Exercises for the course
Machine Learning for Data Science
Winter Semester 2022/23

G. Montavon
Institute of Computer Science
Department of Mathematics and Computer Science
Freie Universität Berlin

Exercise Sheet 2 (theory part)

Exercise 1: Concentration of Squared Distances (10 + 10 + 5 P)

In this exercise, we would like to study analytically the the concentration of (squared) distances in high dimensions. Let $\boldsymbol{x} \in \mathbb{R}^d$ denote an input example. Input examples are drawn iid. from the distribution $\boldsymbol{x} \sim \mathcal{N}(0, I)$, where I is an identity matrix of size $d \times d$. We study the function

$$y(\boldsymbol{x}, \boldsymbol{x}') = \|\boldsymbol{x} - \boldsymbol{x}'\|^2.$$

measuring the squared Euclidean distance between randomly drawn data points.

(a) Express the mean of y as a function of the number of input dimensions d.

(b) Express the standard deviation of y as a function of d and σ.

(c) Show that the ratio $\mathrm{std}[y]/\mathrm{E}[y]$ is given by $\sqrt{\frac{2}{d}}$, and that therefore, square distances concentrate more as d grows.

Exercise 2: Gradient of T-SNE (15 + 15 + 15 P)

T-SNE is an embedding algorithm that operates by minimizing the Kullback-Leibler divergence between two discrete probability distributions p and q representing pairwise similarities of data points in the input space and in the embedding space respectively. Specifically, p_{ij} is a probability value quantifying how similar the points $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ are. Likewise, q_{ij} is a probability score quantifing how similar the same points are in embedded space (denoted by $\boldsymbol{y}_i$ and $\boldsymbol{y}_j$).

The elements of p and q are assumed to be strictly positive. Also, p and q are subject to the constraints of a probability distribution: $\sum_{i=1}^N \sum_{j=1}^N p_{ij} = 1$ and $\sum_{i=1}^N \sum_{j=1}^N q_{ij} = 1$. Once the exact probability functions are defined, the embedding algorithm proceeds by optimizing the function:

$$\begin{aligned} C &= D_{\mathrm{KL}}(p \,\|\, q) \\ &= \sum_{i=1}^N \sum_{j=1}^N p_{ij} \log\left(\frac{p_{ij}}{q_{ij}}\right) \end{aligned} \tag{1}$$

with respect to the coordinates of points in embedded space. Optimization can be performed using gradient descent.

In this exercise, we derive the gradient of the objective, first with respect to the probability scores q_{ij}, and then with respect to the embedding coordinates of which the scores q_{ij} are a function.

(a) *Show* that

$$\frac{\partial C}{\partial q_{ij}} = -\frac{p_{ij}}{q_{ij}} \tag{2}$$

(b) The probability matrix q is now reparameterized using the function

$$q_{ij} = \frac{\frac{1}{1+z_{ij}}}{\sum_{kl} \frac{1}{1+z_{kl}}}$$

Show using the chain rule for derivatives that

$$\frac{\partial C}{\partial z_{ij}} = (p_{ij} - q_{ij}) \cdot \frac{1}{1 + z_{ij}} \tag{3}$$

(c) The scores z_{ij} are finally reparameterized in terms of distances between points in embedded spaced as:

$$z_{ij} = \|\boldsymbol{y}_i - \boldsymbol{y}_j\|^2$$

Show using the chain rule for derivatives that

$$\frac{\partial C}{\partial y_i} = \sum_{j=1}^{N} 4 \cdot (p_{ij} - q_{ij}) \cdot \frac{1}{1 + z_{ij}} \cdot (\boldsymbol{y}_i - \boldsymbol{y}_j) \tag{4}$$