Machine Learning 1, 2025, Homework 1

Instructions & Conventions

- **Show Your Work:** For all questions, provide the intermediate steps of your derivation. Answers without justification will not receive full marks. Please simplify your final answers as much as possible.

- **Gradient Convention:** In this assignment, we use the **numerator layout** convention.

  - The derivative of a scalar function $f(\boldsymbol{x})$ with respect to a vector $\boldsymbol{x} \in \mathbb{R}^n$ is a **row vector**:

    $$\frac{df}{d\boldsymbol{x}} = \nabla_{\boldsymbol{x}} f(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}$$

  - The derivative of a vector function $\boldsymbol{f}(\boldsymbol{x}) \in \mathbb{R}^m$ with respect to a vector $\boldsymbol{x} \in \mathbb{R}^n$ is the $m \times n$ **Jacobian matrix**:

    $$\frac{d\boldsymbol{f}}{d\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{x}} \\ \vdots \\ \frac{\partial f_m}{\partial \boldsymbol{x}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

- **Index Notation Derivations:** When a question asks you to *use index notation for your derivations*, first compute the derivative for arbitrary components (e.g., $\frac{\partial f_i}{\partial x_j}$). After deriving this general form, assemble the final answer as a matrix/vector and simplify.

Machine Learning 1 - HW1 – 2025 – Paper 1

1 Multivariate Calculus (7 points)

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

(*a*) $\nabla_{\boldsymbol{x}} \tanh(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^m$, where tanh is the hyperbolic tangent function applied element-wise. [1 point]

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

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

(*d*) $\frac{d}{d\boldsymbol{x}} \boldsymbol{\varsigma}(\boldsymbol{x})$ with $\boldsymbol{x} \in \mathbb{R}^n$, where $\boldsymbol{\varsigma}$ is the softmax function $\varsigma(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{k=1}^n \exp(x_k)}$. [2 points]

(*e*) $\frac{d}{d\boldsymbol{\theta}} \frac{1}{n} \|X\boldsymbol{\theta} - \boldsymbol{y}\|_2^2$, with $X \in \mathbb{R}^{n \times d}$, $\boldsymbol{\theta} \in \mathbb{R}^d$, $\boldsymbol{y} \in \mathbb{R}^n$. Set this to zero and solve for $\boldsymbol{\theta}$. [2 points]

Machine Learning 1 - HW1 – 2025 – Paper 1

2 Full analysis of a distribution: Poisson distribution (11 points)

The Poisson process is a model for a series of discrete events where the average number of events in a fixed interval is known. 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 Poisson distribution, which is a discrete probability distribution that expresses the probability of a given number of events occurring in a fixed interval of time or space. There are many processes that are modeled this way. For example, it is used to model the number of radioactive decays in a fixed time period, the number of calls arriving at a call center per hour, or the number of typos on a page. Formally, a Poisson distribution with rate parameter $\lambda > 0$ can be modeled as such:

$$p(k; \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}, \quad k \in \{0, 1, 2, \dots\}$$

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:

(a) If $K \sim \text{Poisson}(\lambda)$, prove that $\mathbb{E}[K] = \lambda$. [2 points]

(b) A customer support center receives an average of 4 calls per hour. Assuming the number of calls follows a Poisson distribution, what is the probability of receiving exactly 3 calls in a given hour? [1 point]

(c) Consider a dataset $\boldsymbol{k} = [k_1, k_2, \cdots, k_N]$ which are independent and identically distributed (i.i.d.) random variables from a Poisson distribution with the rate parameter λ. Derive the log-likelihood function for the parameter λ given the dataset. [1 point]

(d) Find the maximum likelihood estimator λ_{ML} for the likelihood function calculated in part c). [2 points]

(e) 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:

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

[2 points]

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

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

(*g*) In the case of a Poisson distribution with a Gamma prior, the resulting posterior distribution can be derived analytically. Show that the posterior distribution is indeed a Gamma distribution. [2 points]

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

3 General Multiple Outputs Linear Regression (6 points)

So far, our linear regression models have assumed that the target t is a single scalar value. In a more general case, we may wish to predict a K-dimensional target vector $\mathbf{t}$. A common approach is to use the same set of basis functions to model all components of the target vector, leading to a model of the form:

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

where $\mathbf{y}$ is the K-dimensional model prediction, $\mathbf{x}$ is the input vector, $\boldsymbol{\phi}(\mathbf{x})$ is an M-dimensional feature vector (with $\phi_0(\mathbf{x}) = 1$ by convention), and $\mathbf{W}$ is the parameter matrix. We assume the conditional distribution of the target vector is a spherical Gaussian:

$$p\left(\mathbf{t} \mid \mathbf{x}, \mathbf{W}, \sigma^2\right) = \mathcal{N}\left(\mathbf{t} \mid \mathbf{y}(\mathbf{x}, \mathbf{W}), \sigma^2\mathbf{I}\right),$$

where σ^2 is the variance and $\mathbf{I}$ is the $K \times K$ identity matrix. Given a dataset of N i.i.d. observations $\{(\mathbf{x}_n, \mathbf{t}_n)\}_{n=1}^{N}$, we can group the targets into an $N \times K$ matrix $\mathbf{T}$ and the feature vectors into an $N \times M$ design matrix $\boldsymbol{\Phi}$.

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

(*b*) Write down the log-likelihood function for the parameters $\mathbf{W}$ and σ^2, given the dataset $(\boldsymbol{\Phi}, \mathbf{T})$. [1 point]

(*c*) Find the maximum likelihood solution $\mathbf{W}_{\mathrm{ML}}$ and show that it is independent of the variance σ^2. [2 points]

(*d*) Show that the maximum likelihood solution for σ^2 is given by:

$$\sigma^2_{\mathrm{ML}} = \frac{1}{NK}||\mathbf{T} - \boldsymbol{\Phi}\mathbf{W}_{\mathbf{ML}}||_F^2$$

[2 points]

Machine Learning 1 - HW1 – 2025 – Paper 1

4 Counting Fish (4 points)

Imagine you are in an aquarium, and you notice that each fish has a tag with a number attached to its fin. You see the numbers 8, 2, 14, 5, and then 14 again. 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 five samples? [1 point]

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

Hint: To find the "most likely" estimate, you first need a likelihood function. This requires assuming a probability distribution for the observed tags. If you have no reason to believe one fish tag is more likely to be seen than another, which distribution should you use?

(c) The MLE estimator you derived in the previous step is the maximum of the observed samples. Let's denote this estimator by $\widehat{M} = X_{\mathrm{max}}$. It is known to underestimate the true value. An adjusted estimator that attempts to correct this is $\widehat{M}_{adj} = \frac{n+1}{n} X_{max} - 1$. Compute the bias of this adjusted estimator. [2 points]

Hint: The bias is $\mathbb{E}[\widehat{M}_{adj}] - M$. Use the linearity of expectation to find $\mathbb{E}[\widehat{M}_{adj}]$. You may use the following known result for the expected value of the maximum of n samples from a discrete uniform distribution $\{1, \dots, M\}$ without proof:

$$\mathbb{E}[X_{max}] = M - \frac{1}{M^n} \sum_{x=0}^{M-1} x^n$$

Fun fact: The Allies used a similar statistical technique during WWII to estimate the monthly production rate of German tanks based on the serial numbers of captured or destroyed tanks. This estimation was significantly more accurate than intelligence reports.