Exercises
Advanced Machine Learning
Fall 2019

Machine Learning Laboratory
Dept. of Computer Science, ETH Zürich
Prof. Joachim M. Buhmann
Web https://ml2.inf.ethz.ch/courses/aml/

Series 4. November 1, 2019
(Newton's Method, Perceptrons and LDA)

Teaching assistant: **Xinrui Lyu**
xlyu@inf.ethz.ch

Problem 1 (Newton's Method):

Newton's method was originally created to find a root $f(x^*) = 0$ of a given function $f(x)$ via the iteration

$$x_{n+1} = x_n - \frac{f(x_n)}{f'(x_n)}, \tag{1}$$

where $f'(x)$ denotes the derivative of f with respect to x. In optimization we are usually interested in finding the minimum of a function. This can be achieved using Newton's method to find a root of the first derivative $f'(x^*) = 0$, the optimization step reads then.

$$x_{n+1} = x_n - \frac{f'(x_n)}{f''(x_n)}. \tag{2}$$

However, to keep the notation simple we will use equation (1) and focus on one dimensional functions $f : \mathbb{R} \mapsto \mathbb{R}$. In this exercise we will examine the convergence of Newton's method and some common pitfalls.

1. **Complete Failure:** Show that Newton's method will never converge for $f(x) = \sqrt[3]{x}$ and $x_0 \neq 0$ (we define the third root of negative numbers $x < 0$ as $\sqrt[3]{x} = -\sqrt[3]{|x|}$). Why does it fail?

2. **Convergence for simple roots:** A sequence x_n converges with order m towards x^*, if there exists a constant C, such that $|x_{n+1} - x^*| \leq C|x_n - x^*|^m$ (as $n \to \infty$). Show that, if Newton's method converges, it will converge quadratic ($m = 2$) for a smooth function f with one root $f(x^*) = 0$ and $f'(x^*) \neq 0$ in a region around x^*.

   Hint: Use the Taylor expansion of f to second order around the point x_n.

3. **Convergence for higher order roots:** A root $f(x^*) = 0$ has order k if all derivatives $f^{(i)}(x^*) = 0$ vanish for $i < k$ and $f^{(k)}(x^*) \neq 0$.

   (a) Show that Newton's method converges linear ($m = 1$) for a smooth function f with one root $f(x^*) = 0$ of order $k > 1$ in a region around the point x^*.

   Hint: Write f as $f(x) = (x - x^*)^k g(x)$ with $g(x^*) \neq 0$ and use the Taylor approximation.

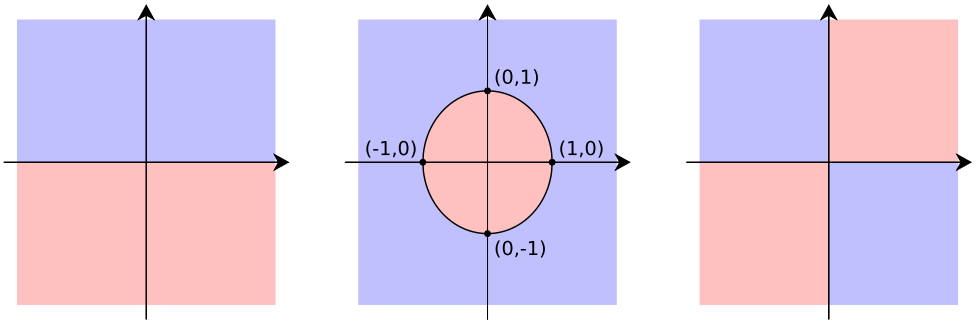
   (b) How should we adapt equation (1) to achieve quadratic convergence?

   (c) What implications does this have for an optimization problem where we want to minimize a function f by finding a root of the derivative f'?

Problem 2 (Perceptrons):

a) Perceptron algorithm Consider four data points $\mathbf{x}_1 = (1, 1)$, $\mathbf{x}_2 = (-1, 1)$, $\mathbf{x}_3 = (-1, -1)$ and $\mathbf{x}_4 = (1, -1)$ in the $\mathbb{R}^2$ space. $\mathbf{x}_1$ and $\mathbf{x}_2$ belong to class 1, while $\mathbf{x}_3$ and $\mathbf{x}_4$ belong to class 2. We want to seperate these two classes with a perceptron. Given initialization for the weight vector $\mathbf{a}_0 = (0, 1, 0)$, threshold $\theta = 10^{-6}$ and the learning rate update function $\eta(k) = 1/(k+1)$ where k is the number of iteration. Find the optimal weight vector $\mathbf{a}_k$ using the perceptron algorithm.
(Write down in details the steps of how you achieve the solution.)

b) Feature transformation Provide weights of a single-layer perceptron that correctly distinguishes the blue and red region (ignore the boundary) in the following three cases respectively.
(*Hint:* If necessary, apply a non-linear transformation $\phi(x_1, x_2)$, such that the resulting features are linear separable in the transformed space.)



Problem 3 (Fisher's Linear Discriminant Analysis):

Given N samples $\mathbf{x}_1, \ldots, \mathbf{x}_N$, where each sample belongs to class either 1 or 2. The solution of Fisher's discriminant function for binary classification ($k = 1, 2$) was $\mathbf{w}_F \propto \mathbf{\Sigma}_W^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$ with

$$\bar{\mathbf{x}}_k = \frac{1}{|\mathcal{C}_k|} \sum_{n \in \mathcal{C}_k} \mathbf{x}_n$$

$$\mathbf{\Sigma}_W = \sum_{k=1}^{2} \sum_{n \in C_k} (\mathbf{x}_n - \bar{\mathbf{x}}_k)(\mathbf{x}_n - \bar{\mathbf{x}}_k)^T,$$

where $\mathcal{C}_k$ $(k = 1, 2)$ denotes the set of samples belonging to class k. A class-dependent value y is assigned to each sample. And $y = \frac{N}{|\mathcal{C}_1|}$ for samples in class 1, and $y = -\frac{N}{|\mathcal{C}_2|}$ for the rest. Show that the solution of the least squares objective

$$E_{LS}(\mathbf{w}) = \frac{1}{2} \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n + w_0 - y_n)^2$$

is just Fisher's solution, that is, $\mathbf{w}_F = \min_{\mathbf{w}} E_{LS}(\mathbf{w})$