Machine Learning 1 - Homework assignment 1

Available: Monday, October 29th, 2018
Deadline: Thursday 23.59, November 8th, 2018

General instructions
Unless stated otherwise, write down a derivation of your solutions. Solutions presented without a derivation that shows how the solution was obtained will not be awared with points.

1 Basic Linear Algebra and Derivatives

Question 1.1

Let $\mathbf{A} = \begin{bmatrix} 2 & 3 & 4 \\ 3 & 4 & 5 \\ 4 & 5 & 6 \end{bmatrix}$ and $\mathbf{B} = \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{bmatrix}$ and $\mathbf{b} = [1, 2, 5]^T$

Answer the following questions with enough intermediate steps to show you did not just simply use a calculator.

a) Compute $\mathbf{AB}$.

b) Are $\mathbf{A}$ and $\mathbf{B}$ invertible? How do you test for it? If so, calculate the inverses. Provide details only for your analysis of $\mathbf{A}$.

c) Is $\mathbf{AB}$ invertible? Why (not)?

d) Compute the solution set for the systems $\mathbf{Bx} = \mathbf{b}$ and $\mathbf{Ax} = \mathbf{b}$. What can we say about the second system due to the invertibility property you determined previously?

Question 1.2

Find the derivative of the following functions with respect to x.

a) $\frac{2}{x^2} + x^{-7} + x^3$

b) $xe^{-\sqrt[5]{x}}$

c) $\frac{1}{x} + \ln(x^2)$

d) $\sigma(x) = \frac{1}{1+e^{-x}}$ (standard logistic function, or "sigmoid function")

e) $\max\{0, x\}$ ("Rectified Linear Unit" (ReLu), which is important in Neural Networks)

What is the shape of the following gradients:

g) $\frac{df(x)}{dx}$ with $f : \mathbb{R} \to \mathbb{R}$, $x \in \mathbb{R}$

h) $\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}$ with $f : \mathbb{R}^n \to \mathbb{R}$, $x \in \mathbb{R}^n$
(We follow the convention that defines this gradient as a **row-vector**).

i) $\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}}$ with $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, $x \in \mathbb{R}^n$

Find the gradient of the following functions. Make their shapes explicit.

j) $\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}$ with $f(\boldsymbol{x}) = 2\exp(x_2 - \ln(x_1^{-1}) - \sin(x_3 x_1^2))$, $\boldsymbol{x} \in \mathbb{R}^3$

k) $\nabla_y h$ with $h(y) = (g \circ f)(y)$, where $g(\boldsymbol{x}) = x_1^3 + exp(x_2)$ and $\boldsymbol{x} := \boldsymbol{f}(y) = [y\sin(y), y\cos(y)]^T$. First show your understanding of the application of the chain rule in this example before "pluggin in" the actual derivatives.

l) We now assume that $\boldsymbol{x} := \boldsymbol{f}(y, z) = [y\sin(y) + z, y\cos(y) + z^2]^T$. Provide $\nabla_{y,z} h$. *Hint:* To determine the correct shape of $\nabla_{y,z} h$, view the input pair y and z as a vector $[y, z]^T$.

Question 1.3

The following questions are good practice in manipulating vectors and matrices and they are very important for solving for posterior distributions.

Compute the following gradients, assuming $\boldsymbol{\Sigma}^{-1}$ is symmetric, positive semi-definite and invertible.

a) $\nabla_{\boldsymbol{\mu}} \mathbf{x}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$

b) $\nabla_{\boldsymbol{\mu}} \boldsymbol{\mu}^T \boldsymbol{\Sigma}^{-1} \boldsymbol{\mu}$

c) $\nabla_{\mathbf{W}} \boldsymbol{f}$ where $\boldsymbol{f} = \mathbf{W}\mathbf{x}$ and assume $\mathbf{W} \in \mathbb{R}^{2x3}$ and $\mathbf{x} \in \mathbb{R}^3$. Follow Example 5.11 of the book *mathematics for machine learning*[1] to solve this.

d) $\nabla_{\mathbf{W}} f$, where $f = (\boldsymbol{\mu} - \mathbf{y}(\mathbf{W}, \mathbf{x}))^T \boldsymbol{\Sigma}^{-1} (\boldsymbol{\mu} - \mathbf{y}(\mathbf{W}, \mathbf{x}))$ and $\mathbf{y}(\mathbf{W}, \mathbf{x}) = \mathbf{W}^T\mathbf{x}$. Assume $\mathbf{W} \in \mathbb{R}^{M \times K}$Follow these steps:

[1] `https://mml-book.com`

1) Compute $\frac{\partial f}{\partial W_{i,j}}$. Hint: Use the chainrule by computing $\frac{\partial f}{\partial \mathbf{y}_k}\frac{\partial \mathbf{y}_k}{\partial W_{i,j}}$. Hint: You might find the notation $\mathbf{\Sigma}^{-1}_{:,k}$ helpful to define the kth row in the matrix $\mathbf{\Sigma}^{-1}$.

2) Convince yourself of the shapes of $\frac{\partial f}{\partial \mathbf{W}_{:,j}}$ and of $\frac{\partial f}{\partial \mathbf{W}_{i,:}}$

3) Combine your insights to answer the original question.

2 Probability Theory

For these questions you will practice manipulating probabilities and probability density functions using the sum and product rules.

Question 2.1

Suppose you meet Bart. Bart is a quiet, introvert man in his mid-fourties. He is very shy and withdrawn, invariably helpful but with little interest in people or in the world of reality. A meek and tidy soul, he has a need for order and structure, and a passion for detail.

a) What do you think: Is Bart a farmer or is he a librarian?

b) Suppose we formulate our belief such that we think that Bart is a librarian, given that he is an introvert, with 0.8 probability. If he were an extrovert, that probability is 0.1. Since the Netherlands is a social country, you assume that the probability of meeting an extrovert is 0.7 Define the random variables and values they can take on, both with symbols and numerically.

c) What is the probability that Bart is a librarian?

d) Suppose you looked up the actual statistics and find out that there are 1000 times more farmers than librarians in the Netherlands. Additionally, Bart's friend is telling you that if Bart were a librarian, the probability of him being a extrovert would be low, and equal to 0.1. Given this information, how should we update our belief in that Bart is a librarian, if he is an introvert? How does does influence the answer to the previous question: What is the probability that Bart is a librarian?

This question was adapted from an example in the book "Thinking Fast and Slow" by Daniel Kahneman. Highly recommended read, although not relevant for ML1.

Question 2.2

For this question you will compute the expression for the posterior parameter distribution for a simple data problem. Assume we observe N univariate data points $\{x_1, x_2, \ldots, x_N\}$. Further, we assume that they are generated by observing the outcome of a coin flip. Therefore the random variable representing the outcome of the coin flip is described by a Bernoulli distribution. The coin takes on the values $0, 1$, where the probability of throwing a 1 is equal to $\rho \in [0, 1]$. A Bernoulli random variable x has the following probability density function: $p(x) = \rho^x (1 - \rho)^{(1-x)}$.

a) Write down the expression for the likelihood of the observed datapoints.

b) Assume you observed the following coin throws: $[1, 1]$. Which value of ρ *most likely* generated this data-set?

We now assume some prior $p(\rho)$ over our model's parameter ρ.

c) Write down the general expression for the posterior over the parameter ρ assuming we observe the dataset D. Indicate which terms correspond to *prior*, *likelihood*, *evidence* and *posterior*. (Do not bother to fill in the (bernoulli) form of the likelihood)

d) Assume our prior $p(\rho)$ reflects our belief that the coin is fair. Under this prior, how will our posterior belief into the value of ρ change, assuming we observed the throws $[1, 1]$?