Machine Learning 1 - Homework assignment 1

Available: Monday September 2, 2019
Deadline: 17:00, Friday September 13, 2019

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. For this problem set, you do not have to hand solutions for question 1.1 and 1.2. All solutions should be typeset using LaTeX or something equivalent[1]. Late submission will *not* be graded.

1 Basic Linear Algebra and Derivatives

Note: this part of the homework is not graded. However, you are encouraged to make sure you know how to do these without any issue.

Question 1.1

$$\mathbf{A} = \begin{bmatrix} 0 & 1 & -1 \\ -1 & -3 & 2 \\ 2 & 1 & 1 \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} -4 & -1 & 2 \\ 3 & -4 & -5 \\ -5 & -2 & 2 \end{bmatrix}, \quad \text{and } \mathbf{b} = [1\ 2\ 5]^T. \tag{1}$$

Given the above, answer the following questions.

a) Compute $\mathbf{AB}$.

b) Are $\mathbf{A}$ and $\mathbf{B}$ invertible? Explain your answer. If invertible, provide the inverse.

c) Is $\mathbf{AB}$ invertible? Explain your answer.

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?

[1] Ask your TA if you want to use something other than LaTeX.

Question 1.2

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

a) $x^{-7} + \dfrac{4}{x^{-5}} + 3^x$

b) $x^3 + e^{\sqrt{x}}$

c) $x \ln(x)$

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:

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

g) $\frac{df(\boldsymbol{x})}{d\boldsymbol{x}}$ 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**).

h) $\frac{d\boldsymbol{f}(\boldsymbol{x})}{d\boldsymbol{x}}$ 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.

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

j) $\nabla_y h$ 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.

k) 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 $[y, z]^T$.

2 Multivariate Calculus

Question 2.1

The following questions are good practice in manipulating vectors and matrices.

Compute the following gradients, assuming $\mathbf{\Sigma}^{-1}$ is symmetric, positive semi-definite and invertible. Simplify your answers as much as possible.

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

b) $\nabla_{\mathbf{q}} - \mathbf{p}^T \log(\mathbf{q})$, where $\log(\cdot)$ is applied element wise.

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

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

3 Probability Theory

Question 3.1

A little warmup. This question is based on one of the early chapters in *Probability Theory: The Logic of Science* by E.T. Jaynes. Consider the following setting. You are driving down the street at night and suddenly you see a man climbing through a broken window of a jewelry store. Then, he runs away carrying a bag over his shoulder. For many of us, our gut reaction would be to think the man in question is a criminal. Why do we draw this conclusion instead of another scenario? Let's explore this using the methods of Probability Theory.

a) Explain in words: why would many people draw the conclussion that the man in question is a criminal? Try to think in terms of probability. [1–3 sentences is sufficient]

b) Show, formally, that the probability of us believing the man is a criminal given our observation is based on our beliefs of making this observation when the man is a criminal and making the observation when the man is *not* a criminal.

Let's assume one in every 10^5 people is in fact a criminal, the probability of making this observation when the man is not a criminal is $\frac{1}{10^6}$, and that of making this observation when the man is a criminal is 0.8.

[2] https://mml-book.com

c) Compute the probability of the man being a criminal based on our observations.

d) The next morning you learn that a group of kids have smashed multiple store fronts in your neighborhood. How does this change your beliefs, i.e., do you still think the man is a criminal? Explicitly state which *belief* updates you make and re-compute the probability of the man being a criminal given our observation. *Note, you do not have to do a Bayesian update or justify your belief update mathematically.*

Question 3.2

We observe N cards drawn from an unknown shuffled stack of cards. For each card, we record the suit of the card, e.g., $\mathbf{x}_i = \heartsuit$. Moreover, we represent each suit as a *one-hot* representation such that:

$$\heartsuit = [1,0,0,0],\ \clubsuit = [0,1,0,0],\ \diamondsuit = [0,0,1,0],\ \text{and } \spadesuit = [0,0,0,1]. \tag{2}$$

Hence, if the jth card is a the ace of $\spadesuit$, then $\mathbf{x}_{j4} = 1$ and $\mathbf{x}_{jk} = 0$ for $k \in \{1,2,3\}$. We can model this data with a Multinomial distribution $\text{Mult}(\rho_\heartsuit, \rho_\clubsuit, \rho_\diamondsuit, \rho_\spadesuit)$ (see page 74 in Bishop). Here, each ρ denotes the probability of drawing a card from a specific suit. For ease of notation we will use $\rho_1, \ldots, \rho_4$ for the remainder of this assignment.

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

b) We observe the following draws: $\{\mathbf{x}_1, \ldots, \mathbf{x}_8\} = \{\heartsuit, \heartsuit, \heartsuit, \heartsuit, \spadesuit, \spadesuit, \spadesuit, \spadesuit\}$. What is the maximum likelihood solutions for $\boldsymbol{\rho} = [\rho_1, \ldots, \rho_4]$? You can simply deduce the answer here. Computing it requires taking account of the constraints on $\boldsymbol{\rho}$, which we will cover later in this course.

Instead of modelling the suit of each card, we can choose to only model the suit color (red or black). This reduces our problem to a two class problem, which we can model using a Bernoulli(p) distribution, where p denotes the probability of a card being red.

c) Write down the expression for the likelihood for the card color in terms of $\mathbf{x}_1, \ldots, \mathbf{x}_N$, and p.

d) We now observe the following draws: $\{\mathbf{x}_1, \ldots, \mathbf{x}_8\} = \{\heartsuit, \heartsuit, \heartsuit, \heartsuit, \diamondsuit, \diamondsuit, \spadesuit, \spadesuit\}$. Which value of p most likely generated these observations? Show your calculations.

e) What is the relationship between $\boldsymbol{\rho}$ and p for a given set of observation $\{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$?

f) Write down the general expression for the posterior over the parameter p assuming we observe the dataset $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$. Indicate which parts correspond to the *prior*, the *likelihood*, *evidence* and *posterior*. (You do not have to fill in the Bernoulli for the likelihood, a general form suffices for this question)

g) Now assume we have chosen a Beta(α, β) prior over p. Compute the maximum a posteriori (MAP) estimate p_{MAP}. What values for α and β would you choose to encode the belief there is an equal probability of drawing a red or black card?