Homework 1 Machine Learning I
Not graded

Introduction

1. This is a group assignment, so sign up in groups of three students.
2. It gives you an indication of the Math used in the Machine Learning I course. If you don't know some terminology used or the meaning of some of the equations then please catch up your Math.There is a page with useful online resources on Canvas.
3. Each group has to submit a **pdf** with their answers and explanation. **Please put your names and group at the top of the hand in**. Moreover put a picture of each group member at the front page of your report. This will make it easier for us to get to know you.
4. For questions about this homework assignment use the Discussion Board on Canvas.
5. Of course you may use a calculator or a programming environment such as Python. But your report should **not contain any code**. Explain your computations and results in English.

Exercise 1: Some geometry

1. Given the points $(4, 1)$ and $(-2, 3)$. Give the equation of the line through these two points. The equation should have the form
$$w_0 + w_1x_1 + w_2x_2 = 0$$
In this equation (w_0, w_1, w_2) are the so-called weights of the line. Observe that this weight vector $\mathbf{w} = (w_0, w_1, w_2)$ is not unique.
2. Make a plot of the line and the vector (w_1, w_2). The direction of (w_1, w_2) is called the positive side of the line.
3. Consider the point $(x_1, x_2) = (4, 3)$. Is this point on the positive side or negative side of the line? Of course this depends on your $\mathbf{w}$. Also compute $w_0 + w_1x_1 + w_2x_2$ for this point.
4. Show that the distance of the point $(x_1, x_2) = (4, 3)$ to the line is equal to
$$|w_0 + w_1x_1 + w_2x_2|/\|(w_1, w_2)\|$$
where $\|(w_1, w_2)\|$ is the length of the vector (w_1, w_2). Observe that the above equation is invariant under multiplying $\mathbf{w}$ with a nonzero scalar.
5. What is the distance of the line to the origin $(0, 0)$ and how can this be computed in terms of (w_0, w_1, w_2)?

Exercise 2: Some calculus

In this exercise we consider the logistic function σ defined by

$$\sigma(y) = 1/(1 + e^{-y})$$

1. Show that the derivative of σ with respect to y, i.e. $\sigma'(y)$, is given by

$$\sigma'(y) = \sigma(y)(1 - \sigma(y))$$

2. Show that:

   - for y large, i.e. $y \to \infty$, $\sigma(y) \approx 1$ (or in other words if $y \to \infty$ then $\sigma(y) \to 1$).
   - for $y = 0$, $\sigma(y) = 0.5$
   - for y small (meaning $y \to -\infty$) $\sigma(y) \approx 0$
   - for y large $\sigma'(y) \approx 0$

3. Now let's combine this with exercise 1: assume that $y = w_0 + w_1x_1 + w_2x_2$. Show that the partial derivative of σ with respect to x_1 is given by

$$\frac{\partial\sigma}{\partial x_1}(y) = \sigma(y)(1 - \sigma(y))w_1$$

   Hint: use the chain rule and part a.

4. Once again assume $y = w_0 + w_1x_1 + w_2x_2$. Show that:

   - $\frac{\partial\sigma}{\partial w_1}(y) = \sigma(y)(1 - \sigma(y))x_1$
   - the gradient of σ with respect to the vector $w = (w_0, w_1, w_2)$ (denoted by $\nabla_w\sigma(y)$) is given by $\nabla_w\sigma(y) = \sigma(y)(1 - \sigma(y))x$, where x is the vector $x = (1, x_1, x_2)$

5. Once again assume $y = w_0 + w_1x_1 + w_2x_2$. Show that:

   - if x_1 is positive then for $w_1 \to \infty$ and other weights and input constant $\frac{\partial\sigma}{\partial w_1}(y) \approx 0$.
   - if x_1 is negative then for $w_1 \to \infty$ $\frac{\partial\sigma}{\partial w_1}(y) \approx 0$.

Exercise 3: Some statistics and probability theory

Consider the very small data set $X = \{(1,4), (2,5), (-2,1), (-2,-3)\}$.

1. Compute the sample mean of X.

2. Compute the covariance matrix of X. This should be a 2x2 matrix.

3. Compute, based on the covariance matrix of part b, the correlation between x_1 (the first component) and x_2.

4. Compute the eigenvalues and eigenvectors of the covariance matrix of part b. What is the covariance matrix if we transform the data points into coordinates with respect to these eigenvectors?

5. Given a normal distribution p with mean $m = (1,1)$ and covariance matrix:

$$\begin{pmatrix} 1 & 1 \\ 1 & 4 \end{pmatrix}$$

   What is the likelihood that the point $x = (2,2)$ is generated by this distribution p?