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 12, 2019 (Newton's Method, Perceptrons and LDA)

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

Problem 1 (Newton's Method):

1. **Complete Failure:** For $x > 0$ we have

$$\frac{f(x)}{f'(x)} = \frac{3x^{\frac{1}{3}}}{x^{-\frac{2}{3}}} = 3x$$

and for $x < 0$ we have

$$\frac{f(x)}{f'(x)} = \frac{-3|x|^{\frac{1}{3}}}{|x|^{-\frac{2}{3}}} = -3|x| = 3x.$$

The optimization step therefore reads

$$x_{n+1} = x_n - 3x_n = -2x_n.$$

This obviously diverges for all $x_0 \neq 0$. The main reason for the failure is the missing derivative at $x = 0$. Otherwise there would be a $\epsilon > 0$ such that the series would converge for $x_0 \in (-\epsilon, \epsilon)$.

2. **Convergence for simple roots:**

We start by expanding $f(x)$ around x_n to second order

$$f(x) = f(x_n) + f'(x_n)(x - x_n) + \frac{1}{2}f''(x_n)(x - x_n)^2,$$

Since x^* is a root of f we have

$$0 = f(x_n) + f'(x_n)(x^* - x_n) + \frac{1}{2}f''(x_n)(x^* - x_n)^2.$$

or equivalently

$$f(x_n) + f'(x_n)(x^* - x_n) = -\frac{1}{2}f''(\xi_n)(x^* - x_n)^2.$$

Since $f'(x^*) \neq 0$ we can assume that $f'(x_n) \neq 0$ for a region around x^*. We can therefore divide by $f'(x_n)$ and rearrange the terms to

$$x^* - \left(x_n - \frac{f(x_n)}{f'(x_n)}\right) = -\frac{f''(x_n)}{2f'(x_n)}(x_n - x^*)^2.$$

Using the optimization step and taking the absolute value we get

$$|x^* - x_{n+1}| = \left|\frac{f''(x_n)}{2f'(x_n)}\right| |x_n - x^*|^2.$$

Since x_k converges to x^*, $f'(x_n)$ will converge to $f'(x^*)$ and $f''(x_n)$ will converge to $f''(x^*)$. For large n we therefore have

$$|x^* - x_{n+1}| \leq C|x_n - x^*|^2,$$

for

$$C \geq \frac{f''(x^*)}{2f'(x^*)}.$$

3. **Convergence for higher order roots:**

(a) Let us define the error $\epsilon_n = x_n - x^*$ and $f(x) = (x - x^*)^k g(x)$ with $g(x^*) \neq 0$. We see that

$$\frac{f(x_n)}{f'(x_n)} = \frac{\epsilon_n^k g(x)}{k\epsilon_n^{k-1} g(x) + \epsilon_n^k g'(x)} = \frac{\epsilon_n g(x)}{kg(x) + \epsilon_n g'(x)} = \frac{1}{\frac{k}{\epsilon_n} + \frac{g'(x_n)}{g(x_n)}}.$$

Further we have

$$\epsilon_{n+1} = \epsilon_n - \frac{f(x_n)}{f'(x_n)} = \epsilon_n - \frac{1}{\frac{k}{\epsilon_n} + \frac{g'(x_n)}{g(x_n)}} = \epsilon_n \left(1 - \frac{1}{k + \epsilon_n \frac{g'(x_n)}{g(x_n)}}\right).$$

We can now approximate the right side of the equation using Taylor's expansion. Since x_n converges to x^* we have $\epsilon_n \ll 1$ for large enough n, using

$$\frac{1}{k+x} = \frac{1}{k} - \frac{x}{k^2} + \mathcal{O}(x^2)$$

we get

$$\epsilon_{n+1} = \epsilon_n \left(1 - \frac{1}{k} + \epsilon_n \frac{g'(x_n)}{k^2 g(x_n)} + \mathcal{O}(\epsilon_n^2)\right).$$

For $k = 1$ we have quadratic convergence since $1 - \frac{1}{k} = 0$, however, for $k > 1$ the convergence is only linear.

(b) If we change the optimization step to

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

we get

$$\epsilon_{n+1} = \epsilon_n \left(1 - \frac{1}{1 + \epsilon_n \frac{g'(x_n)}{kg(x_n)}}\right)$$

and achieve quadratic convergence.

(c) For optimization problems this is quite dangerous. Not only does it mean that Newton's method is attracted to saddle points, but it also converges very slowly towards them.

Problem 2 (Perceptrons):

a) Perceptron algorithm Augment the feature vector $\mathbf{x}$ to with 1 more dimension, resulting $\tilde{\mathbf{x}} = (1, \mathbf{x})$. Assign label $+1$ to class 1, and -1 to class 2. For iteration 1, since $\mathbf{a}_0^T \tilde{\mathbf{x}}_i = 1, \ -1, \ -1, 1$ for $i = 1, 2, 3, 4$ respectively, $\mathbf{x}_2$ and $\mathbf{x}_4$ are misclassified. There for update $a_1 = a_0 + \frac{1}{2}((-1, 1) - (1, -1)) = (0, 1, 0) + (0, -1, 1) = (0, 0, 1)$.

And because $\mathbf{a}_1^T \tilde{\mathbf{x}}_i = 1, \ 1, \ -1, -1$ for $i = 1, 2, 3, 4$ respectively, so no points is misclassified, the optimal solution is found.

b) Feature Transformation

1. This case is obviously linear separable. A possible weight vector would be $a = (0, 0, 1)$, if we define $\tilde{x} = (1, x_1, x_2)^T$. We then have $a^T \tilde{x} > 0$ for the blue region and $a^T \tilde{x} < 0$ for the red region.
2. We apply the feature transformation $\phi(x_1, x_2) = x_1^2 + x_2^2$ and see that $\phi(x_1, x_2) < 1$ for the red region and $\phi(x_1, x_2) > 1$ for the blue region. For $\tilde{x} = (1, \phi(x_1, x_2))^T$ and $a = (-1, 1)$, we have $a^T \tilde{x} > 0$ for the blue region and $a^T \tilde{x} < 0$ for the red region.
3. We apply the feature transformation $\phi(x_1, x_2) = x_1 x_2$ and see that $\phi(x_1, x_2) > 0$ for the red region and $\phi(x_1, x_2) < 0$ for the blue region. For $\tilde{x} = (1, \phi(x_1, x_2))^T$ and $a = (0, 1)$, we have $a^T \tilde{x} < 0$ for the blue region and $a^T \tilde{x} > 0$ for the red region.

Problem 3 (Fisher's Linear Discriminant Analysis):

By setting the gradient of E_{LS} w.r.t w_0

$$\nabla_{w_0} E_{LS} = \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n + w_0 - y_n) = 0$$

We get

$$w_0 = \frac{1}{N}\Big(-\mathbf{w}^T (\sum_{n=1}^{N} \mathbf{x}_n) + \sum_{n=1}^{N} y_n \Big) = \frac{1}{N}\Big(-\mathbf{w}^T (N\mathbf{m}) + \sum_{n \in \mathcal{C}_1} \frac{N}{|\mathcal{C}_1|} - \sum_{n \in \mathcal{C}_2} \frac{N}{|\mathcal{C}_2|} \Big) = -\mathbf{w}^T \mathbf{m},$$

where $\mathbf{m} = \frac{1}{N} \sum_{n=1}^{N} \mathbf{x}_n$.

The gradient of E_{LS} w.r.t. $\mathbf{w}$ is:

$$\begin{aligned}
\nabla_{\mathbf{w}} E_{LS} &= \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n + w_0 - y_n)\mathbf{x}_n \\
&= \sum_{n=1}^{N} (\mathbf{w}^T \mathbf{x}_n - \mathbf{w}^T \mathbf{m} - y_n)\mathbf{x}_n \\
&= \sum_{n=1}^{N} \mathbf{w}^T \mathbf{x}_n \mathbf{x}_n - \sum_{n=1}^{N} \mathbf{w}^T \mathbf{m} \mathbf{x}_n - \sum_{n=1}^{N} y_n \mathbf{x}_n \\
&= \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T \mathbf{w}_n - \sum_{n=1}^{N} \mathbf{x}_n \mathbf{m}^T \mathbf{w} - \sum_{n=1}^{N} y_n \mathbf{x}_n \\
&= (\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T)\mathbf{w}_n - (\sum_{n=1}^{N} \mathbf{x}_n)\mathbf{m}^T \mathbf{w} - \sum_{n \in \mathcal{C}_1}^{N} y_n \mathbf{x}_n - \sum_{n \in \mathcal{C}_2}^{N} y_n \mathbf{x}_n \\
&= (\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T)\mathbf{w}_n - N\mathbf{m}\mathbf{m}^T \mathbf{w} - \sum_{n \in \mathcal{C}_1}^{N} \frac{N}{|\mathcal{C}_1|}\mathbf{x}_n + \sum_{n \in \mathcal{C}_2}^{N} \frac{N}{|\mathcal{C}_2|}\mathbf{x}_n \\
&= (\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x_n}^T)\mathbf{w}_n - N\mathbf{m}\mathbf{m}^T \mathbf{w} - N\bar{\mathbf{x}}_1 + N\bar{\mathbf{x}}_2 \ (*)
\end{aligned}$$

Expand the formula for computing the within class scatter matrix:

$$\begin{aligned}
\mathbf{\Sigma}_W &= \sum_{n \in \mathcal{C}_1} (\mathbf{x}_n - \bar{\mathbf{x}}_1)(\mathbf{x}_n - \bar{\mathbf{x}}_1)^T + \sum_{n \in \mathcal{C}_2} (\mathbf{x}_n - \bar{\mathbf{x}}_2)(\mathbf{x}_n - \bar{\mathbf{x}}_2)^T \\
&= \sum_{n \in \mathcal{C}_1} \mathbf{x}_n \mathbf{x}_n^T - (\sum_{n \in \mathcal{C}_1} \mathbf{x}_n)\bar{\mathbf{x}}_1^T - \bar{\mathbf{x}}_1(\sum_{n \in \mathcal{C}_1} \mathbf{x}_n)^T + \sum_{n \in \mathcal{C}_1} \bar{\mathbf{x}}_1 \bar{\mathbf{x}}_1^T \\
&\quad + \sum_{n \in \mathcal{C}_2} \mathbf{x}_n \mathbf{x}_n^T - (\sum_{n \in \mathcal{C}_2} \mathbf{x}_n)\bar{\mathbf{x}}_2^T - \bar{\mathbf{x}}_2(\sum_{n \in \mathcal{C}_2} \mathbf{x}_n)^T + \sum_{n \in \mathcal{C}_2} \bar{\mathbf{x}}_2 \bar{\mathbf{x}}_2^T \\
&= \sum_{n \in \mathcal{C}_1} \mathbf{x}_n \mathbf{x}_n^T - |\mathcal{C}_1|\bar{\mathbf{x}}_1 \bar{\mathbf{x}}_1^T - \bar{\mathbf{x}}_1(|\mathcal{C}_1|\bar{\mathbf{x}}_1)^T + |\mathcal{C}_1|\bar{\mathbf{x}}_1 \bar{\mathbf{x}}_1^T \\
&\quad + \sum_{n \in \mathcal{C}_2} \mathbf{x}_n \mathbf{x}_n^T - |\mathcal{C}_2|\bar{\mathbf{x}}_2 \bar{\mathbf{x}}_2^T - \bar{\mathbf{x}}_2(|\mathcal{C}_2|\bar{\mathbf{x}}_2)^T + |\mathcal{C}_2|\bar{\mathbf{x}}_2 \bar{\mathbf{x}}_2^T \\
&= \sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T - |\mathcal{C}_1|\bar{\mathbf{x}}_1 \bar{\mathbf{x}}_1^T - |\mathcal{C}_2|\bar{\mathbf{x}}_2 \bar{\mathbf{x}}_2^T.
\end{aligned}$$

By rearanging, we can get

$$\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x}_n^T = \mathbf{\Sigma}_W + |\mathcal{C}_1|\bar{\mathbf{x}}_1 \bar{\mathbf{x}}_1^T + |\mathcal{C}_2|\bar{\mathbf{x}}_2 \bar{\mathbf{x}}_2^T (**)$$

Since the gradient of E_{LS} at the optimal solution $\mathbf{w}^*$ equal to 0, we replace the term $\sum_{n=1}^{N} \mathbf{x}_n \mathbf{x_n}^T$ with the RHS of $(**)$ and get

$$(\mathbf{\Sigma}_W + \underbrace{|\mathcal{C}_1|\bar{\mathbf{x}}_1 \bar{\mathbf{x}}_1^T + |\mathcal{C}_2|\bar{\mathbf{x}}_2 \bar{\mathbf{x}}_2^T - N\mathbf{m}\mathbf{m}^T}_{(***)})\mathbf{w}^* = N(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \tag{1}$$

$$\begin{aligned}
(***) &= |\mathcal{C}_1|\bar{\mathbf{x}}_1\bar{\mathbf{x}}_1^T + |\mathcal{C}_2|\bar{\mathbf{x}}_2\bar{\mathbf{x}}_2^T - \frac{1}{N}(|\mathcal{C}_1|\bar{\mathbf{x}}_1) + |\mathcal{C}_2|\bar{\mathbf{x}}_2)(|\mathcal{C}_1|\bar{\mathbf{x}}_1) + |\mathcal{C}_2|\bar{\mathbf{x}}_2)^T \\
&= \frac{N(|\mathcal{C}_1|\bar{\mathbf{x}}_1\bar{\mathbf{x}}_1^T + |\mathcal{C}_2|\bar{\mathbf{x}}_2\mathbf{x}^{-}{-}_2^T) - (|\mathcal{C}_1|\bar{\mathbf{x}}_1) + |\mathcal{C}_2|\bar{\mathbf{x}}_2)(|\mathcal{C}_1|\bar{\mathbf{x}}_1) + |\mathcal{C}_2|\bar{\mathbf{x}}_2)^T}{N} \\
&= \frac{(|\mathcal{C}_1| + |\mathcal{C}_2|)(|\mathcal{C}_1|\bar{\mathbf{x}}_1\bar{\mathbf{x}}_1^T + |\mathcal{C}_2|\bar{\mathbf{x}}_2\bar{\mathbf{x}}_2^T) - (|\mathcal{C}_1|\bar{\mathbf{x}}_1) + |\mathcal{C}_2|\bar{\mathbf{x}}_2)(|\mathcal{C}_1|\bar{\mathbf{x}}_1) + |\mathcal{C}_2|\bar{\mathbf{x}}_2)^T}{N} \\
&= \frac{|\mathcal{C}_1||\mathcal{C}_2|(\bar{\mathbf{x}}_1\bar{\mathbf{x}}_1^T + \bar{\mathbf{x}}_2\bar{\mathbf{x}}_2^T - \bar{\mathbf{x}}_1\bar{\mathbf{x}}_2^T - \bar{\mathbf{x}}_2\bar{\mathbf{x}}_1^T)}{N} \\
&= \frac{|\mathcal{C}_1||\mathcal{C}_2|(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T}{N}
\end{aligned}$$

The left hand side of the Equation (1) can thus be re-written as

$$\begin{aligned}
LHS &= (\boldsymbol{\Sigma}_W + \frac{|\mathcal{C}_1||\mathcal{C}_2|(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T}{N})\mathbf{w}^* \\
&= \boldsymbol{\Sigma}_W\mathbf{w}^* + \frac{|\mathcal{C}_1||\mathcal{C}_2|}{N}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)\big((\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{w}^*\big) \\
&= \boldsymbol{\Sigma}_W\mathbf{w}^* + C_0(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2),
\end{aligned}$$

where the scaler $C_0 = \frac{|\mathcal{C}_1||\mathcal{C}_2|}{N}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)^T\mathbf{w}^*$
Hence (1) can be transformed into

$$\mathbf{w}^* = \boldsymbol{\Sigma}_W^{-1}(N - C_0)(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2) \propto \boldsymbol{\Sigma}_W^{-1}(\bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2)$$