First assignment in Machine learning 1 – 2024 – Paper 1

1 Multivariate Calculus (Recommended timeline: TBD)

In this exercise, you are going to compute several gradients. Simplify your answers as much as possible, *and use index-notation for all your derivations.* Consider $\nabla_{\boldsymbol{x}} f(\boldsymbol{x})$ as the same with $\frac{df}{d\boldsymbol{x}}$.

Compute the following:

(*a*) $\nabla_{\boldsymbol{x}} \boldsymbol{\sigma}(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^m$ where $\boldsymbol{\sigma}$ denotes the Sigmoid function applied element-wise. [1 point]

(*b*) $\frac{d}{d\boldsymbol{w}} \boldsymbol{f}$ with $\boldsymbol{f} = \boldsymbol{X}\boldsymbol{w}$ with $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{w} \in \mathbb{R}^n$ [1 point]

(*c*) $\frac{d}{d\boldsymbol{w}} f$ with $f = \boldsymbol{w}^T \boldsymbol{X} \boldsymbol{w}$ with $\boldsymbol{X} \in \mathbb{R}^{n \times n}$ and $\boldsymbol{w} \in \mathbb{R}^n$ [1 point]

(*d*) $\frac{d}{d\boldsymbol{x}} \boldsymbol{\varsigma}(\boldsymbol{x}), \boldsymbol{x} \in \mathbb{R}^n$ where $\boldsymbol{\varsigma}(\boldsymbol{x})_i = \frac{\exp x_i}{\sum_{j=1}^{n} \exp x_j}$. Try to write it in matrix form making use of the diag function. [2 points]

(*e*) $\frac{d}{d\boldsymbol{\theta}} \|X\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$, with $X \in \mathbb{R}^{n \times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, and $\boldsymbol{y} \in \mathbb{R}^n$. Can you set this to zero and solve for $\boldsymbol{\theta}$? Congratulations, you derived the linear regression closed form on your own! [2 points]

(*f*) **BONUS:** When does the formula that you just obtained fail? Explain how L2 regularization (i.e., adding the penalty term $\lambda \|\boldsymbol{\theta}\|_2^2$, where $\lambda > 0$ is a hyperparameter) will fix this issue. [1 point]

First assignment in Machine learning 1 – 2024 – Paper 1

2 Full analysis of a distribution: Exponential distribution (Recommended timeline: TBD)

The Poisson process is a model for series of discrete events where an average time between events is known, but the exact time at which they occur is random. It is also assumed that the process is memoryless or Markovian, i.e. the occurrence of a new event is independent of the previous events.

In this exercise, we are interested in analyzing the exponential distribution, which is a continuous probability distribution commonly used to model the time between events in a Poisson process. There are many processes that are modeled this way. For example, it is often used to estimate the lifespan of electronic devices, where failure rates do not change as the device ages. In manufacturing, the time between the completion of successive products on an assembly line can be modeled using an exponential distribution, when the production rate is constant. Even the amount of money customers spend in one trip to the supermarket follows an exponential distribution!

Formally, an exponential distribution with rate parameter $\lambda > 0$ can be modeled as such:

$$p(x) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

The aim of this exercise is to familiarize you with arbitrary distributions. Note that by no means the following questions are the only things you might want to know about a distribution, but rather serve as a starting point for further research. Using these insights, answer the following questions:

Useful results:

- The cumulative distribution function (cdf) of the exponential distribution is:

$$P(X \leq x) = \begin{cases} 1 - e^{-\lambda x}, & x \geq 0 \\ 0, & x < 0 \end{cases}$$

(a) If $X \sim \text{Exp}(\lambda)$, prove that $\mathbb{E}[X] = \frac{1}{\lambda}$. If you learned Calculus in English, remember "ultraviolet voodoo". If you speak Spanish, remember "un día vi una vaca vestida de uniforme". [2 points]

(b) Students arrive to the ML1 lecture in the morning according to an exponential distribution with an average time between arrivals of 2 minutes. What is the probability that the time between two consecutive arrivals is less than 1 minutes?

[1 point]

(*c*) Consider a dataset $\boldsymbol{x} = [x_1, x_2, \cdots, x_N]$ which are independent and identically distributed (i.i.d.) non-negative random variables from an exponential distribution with the rate parameter λ. Derive the log-likelihood function for the parameter λ given the dataset. [1 point]

(*d*) One common method to estimate the parameters of the assumed probability distribution is called maximum likelihood optimization. In this approach, we wish to find the parameters of the distribution (in our case λ) that will maximize the likelihood function. Find the maximum likelihood estimator λ_{ML} for the likelihood function calculated in part c). [2 points]

(*e*) In general it is difficult to find a closed form expression for the posterior distribution of our model parameters because of the integral in the evidence. Rather than finding the full posterior distribution, we can find a point estimate of the model parameters. A common point estimate is the maximum a posteriori estimation, or the MAP, which estimates the model parameters as the mode of the posterior distribution, i.e.

$$\lambda_{\text{MAP}} = \arg\max_{\lambda} p(\lambda \mid \boldsymbol{x}).$$

Assume that the prior for the parameter λ is given by the Gamma distribution with hyperparameters α_1 and α_2:

$$p(\lambda|\alpha_1, \alpha_2) = \frac{\alpha_2^{\alpha_1}}{\Gamma(\alpha_1)} \lambda^{\alpha_1 - 1} e^{-\alpha_2 \lambda},$$

where Γ denotes the gamma function. Show that we can find λ_{MAP} by optimizing:

$$\lambda_{\text{MAP}} = \arg\max_{\lambda} \ (N + \alpha_1 - 1) \log \lambda - (\alpha_2 + \sum_{i=1}^{N} x_i)\lambda.$$

[2 points]

Hint: Show first that $\lambda_{\text{MAP}} = \arg\max_{\lambda} \log p(\boldsymbol{x} \mid \lambda) + \log p(\lambda)$.

(*f*) Find the MAP estimator λ_{MAP}. [1 point]

(*g*) In the case of a Exponential distribution with a Gamma prior, the resulting posterior distribution can be derived analytically. The resulting distribution is a Gamma distribution $\text{Gamma}(\alpha_1', \alpha_2')$.

Show that the posterior distribution is indeed a Gamma distribution.

[2 points]

Hint: The resulting distribution follows $\alpha_1' = N + \alpha_1$ and $\alpha_2' = (\sum_{i=1}^{N} x_i) + \alpha_2$.

3 General Multiple Outputs Linear Regression (Recommended timeline: TBD)

So far, all linear regression models assumed that the target t is a single target. In a more general case, however, we may wish to predict multiple targets $\mathbf{t}$.

One possibility is to perform an independent linear regression for each component of the target vector $\mathbf{t}$ by introducing a different set of basis functions for each component. In other words, if the target is a K-dimensional vector, we would perform a separate linear regression for each component t_i, $i \in 1, \dots, K$ of the target vector $\mathbf{t}$.

The other more common approach is to use the same set of basis function to model the target vector directly in the following form:

$$\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x}),$$

where $\mathbf{y}$ is the model prediction, $\mathbf{x}$ is a M-dimensional input vector, and $\mathbf{W}$ is matrix of parameters. $\boldsymbol{\phi}(\mathbf{x})$ is an M-dimensional vector with elements $\phi_j(\mathbf{x})$, and $\phi_0(\mathbf{x}) = 1$ as usual. Assume that the conditional distribution of the target vector to be an Gaussian of the form:

$$p\left(\mathbf{t} \middle| \mathbf{W}, \boldsymbol{\Sigma}\right) = \mathcal{N}\left(\mathbf{t} \middle| \mathbf{y}(\mathbf{x}, \mathbf{W}), \boldsymbol{\Sigma}\right),$$

where $\boldsymbol{\Sigma}$ is the covariance matrix, and $\mathbf{y}(\mathbf{x}, \mathbf{W}) = \mathbf{W}^{\mathrm{T}} \boldsymbol{\phi}(\mathbf{x})$. Assume that we have N observations $\mathbf{t_1}, \dots, \mathbf{t_N}$, which can be combined into a matrix $\mathbf{T}$ of size $N \times K$, such that the n^{th} row is given by $\mathbf{t}_n^{\mathrm{T}}$.

(a) What are the dimensions of the parameter matrix $\mathbf{W}$? [1 point]

(b) Write down the log-likehood. [1 point]

(c) Find the maximum likelihood solution $\mathbf{W}_{\mathrm{ML}}$ in the terms of feature matrix $\boldsymbol{\Phi}$ and the target matrix $\mathbf{T}$, and show that it is independent of the covariance matrix $\boldsymbol{\Sigma}$.

Hint: Make use of the following derivation identity:

$$\frac{\partial}{\partial \mathbf{A}} (\mathbf{x} - \mathbf{A}\mathbf{s})^T \mathbf{W} (\mathbf{x} - \mathbf{A}\mathbf{s}) = -2\mathbf{W}(\mathbf{x} - \mathbf{A}\mathbf{s})\mathbf{s}^T$$

[2 points]

(*d*) Show that the maximum likelihood solution for $\boldsymbol{\Sigma}$ is given by:

$$\boldsymbol{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\phi(\mathbf{x}_n)\right)\left(\mathbf{t}_n - \mathbf{W}_{\mathrm{ML}}^{\mathrm{T}}\phi(\mathbf{x}_n)\right)^{\mathrm{T}}$$

[2 points]

Substitute $\boldsymbol{\Omega} := \boldsymbol{\Sigma}^{-1}$ and differentiate with respect to $\boldsymbol{\Omega}$. Moreover, make use of the identities $\frac{d}{d\mathbf{A}} \log|\mathbf{A}| = (\mathbf{A}^{\mathbf{-1}})^{\mathbf{T}}$ and $\frac{d}{d\mathbf{X}}\mathbf{a}^{\mathbf{T}}\mathbf{X}^{\mathbf{T}}\mathbf{b} = \mathbf{b}\mathbf{a}^{\mathbf{T}}$.

4 Counting Fish (Recommended timeline: TBD)

Imagine you are in an aquarium, and you notice that each fish has a tag with a number attached in a fin. You see number 3, then 6, then 3 again, and finally 9. After asking an employee, they tell you that each fish has a consecutive number assigned, and that the sequence starts at one.

(a) Intuitively, how many fishes do you think there are after observing these four samples? [1 point]

(b) What is the most likely estimate? State your assumptions clearly. [1 point]

(c) Is this estimation unbiased, or does it overestimate / underestimate the true number of fishes? If the estimation is biased, compute the bias. [2 points]

Fun fact: the British used a similar technique to estimate how many tanks Nazis were building per month. This estimation was significantly better than the figures that the British spies were reporting.