Machine Learning 2 – Group ESHG
Assignment 04

Willi Gierke, Arik Elimelech, Mehmed Halilovic, Leon Sixt

May 17, 2017

Exercise 1: Sparse Coding (5+5 P)

Let $\boldsymbol{x}_1, \dots, \boldsymbol{x}_N \in \mathbb{R}^d$ be a dataset of N examples. Let $\boldsymbol{s}_i \in \mathbb{R}^h$ be the source associated to example $\boldsymbol{x}_i$, and $W \in \mathbb{R}^{d \times h}$ be a matrix of size $d \times h$ that linearly projects the source onto the reconstructed example $\widehat{\boldsymbol{x}}_i$. We optimize the following sparse coding objective:

$$\min_{W, \boldsymbol{s}_1, \dots, \boldsymbol{s}_N} \quad \eta \|W\|_F^2 + \sum_{i=1}^N \|\boldsymbol{x}_i - W\boldsymbol{s}_i\|^2 + \lambda \|\boldsymbol{s}_i\|_1 \qquad \text{where} \quad \forall_{i=1}^N : \ \boldsymbol{s}_i \geq 0$$

(a) *Compute* the gradient of the objective with respect to the model parameters W. (i.e. compute the matrix $\frac{\partial E}{\partial W}$).

(b) *Compute* the gradient of the objective with respect to the sources $\boldsymbol{s}_i$ for each data point.

Exercise 2: Sparsifying Non-Linearities (10+10 P)

As an alternative to the sparse coding problem above, we would like to minimize the reparameterized objective of the form:

$$\min_{W, \boldsymbol{r}_1, \dots, \boldsymbol{r}_N} \quad \eta \|W\|_F^2 + \sum_{i=1}^N \|\boldsymbol{x}_i - W\boldsymbol{g}(\boldsymbol{r}_i)\|^2 + \lambda \|\boldsymbol{r}_i\|^2 \qquad \text{where} \quad \forall_{i=1}^N : \ \boldsymbol{r}_i \in \mathbb{R}^h$$

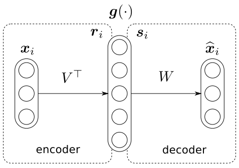
We call $\boldsymbol{r}_i$ the source parameter, and $\boldsymbol{s}_i = \boldsymbol{g}(\boldsymbol{r}_i)$ the reparameterized source associated to example $\boldsymbol{x}_i$. Note that the new objective no longer involves the minimization of an L_1-norm, and also does not include positivity constraints.

(a) *Find* a reparameterization function $\boldsymbol{g} : \mathbb{R}^h \to \mathbb{R}^h$ for which the optimization problem is equivalent to the one of Exercise 1.

(b) *Explain* what are the advantages and disadvantages of using such formulation of the optimization problem when compared to the original sparse coding problem. Your answer may include: (1) Applicability of gradient descent to find sources $\boldsymbol{s}_i$. (2) Ease of using an encoder to initialize the search for optimal sources.

Exercise 3: Auto-Encoders (10+10 P)

We now give an explicit definition of the encoder $\boldsymbol{r}_i = V^\top \boldsymbol{x}_i$, where $V \in \mathbb{R}^{d \times h}$ is a matrix of size $d \times h$. A graphical depiction of the resulting auto-encoder for $\boldsymbol{x}_i \in \mathbb{R}^3$ and $\boldsymbol{r}_i, \boldsymbol{s}_i \in \mathbb{R}^5$ is given below:



(a) Assuming the same error function as in Exercise 2, *use* the chain rule to express the gradient of the objective with respect to the encoder parameter $\frac{\partial E}{\partial V}$.

(b) *Explain* what are the advantages and disadvantages of using an autoencoder instead of directly optimizing $\boldsymbol{s}_i$ or $\boldsymbol{r}_i$. Your answer should include the following aspects: (1) Computational requirements of inferring sources $\boldsymbol{s}_i$ from observations $\boldsymbol{x}_i$. (2) Difficulty of the optimization problem. (3) Computational requirements at training time.

Exercise 4: Programming Exercise (50 P)

Download the code for Exercise sheet 4 on ISIS and follow the instructions.

Task 1

a

$$E = \eta ||W||_F^2 + \sum_{i=1}^{N} ||x_i - W s_i||^2 + \lambda ||s_i||_1$$

$$\frac{\partial E}{\partial W} = 2\eta W + \sum_{i=1}^{N} 2(x_i - W s_i)(-s_i)^T$$

$$\frac{\partial E}{\partial W} = 2\eta W - 2\sum_{i=1}^{N} (x_i - W s_i) s_i^T$$

b

$$\frac{\partial E}{\partial s_i} = \sum_{i=1}^{N} 2(x_i - W s_i)(-W) + \lambda$$

Task 2

a

To get the same optimization problem, the only thing that has to be adapted is the last term, since instead of $||\vec{s}_i||_1$ we have $||\vec{r}_i||^2$. The connection between $\vec{r}$ and $\vec{s}$ is given as $\vec{s}_i = \vec{g}(\vec{r}_i)$, so the condition that needs to be fulfilled is

$$\forall_{i=1}^{N} : \quad ||\vec{r}_i||^2 \overset{!}{=} ||\vec{g}(\vec{r}_i)||_1 \quad ,$$

where the $\vec{s}_i$ only have positive components.
This condition can be met by choosing the components of $\vec{g}(\vec{r}_i)$ as

$$g_j(\vec{r}_i) = r_{ij}^2$$

On the one hand, the square ensures that all components are positive, and on the other hand we get

$$||\vec{s}_i||_1 = \sum_{j=1}^{h} |s_{ij}| = \sum_{j=1}^{h} |g_j(\vec{r}_i)| = \sum_{j=1}^{h} |r_{ij}^2| = \sum_{j=1}^{h} r_{ij}^2 = ||\vec{r}_i||^2 \quad .$$

b

If $g(r)$ is differentiable, we can apply backpropagation as before to find the r's. If we pick $g(r)$ such that it depends on x i.e. $g(x)$, then we can also use it as a good initialization for the sources.

Task 3

a

$$\frac{\partial E}{\partial V^T} = \sum_{i=1}^{N} 2(x_i - WV^T x_i)(-Wx_i)$$

Therfore, deriving wrt. V:

$$\begin{aligned}\frac{\partial E}{\partial V} &= \sum_{i=1}^{N} (2(x_i - WV^T x_i)(-Wx_i)^T \\ &= -2\sum_{i=1}^{N} (Wx_i)^T (x_i - WV^T x_i)^T\end{aligned}$$

b

Before, we had to solve a optimization problem to find suitable sources. Now, we can utilize the encoder network to find the source codes. As long as the encoder network is small, we can find the codes more efficiently than by solving the optimization problem. The computational costs grow linearly with the complexity of the networks.