

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 \mathbf{x} and let w_{ij} denote the entry in the i -th row and j -th column in matrix \mathbf{W} . We then have that:

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

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

$$\mathbf{A}_\Theta = \begin{bmatrix} -\mathbf{a}_1^T & - \\ -\mathbf{a}_2^T & - \\ \vdots & \\ -\mathbf{a}_K^T & - \end{bmatrix}$$

where

$$\mathbf{a}_i = \left[w_{i1}^2 \quad w_{i1}W_{i2} \quad \dots \quad w_{i1}w_{iD} \quad w_{i2}^2 \quad w_{i2}w_{i3} \quad \dots \quad w_{i(D-1)}^2 \quad w_{i(D-1)}w_{iD} \quad w_{iD}^2 \right]^T$$

and so $\mathbf{A}_\Theta \in \mathbb{R}^{K \times \frac{D(D+1)}{2}}$ (since $\sum_{k=1}^D k = \frac{D(D+1)}{2}$). Furthermore, we can define the feature transformation $\phi : \mathbb{R}^D \mapsto \mathbb{R}^{\frac{D(D+1)}{2}}$ as:

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

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

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

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

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

where we take \mathbf{c}_Θ to be equal to $\mathbf{A}_\Theta^T \mathbf{v}$, thereby proving that \hat{y} is also a linear transformation of $\phi(\mathbf{x})$. However, \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 \mathbf{W} and \mathbf{v} rather than being linear in **each** individual entry:

$$\hat{y} = \mathbf{v}^T \mathbf{A}_\Theta \phi(\mathbf{x}) = \sum_{i=1}^D v_i (\mathbf{w}_i^T \mathbf{x})^2 = \sum_{i=1}^K \sum_{j=1}^D \sum_{k=1}^D v_i w_{ij} w_{ik} x_i x_k$$

where we define \mathbf{w}_i to be the vector in the i -th row of matrix \mathbf{W} .

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

$$\begin{aligned} \mathbf{c}_\Theta^T \phi(\mathbf{x}) &= \mathbf{v}^T \mathbf{h} = \sum_{i=1}^K v_i (\mathbf{A}_\Theta \phi(\mathbf{x}))_i^2 = \sum_{i=1}^K v_i (\mathbf{w}_i^T \mathbf{x})^2 \\ &= [\mathbf{w}_1^T \mathbf{x} \quad \mathbf{w}_2^T \mathbf{x} \quad \dots \quad \mathbf{w}_K^T \mathbf{x}] \text{diag}(\mathbf{v}) \begin{bmatrix} \mathbf{w}_1^T \mathbf{x} \\ \mathbf{w}_2^T \mathbf{x} \\ \vdots \\ \mathbf{w}_K^T \mathbf{x} \end{bmatrix} \\ &= (\mathbf{W} \mathbf{x})^T \text{diag}(\mathbf{v}) \mathbf{W} \mathbf{x} = \mathbf{x}^T \mathbf{W}^T \text{diag}(\mathbf{v}) \mathbf{W} \mathbf{x} \end{aligned}$$

and

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

where $\mathcal{M}(\mathbf{c}) \in \mathbb{R}^{D \times D}$ is a symmetric matrix obtained from \mathbf{c} such that:

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

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

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

Furthermore, we also recur to the following lemma:

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

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

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

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

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

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

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

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

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

- we make \mathbf{v} to be equal to $(\lambda_1, \lambda_2, \dots, \lambda_D, \underbrace{0, \dots, 0}_{K-D \text{ times}})$;
- we make \mathbf{W} to be equal to the vertical concatenation of \mathbf{Q}^T with a $(K-D) \times D$ matrix of zeros, i.e.:

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

We then have that:

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

and, by Lemma 1, we prove that the previous choice of \mathbf{W} and \mathbf{v} originate a vector \mathbf{c}_Θ such that $\mathbf{c}_\Theta = \mathbf{c}$. Furthermore, we have proven that the sets of classifiers $\mathcal{C}_1 = \{\mathbf{c}_\Theta^T \phi(\mathbf{x}) : \Theta = (\mathbf{W}, \mathbf{v}) \in \mathbb{R}^{K \times D \times K}\}$ and $\mathcal{C}_2 = \{\mathbf{c}^T \phi(\mathbf{x}) : \mathbf{c} \in \mathbb{R}^{\frac{D(D+1)}{2}}\}$ are exactly the same, and so the original neural network reduces to a **linear model** in terms of \mathbf{c}_Θ . \square

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 $\mathcal{M}(\mathbf{c})$. Since $\mathcal{M}(\mathbf{c})$ is symmetric and thus diagonalizable, the set of indices \mathcal{I} such that the eigenvectors $\{\mathbf{q}_i : i \in \mathcal{I}\}$ are not associated with eigenvalue zero has $D - N$ elements. As such:

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

and thus we can take \mathbf{W} to be equal to the concatenation of the matrix $\tilde{\mathbf{Q}}^T$ with a matrix of $(K - (D - N)) \times D$ zeros (where $\tilde{\mathbf{Q}} \in \mathbb{R}^{D \times (D - N)}$ and each column vector $\tilde{\mathbf{q}}_k$ is equal to \mathbf{q}_{i_k} , where $1 \leq k \leq D - N$ and i_k is the k -th index present in \mathcal{I}). Following the previously made argument, we would obtain again a vector \mathbf{c}_Θ equal to \mathbf{c} .

Now, if $K < D - N$, then the nullspace of \mathbf{W} has at least dimension $D - (D - N - 1) = N + 1$. Since the rank of $\mathcal{M}(\mathbf{c})$ is $D - N$, the dimensions of the row space of $\mathcal{M}(\mathbf{c})$ and the nullspace of \mathbf{W} sum up to at least $D + 1$ and thus there is some non-null vector \mathbf{x}^* that is in the row space of $\mathcal{M}(\mathbf{c})$ and in the nullspace of \mathbf{W} . Choosing \mathbf{c} to be a vector such that $\mathcal{M}(\mathbf{c})$ is positive semidefinite, for example:

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

we have:

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

and

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

since $\lambda_i > 0$ (for $i \in \mathcal{I}$), and $(\mathbf{q}_i^T \mathbf{x}^*)^2 > 0$ for at least one $i \in \mathcal{I}$, as \mathbf{x}^* belongs to the row space of $\mathcal{M}(\mathbf{c})$ (which is spanned by $\{\mathbf{q}_i : i \in \mathcal{I}\}$).

We have thus constructed an instance where $K < D$ and there is no choice of parameters \mathbf{W} and \mathbf{v} that make \mathbf{c}_Θ^T equal to \mathbf{c} and so the model cannot be parametrized by \mathbf{c}_Θ^T in this case.

4. Given that $\hat{y} = \mathbf{c}_\Theta^T \phi(\mathbf{x})$, we can write the squared loss as:

$$L(\mathbf{c}_\Theta^T; \mathcal{D}) = \sum_{n=1}^N (\hat{y}_n(\mathbf{x}_n; \mathbf{c}_\Theta^T) - y_n)^2 = \sum_{n=1}^N (\mathbf{c}_\Theta^T \phi(\mathbf{x}_n) - y_n)^2 = \|\mathbf{X} \mathbf{c}_\Theta - \mathbf{y}\|_2^2$$

where $\mathbf{y} = (y_1, y_2, \dots, y_N)$. As such, the minimization of the squared loss corresponds to a linear least squares problem, which in this case has a unique solution \mathbf{c}_Θ^* that is found simply by setting the gradient to zero (since \mathbf{X} has rank N , $\mathbf{X}^T \mathbf{X}$ has also rank N and so it is invertible):

$$\begin{aligned} \nabla_{\mathbf{c}_\Theta} L(\mathbf{c}_\Theta^T; \mathcal{D}) = \mathbf{0} &\Leftrightarrow \nabla_{\mathbf{c}_\Theta} (\mathbf{X} \mathbf{c}_\Theta) \nabla_{\mathbf{z}} (\|\mathbf{z}\|_2^2)|_{\mathbf{z}=\mathbf{X} \mathbf{c}_\Theta - \mathbf{y}} = \mathbf{0} \Leftrightarrow 2 \mathbf{X}^T (\mathbf{X} \mathbf{c}_\Theta - \mathbf{y}) = \mathbf{0} \\ &\Rightarrow \mathbf{c}_\Theta^* = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y} \end{aligned}$$

Usually, loss functions of feedforward neural networks are non-convex in their parameters due to multiple compositions of non-linear activation functions, making the global minimization of said functions especially hard. In spite of that, since the presented architecture only has a single hidden layer with at least as many units as the input layer and uses only quadratic activations, we managed to reduce the underlying model to a **linear regression** by defining the underlying **feature transformation** ϕ and **weight vector** \mathbf{c}_Θ . Hence, the minimization of the loss function admits a closed form solution.