Exercises for the course
Machine Learning 2
Summer semester 2016

Abteilung Maschinelles Lernen
Institut für Softwaretechnik und theoretische Informatik
Fakultät IV, Technische Universität Berlin
Prof. Dr. Klaus-Robert Müller
Email: klaus-robert.mueller@tu-berlin.de

Exercise Sheet 2

This exercise sheet is based on the paper *Visualizing Data using t-SNE* by Laurens van der Maaten and Geoffrey Hinton, 2008, which is linked via ISIS.

Exercise 1: Kullback-Leibler Divergence (30 P)

The objective of t-SNE is is based on minimization of the Kullback-Leibler divergence between two probability distributions p and q.

$$C = D_{\mathrm{KL}}(P||Q) = \sum_j p_j \log\left(\frac{p_j}{q_j}\right)$$

where $\sum_j p_j = 1$ and $\sum_j q_j = 1$. Minimization of such quantity also intervenes in various probabilistic machine learning models. In this exercise, we derive the gradient of the Kullback-Leibler divergence, both with respect to the probability distribution itself, and a reparameterization of it.

- *Show* that
$$\frac{\partial C}{\partial q_i} = -\frac{p_i}{q_i}.$$
- The probability q_i that has to be optimized for all i is now reparameterized as $q_i = \frac{e^{x_i}}{\sum_k e^{x_k}}$. Here, x_i can be interpreted as the unnormalized log-probability associated to q_i. *Show* that
$$\frac{\partial C}{\partial x_i} = -p_i + q_i$$
- *Explain* which of these two gradients is the most appropriate for practical use in a learning algorithm. Motivate your choice (1) in terms of stability or boundedness of the gradient, and (2) in terms of ability to produce a valid probability distribution.

Exercise 2: Reversing t-SNE (20 P)

Section 3.2 of the t-SNE paper discusses the crowding problem and show how it can be useful to replace the distribution of pairs q_{ij} by a heavy tailed distribution, that make large distances more likely in the embedded space. As an experiment of thought, we now consider the opposite case where we set p_{ij} to be heavy-tailed and q_{ij} to be Gaussian distributed, that is:

$$p_{ij} = \frac{(1 + \|x_i - x_j\|^2/\sigma^2)^{-1}}{\sum_{k\neq l}(1 + \|x_k - x_l\|^2/\sigma^2)^{-1}}$$
$$q_{ij} = \frac{e^{-\|y_i - y_j\|^2}}{\sum_{k\neq l} e^{-\|y_k - y_l\|^2}}.$$

Explain how the learned embedding would relate qualitatively to the input data. *Discuss* a practical scenario where such choice of probability distributions would be useful.

Exercise 3: Programming Exercise (50 P)

Download the code for Exercise sheet 2 on ISIS and follow the instructions.