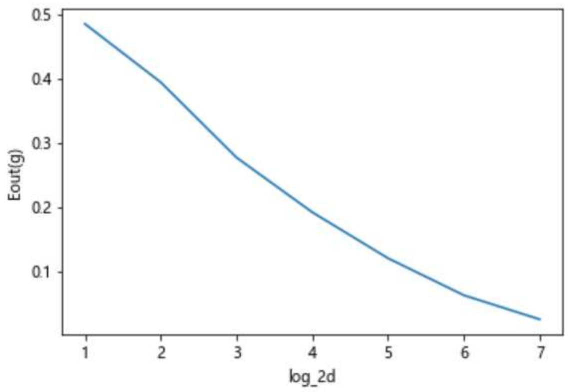


15.
As d goes bigger, the size of hidden layer goes bigger, Ein(g) gets smaller.
Every Ein(g) in PCA perform better than Ein(g) in 13. in each different d.

16.

```
[0.4847186, 0.3939795, 0.2769158, 0.19177727, 0.12019751, 0.06265429, 0.025237955]
```

Out[14]: `Text(0, 0.5, 'Eout(g)')`



15.
As d goes bigger, the size of hidden layer goes bigger, Ein(g) gets smaller.
Compares to Ein, Eout is bigger(error is larger).
Every Ein(g) in PCA perform better than Ein(g) in 14. in each different d.

17.

17.

We want to prove $N^\Delta < 2^N$

First, we change it to $\Delta \log N < N \log 2$ by using log.

So, we know $f(N) = \Delta \ln N - N \ln 2 < 0$.

$$f'(N) = \frac{\Delta}{N} - \ln 2$$

while $\frac{\Delta}{N} - \ln 2 > 0$, $f'(x) > 0$

$$\Rightarrow \frac{\Delta}{N} > \ln 2$$

$$\Rightarrow N < \frac{\Delta}{\ln 2}, \quad f'(x) > 0$$

while $\frac{\Delta}{N} - \ln 2 < 0$, $f'(x) < 0$

$$\Rightarrow N > \frac{\Delta}{\ln 2}, \quad f'(x) < 0$$

$$N = 3\Delta \log_2 \Delta = \frac{3\Delta \ln \Delta}{\ln 2} > \frac{\Delta}{\ln 2} \text{ while } \Delta \ge 2,$$

so we can know that

$$\underbrace{f(N) \le f(3\Delta \log_2 \Delta)}_{f'(x) < 0} < 0$$

Finally, we can know that while $N \ge 3\Delta \log_2 \Delta$,

$$N^\Delta < 2^N$$

and while Δ, N are integers, we can prove that

$$N^\Delta + 1 < 2^N \quad \#$$

18.

18.

First, we know that

$$B(N,k) \le \sum_{i=0}^{k-1} C_i^N \quad \text{and} \quad \sum_{i=0}^{D} \binom{N}{i} \le N^D+1$$

From the question, we know that the hypothesis H_{3A} that consists of all d-3-1 neural networks' maximum combination amount is

$$B(N, d+1)^3$$

Because

$$B(N,d+1) \le \sum_{i=0}^{d} C_i^N \le N^d+1 \le N^{d+1}+1$$, we can know that

$$B(N,d+1)^3 = (N^{d+1}+1)^3 = N^{3(d+1)} + 3N^{2(d+1)} + 3N^{d+1} + 1,$$

and while $\Delta = 3(d+1)+1$, $N \ge \Delta \ge 4$,

$$B(N,d+1)^3 = N^{3(d+1)} + 3N^{2(d+1)} + 3N^{d+1} + 1$$

$$< N^{3(d+1)} + N^{3(d+1)} + N^{3(d+1)} + 1 = 3N^{3(d+1)} + 1 < N^{3(d+1)+1} + 1$$

while $N \ge 3\Delta \log_2 \Delta$,

$$(N^{d+1}+1)^3 < N^{3(d+1)+1} + 1 \le 2^N$$

$$VC < 3\Delta \log_2 \Delta = 3(3(d+1)+1)\log_2(3(d+1)+1)$$

#