Problem Set 1

Problem 1

Suppose we have n independent and identically-distributed samples $x_1, ..., x_n$ from a Gaussian distribution with unknown mean μ and known variance σ^2. The log-likelihood function for this data is

$$\begin{aligned}
\mathcal{L}(\mathbf{x}; \sigma^2, \mu) &= \sum_{i=1}^{n} \log p(x_i; \sigma^2, \mu) \\
&= \sum_{i=1}^{n} \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x_i - \mu)^2}{2\sigma^2}\right)\right) \\
&= \sum_{i=1}^{n} \left(-\frac{1}{2}\log(2\pi\sigma^2) - \frac{(x_i - \mu)^2}{2\sigma^2}\right) \\
&= -\frac{n}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i - \mu)^2.
\end{aligned}$$

Compute the gradient of this function $\nabla\mathcal{L}(\mathbf{x}; \sigma^2, \mu)$ with respect to the parameters μ and σ^2.

Problem 2

Suppose that we have data points $x_1, x_2, \ldots, x_n$ sampled independent and identically-distributed from a Gaussian distribution with unknown mean μ and unknown variance σ^2. Symbolically, for each i, $x_i \sim \mathcal{N}(\mu, \sigma^2)$.

Calculate the maximum-likelihood estimates $\hat{\mu}$ and $\hat{\sigma}^2$ of the parameters μ and σ^2 by solving the equation $\nabla \ell(\mu, \sigma^2) = \mathbf{0}$. Justify your calculations.

Problem 3

Derive the maximum-likelihood estimates $\hat{w}_0$ and $\hat{w}_1$ for the 1-dimensional linear-Gaussian model

$$y_i \sim \mathcal{N}(w_1 x_i + w_0, \sigma^2)$$

by setting the gradient to zero, and solving for the parameters. You may find it useful to recall the relationship between the log-likelihood and the mean-squared error loss.

Problem 4

Recall that, if $f : \mathbb{R} \to \mathbb{R}$ is a differentiable function, then x^* is a *critical point* of f if $\frac{df}{dx}(x^*) = f'(x^*) = 0$.

> **Definition 0.1**: A function $g : \mathbb{R} \to \mathbb{R}$ is *monotonic increasing* on an interval $I \subseteq \mathbb{R}$ if for all $x_1, x_2 \in I$ such that $x_1 < x_2$, we have $g(x_1) \leq g(x_2)$. Function g is *strictly monotonic increasing* if $g(x_1) < g(x_2)$ for all such x_1, x_2.

An important mathematical property we'll use in class is the following theorem:

> **Theorem 0.1**: Let $f : \mathbb{R} \to \mathbb{R}$ be a differentiable function such that $f(x) > 0$ for all $x \in \mathbb{R}$. Then, x^* is a critical point of f **if and only if** x^* is also a critical point of the function $h(x) = \log f(x)$.

(You may assume that the log is base e, which is sometimes also written $\ln f(x)$.)

Informally, this theorem says that if we want to find a critical point of f, it's ok to take the logarithm of f first and then find the critical points of $\log f$ instead. In many machine learning contexts it's much easier to work with $\log f$.

Part A

Prove that the function $k(x) = \log x$ is strictly monotonic increasing on the interval $(0, \infty)$. It's sufficient to evaluate the derivative of k and apply the mean value theorem.

Part B

Use Part A to prove Theorem 0.1. It's sufficient to calculate $\frac{dh}{dx}$ in terms of $\frac{df}{dx}$ (use the chain rule!). Can it be true that one of $\frac{df}{dx}(x^*)$ or $\frac{dh}{dx}(x^*)$ is zero while the other is non-zero?