SET-6

6.1) Deep Learning

2) Exercise 5.18 (Bishop (2006))

* On adding a new skip layer with connections directly from the inputs to the outputs, our model can be represented as:

$$x_i,\ i=1,\dots,D \xrightarrow{w_{ji}^{(1)}} z_j,\ j=1,\dots,M \xrightarrow{w_{kj}^{(2)}} y_k,\ k=1,\dots,K$$

$$x_i \xrightarrow{w_{ki}^{(3)}} y_k$$

* The equations in Section 5.3.2 mostly remain the same, except Eq. (5.64), which is now

$$y_k = \sum_{j=0}^{M} w_{kj}^{(2)} z_j + \sum_{i=0}^{D} w_{ki}^{(3)} x_i$$

* The eq. for the derivative of the error function w.r.t. $w_{ki}^{(3)}$ is

$$\frac{\partial E_n}{w_{ki}^{(3)}} = (y_n - t_n)\, \partial_{w_{ki}^{(3)}} y_k = (y_n - t_n) x_i$$

$$\equiv \delta_k x_i$$

6.2) Representation Learning

1) a) Exercise 23.1

1) * We know that if V, W are finite dimensional vector spaces, & T is a linear transformation from V to W, then the image of T is a finite-dimensional subspace of W and

$$\dim(V) = \dim(\text{null}(T)) + \dim(\text{image}(T))$$

* $\therefore$ for $\dim(\text{null}(A)) \geq 1$, $\exists\ v \neq 0$ s.t. $Av = A0 = 0$.

$\therefore \exists\ u \neq v \in \mathbb{R}^n$ s.t. $Au = Av$.

2) We have that $Au = Av$ for some $u \neq v \in \mathbb{R}^n$.

$\Rightarrow$ For any recovery function f, $f(Au) = f(Av)$

$\therefore$ Exact recovery of a linear compression scheme is infeasible.

b) <u>Exercise 23.3</u>

* Let X be a matrix with its j-th column as $\psi(x_j)$.
* We find the spectral decomposition of $X^T X$, using the results from section 23.1.1 (A more efficient solution for the case $d \gg m$).
* Note that $(X^T X)_{ij} = K(x_i, x_j)$, thus the eigendecomposition of $X^T X$ can be found in polynomial time.
* Let V be the matrix whose columns are the n leading eigenvectors of $X^T X$, and let D be a diagonal $n \times n$ matrix whose diagonal consists of the corresponding eigenvalues.
* Denote by U be the matrix whose columns are the n leading eigenvectors of $X X^T$. ~~We now show how to~~
* We now show how to project the data without maintaining the matrix U. For every $x \in \mathcal{X}$, the projection $U^T \phi(x)$ is calculated as follows:

$$U^T \phi(x) = D^{-\frac{1}{2}} V^T X^T \phi(x) = D^{-\frac{1}{2}} V^T \begin{pmatrix} K(x_1, x) \\ \vdots \\ K(x_m, x) \end{pmatrix}$$

c) Exercise 23.4

1) * We have that $\forall w \in \mathbb{R}^d$ s.t. $\|w\|=1$ & every $i \in [m]$,

$$(\langle w, x_i \rangle)^2 = tr(w^T x_i x_i^T w)$$

* Hence

$$\underset{w:\|w\|=1}{\operatorname{argmax}} Var[\langle w, x\rangle] = \underset{w:\|w\|=1}{\operatorname{argmax}} \frac{1}{m}\sum_{i=1}^{m} (\langle w, x_i\rangle)^2$$

$$= \underset{w:\|w\|=1}{\operatorname{argmax}} \frac{1}{m}\sum_{i=1}^{m} tr(w^T x_i x_i^T w)$$

* We see that our equation reduces to the PCA problem for $n=1$.
* Hence the optimal solution of our the variance maximization problem is to set w to be the first principle vector of $x_1, \dots, x_m$.

2) * We know that

$$E[(\langle w_1, x\rangle)(\langle w, x\rangle)] = w_1^T E[xx^T] w = m w_1^T A w, \text{ where } A = \sum_i x_i x_i^T$$

Since w is an eigen vector of A and $E[(\langle w_1, x\rangle)(\langle w, x\rangle)] = 0$

$$\Leftrightarrow \langle w_1, w\rangle = 0$$

* $\therefore$ Our problem reduces to

$$w^* = \underset{w:\|w\|=1,\, E[(\langle w_1, x\rangle)(\langle w, x\rangle)]=0}{\operatorname{argmax}} Var[\langle w, x\rangle]$$

$$= \underset{w:\|w\|=1,\, \langle w_1, w\rangle = 0}{\operatorname{argmax}} Var[\langle w, x\rangle]$$

$$\Rightarrow w^* = \underset{w:\|w\|=1,\, \langle w, w_1\rangle = 0}{\operatorname{argmax}} \frac{1}{m}\sum_{i=1}^{m} (\langle w, x_i\rangle)^2$$

$$= \underset{w:\|w\|=1,\, \langle w, w_1\rangle = 0}{\operatorname{argmax}} tr(w^T \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T w)$$

* Note that the PCA problem for $n=2$ is equivalent to finding a unitary matrix $W \in \mathbb{R}^{d\times 2}$ such that

$$(W^T \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T W) \text{ is maximized.}$$

* We know that the columns of the optimal matrix, w_1, w_2 are the two first principal vectors of $x_1, \dots, x_m$.

$$\Rightarrow W^T \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T W = w_1^T \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T w_1 + w_2^T \frac{1}{m}\sum_{i=1}^{m} x_i x_i^T w_2$$

- Since w^* and w_1 are orthonormal, we obtain that

$$w_1^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w_1 + w_2^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w_2 \geq w_1^T \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w_1 + w^{*T} \frac{1}{m} \sum_{i=1}^{m} x_i x_i^T w^*$$

$\therefore$ we conclude that $w^* = w_2$.

afthy (2022) Exercise 20.5

20.5) a) Covariance of the deflated matrix is given by

$$\tilde{C} = \frac{1}{n}\left((I - v_1 v_1^T) X^T X (I - v_1 v_1^T)\right)$$

$$= \frac{1}{n}\left((X^T X - v_1 v_1^T X^T X)(I - v_1 v_1^T)\right)$$

$$= \frac{1}{n}\left((X^T X - v_1 n \lambda_1 v_1^T)(I - v_1 v_1^T)\right)$$

$$= \frac{1}{n}\left((X^T X - X^T X v_1 v_1^T - \cancel{v_1 n \lambda_1 v_1^T} + \cancel{v_1 n \lambda_1 \overbrace{v_1^T v_1}^{I} v_1^T}\right)$$

$$= \frac{1}{n}\left(X^T X - n \lambda_1 v_1 v_1^T\right)$$

$$= \frac{1}{n} X^T X - \lambda_1 v_1 v_1^T$$

b) As $\tilde{X} \in (d-1)$ subspace orthogonal to v_1,

$\Rightarrow$ u must be orthogonal to v_1

$\therefore \Rightarrow u^T v_1 = 0$ & $u^T u = 1$

$\therefore u = v_2$

c)
```
def function [V, λ] = simplePCA(C, K, f) {
    d = C.length();
    V = zeroes(d, K);
    for j in range(1, K) {
        [λ(j), V(:,j)] = f(C);
        C = C - λ(j) * V(:,j) * V(:,j)';
    }
}
```