DATA.ML.200 Pattern Recognition and Machine Learning

Homework 3: Linear models in machine learning

This homework prepares you for the next week exercises.

1. **pen&paper** *Logistic regression with sum of squared error (SSE).*
   During the lectures we introduced the sigmoid decision rule

$$y = logsig\left(\mathbf{w}^T\mathbf{x}\right) = \frac{1}{1+e^{\mathbf{w}^T\mathbf{x}}} \ , \tag{1}$$

   where $\mathbf{w}$ is the vector of weights to be optimized. For example, for 2-dim data, $\mathbf{x} = (x_1\ x_2)^T$, the weights are $\mathbf{w} = (w_1\ w_2)^T$ or with intercept $\mathbf{w} = (w_0\ w_1\ w_2)^T$ ($\mathbf{x} = (1\ x_1\ x_2)^T$).

   Instead of the *maximum likelihood* (ML) loss derived during the lectures, we can also formulate the sum of squared error (SSE) loss, ℓ_{SSE}, which can be used to update the weights as

$$\mathbf{w}^{(t+1)} = \mathbf{w}^t - \mu\frac{\nabla\ell_{SSE}}{\nabla\mathbf{w}} \ . \tag{2}$$

   a) Define the SSE loss between the ground truth class labels $y[n]$ and sigmoid predictions $\hat{y}[n]$.
   b) Derive the gradient of the SSE loss to be used in the update rule (2).

2. **pen&paper** *Logistic regression with maximum likelihood (ML).*
   Let's now return to the *maximum likelihood* solution used in logistic regression. During the lectures we derived that the likelihood of samples from the class C_1 given $\mathbf{w}$ is

$$p(\mathbf{x} \in \mathcal{X}_{C_1}|\mathbf{w}) = \prod p(\mathbf{x_i}|\mathbf{w}) = \prod logsig(\mathbf{w}^T\mathbf{x}_i) = \prod \frac{1}{1+e^{\mathbf{w}^T\mathbf{x}_i}} \tag{3}$$

   and those of C_2

$$p(\mathbf{x} \in \mathcal{X}_{C_2}|\mathbf{w}) = \prod p(\mathbf{x_j}|\mathbf{w}) = \prod(1-p(\mathbf{x_j}|\mathbf{w})) = \prod(1-logsig(\mathbf{w}^T\mathbf{x}_j)) \tag{4}$$

   and the total likelihood of samples from the both class is

$$p(\mathbf{w}|\mathcal{X}) = \prod p(\mathbf{x}_i \in \mathcal{X}_{C_1}|\mathbf{w}) \prod p(\mathbf{x}_j \in \mathcal{X}_{C_2}|\mathbf{w}) = \ell_{ML} \ . \tag{5}$$

   a) Define the log-loss, $\ln \ell_{ML}$, that makes computation of the gradient easier.
   b) Derive the gradient of $\ln(\ell_{ML})$ that can be used in gradient descent (ascent)

$$\mathbf{w}^{(t+1)} = \mathbf{w}^t + \mu\frac{\nabla \ln \ell_{ML}}{\nabla\mathbf{w}} \ . \tag{6}$$

   Note: its probably easier to derive the log-loss gradients separately for C_1 and C_2 and the final gradient is the sum of the two.