Homework 1

Deep Learning

Duarte Calado de Almeida
95565

André Lopes Rodrigues
96576

Question 1

Question 2

Question 3

1. Let x_i denote the i-th component of $\boldsymbol{x}$ and let W_{ij} denote the entry in the i-th row and j-th column in matrix $\boldsymbol{W}$. We then have that:

$$
\begin{aligned}
h_i(\boldsymbol{x}) &= g\left(\sum_{j=1}^{D} W_{ij}x_j\right) = \left(\sum_{j=1}^{D} W_{ij}x_j\right)^2 = \left(\sum_{j=1}^{D} W_{ij}x_j\right)\left(\sum_{k=1}^{D} W_{ik}x_k\right) \\
&= \sum_{j=1}^{D}\sum_{k=1}^{D} W_{ij}x_jW_{ik}x_k = \sum_{j=1}^{D} W_{ij}^2x_j^2 + \sum_{j=1}^{D}\sum_{k\neq j} W_{ij}W_{ik}x_jx_k \\
&= \sum_{j=1}^{D} W_{ij}^2x_j^2 + \sum_{j=1}^{D}\sum_{k=1}^{j-1} W_{ij}W_{ik}x_jx_k + \sum_{j=1}^{D}\sum_{k=j+1}^{D} W_{ij}W_{ik}x_jx_k \\
&= \sum_{j=1}^{D} W_{ij}^2x_j^2 + \sum_{k=1}^{D}\sum_{j=k+1}^{D} W_{ij}W_{ik}x_jx_k + \sum_{j=1}^{D}\sum_{k=j+1}^{D} W_{ij}W_{ik}x_jx_k \\
&= \sum_{j=1}^{D} W_{ij}^2x_j^2 + \sum_{k=1}^{D}\sum_{j=k+1}^{D} W_{ij}W_{ik}(2x_jx_k) \\
&= \begin{bmatrix} W_{i1}^2 & W_{i1}W_{i2} & \dots & W_{i1}W_{iD} & W_{i2}^2 & W_{i2}W_{i3} & \dots & W_{i(D-1)}^2 & W_{i(D-1)}W_{iD} & W_{iD}^2 \end{bmatrix}
\begin{bmatrix} x_1^2 \\ 2x_1x_2 \\ \dots \\ 2x_1x_D \\ x_2^2 \\ 2x_2x_3 \\ \dots \\ x_{D-1}^2 \\ 2x_{D-1}x_D \\ x_D^2 \end{bmatrix}
\end{aligned}
$$

As such, $\boldsymbol{h}$ is linear in some feature transformation $\boldsymbol{\phi}$, that is, $\boldsymbol{h}$ can be written as $\boldsymbol{A}_\Theta\boldsymbol{\phi}(\boldsymbol{x})$. In particular, we have that such matrix $\boldsymbol{A}_\Theta$ can be defined as:

$$\boldsymbol{A}_\Theta = \begin{bmatrix} - \ \boldsymbol{a}_1^T \ - \\ - \ \boldsymbol{a}_2^T \ - \\ \vdots \\ - \ \boldsymbol{a}_D^T \ - \end{bmatrix}$$

where

$$\boldsymbol{a}_i = \begin{bmatrix} W_{i1}^2 & W_{i1}W_{i2} & \dots & W_{i1}W_{iD} & W_{i2}^2 & W_{i2}W_{i3} & \dots & W_{i(D-1)}^2 & W_{i(D-1)}W_{iD} & W_{iD}^2 \end{bmatrix}^T$$

and we can define the feature transformation $\boldsymbol{\phi}$ as:

$$\boldsymbol{\phi}(\boldsymbol{x}) = (x_1^2, 2x_1x_2, \dots, 2x_1x_D, x_2^2, 2x_2x_3, \dots, x_{D-1}^2, 2x_{D-1}x_D, x_D^2)$$

2. Given that the predicted output $\hat{y}$ is defined as:

$$\hat{y} = \boldsymbol{v}^T\boldsymbol{h}$$

the linearity of $\boldsymbol{h}$ in the feature transformation $\boldsymbol{\phi}(\boldsymbol{x})$ proven above leads to following equality:

$$\hat{y} = \boldsymbol{v}^T\boldsymbol{A}_\Theta\boldsymbol{\phi}(\boldsymbol{x}) = (\boldsymbol{A}_\Theta^T\boldsymbol{v})^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x})$$

where we take $\boldsymbol{c}_\Theta$ to be equal to $\boldsymbol{A}_\Theta^T\boldsymbol{v}$, thereby proving that $\hat{y}$ is also a linear transformation of $\boldsymbol{\phi}(\boldsymbol{x})$. Nevertheless, $\hat{y}$ is **not** linear in terms of the original parameters Θ. To see this, note that the model is now a linear combination of **products** of entries of $\boldsymbol{W}$ and $\boldsymbol{v}$ rather than the entries by themselves:

$$\hat{y} = \boldsymbol{v}^T\boldsymbol{A}_\Theta\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{D} v_i(w_i^T\boldsymbol{x})^2 = \sum_{i=1}^{K}\sum_{j=1}^{D}\sum_{k=1}^{D} v_iW_{ij}W_{ik}x_ix_k$$

where we define w_i to be the i-th row vector of matrix $\boldsymbol{W}$.

3. To prove the desired result, for $\boldsymbol{c}_\Theta$ defined in the previous subquestion and for any $\boldsymbol{c} \in \mathbb{R}^{\frac{D(D+1)}{2}}$, we make the observation that the inner products $\boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x})$ and $\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x})$ actually correspond to quadratic forms in $\boldsymbol{x}$:

$$\begin{aligned} \boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x}) &= \sum_{i=1}^{K} v_i(\boldsymbol{A}_\Theta\boldsymbol{\phi}(\boldsymbol{x}))_i^2 = \sum_{i=1}^{K} v_i(w_i^T\boldsymbol{x})_i^2 \\ &= \begin{bmatrix} w_1^T\boldsymbol{x} & w_2^T\boldsymbol{x} & \dots & w_K^T\boldsymbol{x} \end{bmatrix} \operatorname{diag}(\boldsymbol{v}) \begin{bmatrix} w_1^T\boldsymbol{x} \\ w_2^T\boldsymbol{x} \\ \dots \\ w_K^T\boldsymbol{x} \end{bmatrix} \\ &= (\boldsymbol{W}\boldsymbol{x})^T\operatorname{diag}(\boldsymbol{v})\boldsymbol{W}\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{W}^T\operatorname{diag}(\boldsymbol{v})\boldsymbol{W}\boldsymbol{x} \end{aligned}$$

and

$$\begin{aligned} \boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x}) &= \sum_{i=1}^{\frac{D(D+1)}{2}} c_i\phi_i(x) = \sum_{i=1}^{D} c_{(i-1)D+i}\, x_i^2 + \sum_{i=1}^{D}\sum_{j=i+1}^{D} c_{(i-1)D+j}\ (2x_ix_j) \\ &= \boldsymbol{x}^T\mathcal{M}(\boldsymbol{c})\boldsymbol{x} \end{aligned}$$

where $\mathcal{M}(\boldsymbol{c}) \in \mathbb{R}^{D\times D}$ is a symmetric matrix obatined through $\boldsymbol{c}$ such that:

– the diagonal and the part above the diagonal of the matrix $\mathcal{M}(\boldsymbol{c})$ is filled row-wise with the elements of vector $\boldsymbol{c}$, i.e.:

$$\mathcal{M}(\boldsymbol{c}) = \begin{bmatrix} c_1 & c_2 & c_3 & \dots & c_D \\ c_2 & c_{D+1} & c_{D+2} & \dots & c_{2D-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_D & c_{D-1} & c_{D-2} & \dots & c_{\frac{D(D+1)}{2}} \end{bmatrix}$$

– for $1 \leq 1 \leq D$ and $j < i$, $(\mathcal{M}(\boldsymbol{c}))_{ij} = (\mathcal{M}(\boldsymbol{c}))_{ji}$

Furthermore, we also recurr to the following lemma:

Lemma 1. *Two vectors* $\boldsymbol{a}$, $\boldsymbol{b} \in \mathbb{R}^{\frac{D(D+1)}{2}}$ *are equal if and only if* $\boldsymbol{a}^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{b}^T\boldsymbol{\phi}(\boldsymbol{x})$, *for all* $\boldsymbol{x} \in \mathbb{R}^D$

Proof. If $\boldsymbol{a} = \boldsymbol{b}$, then $\boldsymbol{a}^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{b}^T\boldsymbol{\phi}(\boldsymbol{x})$ is trivially verified. For the reverse implication, note that, according to the previously made observation, for any $\boldsymbol{x} \in \mathbb{R}^D$:

$$\boldsymbol{a}^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{b}^T\boldsymbol{\phi}(\boldsymbol{x}) \Rightarrow \boldsymbol{x}^T\mathcal{M}(\boldsymbol{a})\boldsymbol{x} = \boldsymbol{x}^T\mathcal{M}(\boldsymbol{b})\boldsymbol{x}$$

Since both $\mathcal{M}(\boldsymbol{a})$ and $\mathcal{M}(\boldsymbol{a})$ are symmetric and the associated quadratic forms are twice differentiable continuous, taking the hessian on both sides of the equation yields:

$$\mathcal{M}(\boldsymbol{a}) = \mathcal{M}(\boldsymbol{b})$$

that is, $\mathcal{M}(\boldsymbol{a})$ and $\mathcal{M}(\boldsymbol{b})$ are entrywise equal. In particular, we have that $a_i = b_i$, for $i = 1, \dots, \frac{D(D+1)}{2}$. ∎

We are now equipped with the tools needed for the proof. Since $\mathcal{M}(\boldsymbol{c})$ is symmetric, the **Spectral Decomposition Theorem** tells us that there is a matrix $\boldsymbol{Q}$ orthonormal and $\boldsymbol{\Lambda}$ diagonal such that $\mathcal{M}(\boldsymbol{c}) = \boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T$. Let $\boldsymbol{q_i}$ denote the eigenvector of $\mathcal{M}(\boldsymbol{c})$ that is present in i-th column of $\boldsymbol{Q}$ and let λ_i be the corresponding eigenvector (note that $\{\boldsymbol{q}_i\}_{i=1}^D$ forms an orthonormal basis of $\mathbb{R}^D$). Then, we can write $\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x})$ as:

$$\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x}) = \boldsymbol{x}^T\mathcal{M}(\boldsymbol{c})\boldsymbol{x} = (\boldsymbol{Q}^T\boldsymbol{x})^T\boldsymbol{\Lambda}(\boldsymbol{Q}^T\boldsymbol{x}) = \sum_{i=1}^{D}\lambda_i(\boldsymbol{q}_i^T\boldsymbol{x})^2$$

Now, if we assume that $K \geq D$, then we can construct the matrix $\boldsymbol{W}$ and vector $\boldsymbol{v}$ that make $\boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x})$ equal to $\boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x})$ in the following way:

– we make $\boldsymbol{v}$ to be equal to $(\lambda_1, \lambda_2, \dots, \lambda_D, \underbrace{0, \dots, 0}_{K-D\text{ times}})$;

– we make $\boldsymbol{W}$ to be equal to the vertical concatenation of $\boldsymbol{Q}^T$ with a $(K-D) \times D$ matrix of zeros, i.e.:

$$\boldsymbol{W} = \begin{bmatrix} \boldsymbol{Q}^T \\ \boldsymbol{0}_{(K-D)\times D} \end{bmatrix}$$

We then have that:

$$\boldsymbol{c}_\Theta^T\boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{K}\boldsymbol{v}_i(\boldsymbol{w}_i^T\boldsymbol{x})^2 = \sum_{i=1}^{D}\lambda_i(\boldsymbol{q}_i^T\boldsymbol{x})^2 = (\boldsymbol{Q}^T\boldsymbol{x})^T\boldsymbol{\Lambda}(\boldsymbol{Q}^T\boldsymbol{x}) = \boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{\Lambda}\boldsymbol{Q}^T\boldsymbol{x} = \boldsymbol{c}^T\boldsymbol{\phi}(\boldsymbol{x})$$

and, by Lemma 1, we prove that the previous choice of $\boldsymbol{W}$ and $\boldsymbol{v}$ originate a vector $\boldsymbol{c}_\Theta$ such that $\boldsymbol{c}_\Theta = \boldsymbol{c}$ ∎

In fact, we can relax the requirement $K \geq D$ to be $K \geq D - N$, where N is the dimension of the nullspace of $\mathcal{M}(\boldsymbol{c})$. Since $\mathcal{M}(\boldsymbol{c})$ is symmetric and thus diagonalizable, the set of indices $\mathcal{I}$ such that the eigenvectors $\{\boldsymbol{q}_i | i \in \mathcal{I}\}$ are not associated with eigenvalue zero has $D - N$ elements. As such:

$$\boldsymbol{c}^T \boldsymbol{\phi}(\boldsymbol{x}) = \sum_{i=1}^{D} \lambda_i (\boldsymbol{q}_i^T \boldsymbol{x})^2 = \sum_{i \in \mathcal{I}} \lambda_i (\boldsymbol{q}_i^T \boldsymbol{x})^2$$

and thus we can take $\boldsymbol{W}$ to be equal to the concatenation of the matrix $\tilde{\boldsymbol{Q}}^T$ with a matrix of $N \times K$ zeros (where $\tilde{\boldsymbol{Q}} \in \mathbb{R}^{D \times (D-N)}$ and $\tilde{\boldsymbol{q}}_k = \boldsymbol{q}_{i_k}$, $1 \leq k \leq D - N$). Following the previously made argument, we would obtain again a vector $\boldsymbol{c}_\Theta$ equal to $\boldsymbol{c}$.
Now, if $K < D - N$, then the nullspace of $\boldsymbol{W}$ has at least dimension $D - (D - N - 1) = N + 1$. Since the rank of $\mathcal{M}(\boldsymbol{c})$ is $D - N$, the dimensions of the row space of $\mathcal{M}(\boldsymbol{c})$ and the nullspace of $\boldsymbol{W}$ sum up to $D + 1$ and thus there is exists some non-null vector $\boldsymbol{x}^*$ that is in the rowspace of $\mathcal{M}(\boldsymbol{c})$ and in the nullspace of $\boldsymbol{W}$. Choosing $\boldsymbol{c}$ to be a vector such that $\mathcal{M}(\boldsymbol{c})$ is positive semidefinite, for example:

$$\mathcal{M}(\boldsymbol{c}) = \text{diag}(\underbrace{1, \dots, 1}_{D-N\text{times}}, \underbrace{0, \dots, 0}_{N\text{times}})$$

we have:

$$\boldsymbol{c}_\Theta^T \boldsymbol{\phi}(\boldsymbol{x}^*) = \boldsymbol{x}^{*T} \boldsymbol{W}^T \text{diag}(\boldsymbol{v}) \boldsymbol{W} \boldsymbol{x}^* = \boldsymbol{0}$$

and

$$\boldsymbol{c}^T \boldsymbol{\phi}(\boldsymbol{x}^*) = \sum_{i \in \mathcal{I}} \lambda_i (\boldsymbol{q}_i^T \boldsymbol{x}^*)^2 > 0$$

since $\lambda_i > 0$ ($i \in \mathcal{I}$), and $(\boldsymbol{q}_i^T \boldsymbol{x}^*)^2 > 0$ for at least one $i \in \mathcal{I}$, as x^* belongs to the rowspace of $\mathcal{M}(\boldsymbol{c})$ (which is spanned by $\{\boldsymbol{q}_i | i \in \mathcal{I}\}$).
We have thus constructed an instance where $K < D$ and there is no choice of parameters $\boldsymbol{W}$ and $\boldsymbol{v}$ that make $\boldsymbol{c}_\Theta^T$ equal to $\boldsymbol{c}$.

4.