# CS689: Machine Learning - Spring 2023

## Homework 1: Part 1 Solutions

**1.** (*5 points*) Consider the linear regression prediction function $f_\theta(\mathbf{x}) = \mathbf{x}\mathbf{w} + b$ with $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathbb{R}$. What is the computational complexity of computing the risk when the prediction loss function is the squared loss and the data set has $N$ data cases? Explain your answer.

**Example Solution:** The risk for this problem is defined as $R(f_\theta, \mathcal{D}) = \frac{1}{N}\sum_{n=1}^{N}(y_n - f_\theta(x_n))^2$ where $f_\theta(\mathbf{x}) = \mathbf{x}\mathbf{w} + b$. We have $\mathbf{x} \in \mathbb{R}^D$ and $y \in \mathbb{R}$. Computing the vector product $\mathbf{x}_n\mathbf{w}$ takes $O(D)$ time since $\mathbf{x}_n$ and $\mathbf{w}$ are $D$-dimensional vectors. Computing the addition for $\mathbf{x}_n\mathbf{w} + b$ takes constant time since both $\mathbf{x}_n\mathbf{w}$ and $b$ are scalars. Computing the subtraction for $y_n - (\mathbf{x}_n\mathbf{w} + b)$ takes constant time since both $y_n$ and $\mathbf{x}_n\mathbf{w} + b$ are scalars. Computing the square $(y_n - (\mathbf{x}_n\mathbf{w} + b))^2$ takes constant time since $y_n - (\mathbf{x}_n\mathbf{w} + b)$ is a scalar. Thus, the computation of the loss for each data case $n$ takes $O(D)$ time and the time to compute the loss for all $N$ data cases is $O(N \cdot D)$. Averaging the loss values over the $N$ data cases requires $N$ additional operations, so the total computational complexity is $O(N \cdot D)$.

**2.** (*5 points*) Suppose we have a regression task where $x \in \mathbb{R}$ and $y \in \mathbb{R}$ and we expect that the optimal prediction function is a polynomial of order 3 or lower. Provide a definition for a prediction function model $\mathcal{F}$ matching these assumptions. Explain your answer.

**Example Solution:** $\mathcal{F}$ must be a set containing all polynomial functions $f_\theta : \mathbb{R} \to \mathbb{R}$ of order at-most 3. We can write a polynomial of order three as $ax^3 + bx^2 + cx + d$. In general, the parameter values $a$, $b$, $c$, $d$ can take any real values. We obtain polynomials of second degree by setting $a = 0$, polynomials of first order by setting $a = 0$ and $b = 0$, and so on. The prediction function model expressing all polynomials of order up to three is thus: $\mathcal{F} = \{f_\theta(x) = ax^3 + bx^2 + cx + d | \theta \in \mathbb{R}^4\}$ where $\theta = [a, b, c, d]$.

**3.** (*10 points*) Suppose we have a regression task where $\mathbf{x} \in \mathbb{R}^D$. Suppose the true data generating distribution satisfies $\mathbb{E}_{p_*(\mathbf{X}=\mathbf{x})}[\mathbf{x}] = \mu_*$ for some $\mu_* \in \mathbb{R}^D$. Prove that the statistic $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ is an unbiased estimator of $\mu_*$ when the inputs $\mathbf{x}_1, ..., \mathbf{x}_N$ are sampled from $p_*$.

**Example Solution:** We are given that $\mathbb{E}_{p_*(\mathbf{X}=\mathbf{x})}[\mathbf{x}] = \mu_*$ where $\mathbf{x}$ is a single data vector sampled from $p_*$ and $\mathbf{X}$ is the corresponding random variable. We will let $\mathcal{D}_x = \{\mathbf{x}_1, ..., \mathbf{x}_N\}$. Under the IID assumption, we have that $p_*(\mathcal{D}_x) = \prod_{n=1}^{N} p_*(\mathbf{X}_n = \mathbf{x}_n)$. We need to prove that $\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n$ is an unbiased estimator of $\mu_*$ under the assumption that $\mathbb{E}_{p_*(\mathbf{X}=\mathbf{x})}[\mathbf{x}] = \mu_*$. To prove that the estimator is unbiased assuming the data in $\mathcal{D}_x$ are sampled IID from $p_*$, we need to show that $\mathbb{E}_{p_*(\mathcal{D})}[\hat{\mu}] = \mu_*$.

To begin, we will show that $\mathbb{E}_{p_*(\mathcal{D}_x)}[\mathbf{x}_n] = \mu_*$ for all $n$. To do so, consider the data set $\mathcal{D}_x^{-n}$ that contains all of the data cases except for $\mathbf{x}_n$. We note that:

$$p_*(\mathcal{D}_x^{-n}|\mathbf{x}_n) = p_*(\mathcal{D}_x)/p_*(\mathbf{X}_n = \mathbf{x}_n) = \prod_{n' \neq n} p_*(\mathbf{X}_{n'} = \mathbf{x}_{n'}) = p_*(\mathcal{D}_x^{-n})$$

Now consider the following decomposition of the joint expectation over the data set:

$$\mathbb{E}_{p_*(\mathcal{D}_x)}[\mathbf{x}_n] = \mathbb{E}_{p_*(\mathbf{x}_n)}[\mathbb{E}_{p_*(\mathcal{D}_x^{-n}|\mathbf{x}_n)}[\mathbf{x}_n]] \tag{1}$$

$$= \mathbb{E}_{p_*(\mathbf{x}_n)}[\mathbb{E}_{p_*(\mathcal{D}_x^{-n})}[\mathbf{x}_n]] \tag{2}$$

Since $\mathbf{x}_n$ is constant with respect to the variables in $\mathcal{D}_x^{-n}$ (which does not contain $\mathbf{x}_n$ by definition), we have that $\mathbb{E}_{p_*(\mathcal{D}_x^{-n})}[\mathbf{x}_n] = \mathbf{x}_n$. This means that $\mathbb{E}_{p_*(\mathcal{D}_x)}[\mathbf{x}_n] = \mathbb{E}_{p_*(\mathbf{x}_n)}[\mathbf{x}_n]$. By the expectation assumption stated in the question, we have $\mathbb{E}_{p_*(\mathbf{x}_n)}[\mathbf{x}_n] = \mu_*$. We use linearity of expectation combined with this result to complete the proof:

$$\mathbb{E}_{p_*(\mathcal{D})}[\hat{\mu}] = \mathbb{E}_{p_*(\mathcal{D}_x)}\left[\frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n\right] \tag{3}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{p_*(\mathcal{D}_x)}\left[\mathbf{x}_n\right] \tag{4}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}_{p_*(\mathbf{x}_n)}\left[\mathbf{x}_n\right] \tag{5}$$

$$= \frac{1}{N}\sum_{n=1}^{N}\mu_* = \mu_* \tag{6}$$

**4.** (*10 points*) Consider a regression task where $\mathcal{X} = \mathbb{R}^D$. Suppose that for all data cases, dimension $j$ is a scaled copy of dimension $i$. In other words, for some $a \neq 0$ and for all $n$, we have $\mathbf{x}_{ni} = a \cdot \mathbf{x}_{nj}$. Explain why the standard OLS estimator for the linear regression model $\hat{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ can not be used to find the optimal parameters of the linear regression model under squared prediction loss in this case.

**Example Solution:** Consider the data matrix $\mathbf{X}$ where each row is a data case with bias absorption applied as in the OLS estimator derivation. This matrix has size $N \times (D+1)$. Since there is a non-zero scalar $a$ such that $\forall n$ the condition $\mathbf{x}_{ni} = a \cdot \mathbf{x}_{nj}$ holds, it will be the case that column $j$ of $\mathbf{X}$ is a scaled copy of column $i$. This means that there is at least one column of $\mathbf{X}$ that is not linearly independent of the other columns in $\mathbf{X}$ so the column rank of $\mathbf{X}$ is at most $D$ and thus the rank of $\mathbf{X}$ is at most $D$.

Now consider the term $\mathbf{X}^T\mathbf{X}$. Using the result that the rank of a product of two matrices $AB$ is less than or equal to the the minimum of the rank of $A$ and the rank of $B$, we have that the rank of $\mathbf{X}^T\mathbf{X}$ is at most $D$. Since $\mathbf{X}^T\mathbf{X}$ is a $(D+1) \times (D+1)$ matrix with rank at most $D$, by the invertible matrix theorem, the matrix $\mathbf{X}^T\mathbf{X}$ is not invertible and thus the standard OLS formula can not be used to find the optimal model parameters.

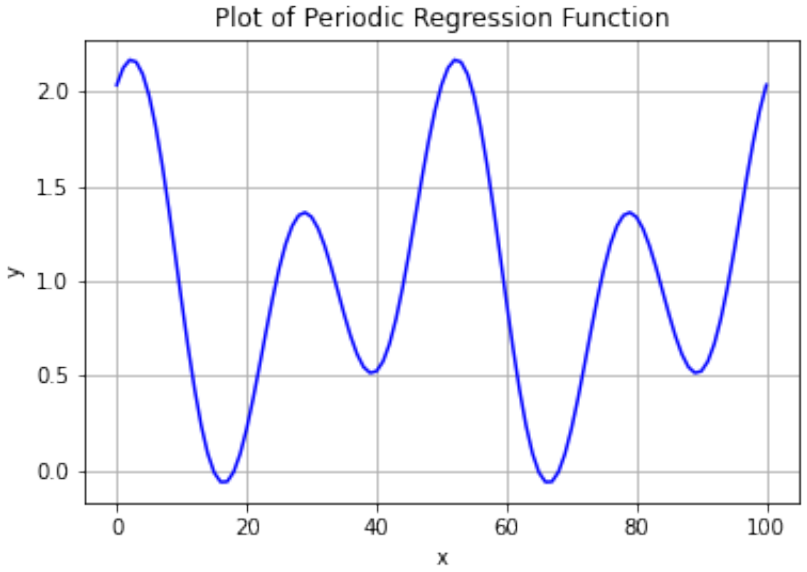
**5.** (*20 points*) Consider the problem of building a regression model for periodic time series data. In this problem, $y \in \mathbb{R}$ and $x \in \mathbb{R}$. One way to model such data is with a regression function built from a sum of $K$ cosine-based components. Each component $k$ has an amplitude $w_k$, a phase $\phi_k$ and a period $\rho_k$. We also include a bias parameter $b$. The form of the prediction function is shown below:

$$f_\theta(x) = b + \sum_{k=1}^{K} w_k \cos\left(\frac{2\pi}{\rho_k} x - \phi_k\right)$$

In this prediction function, the parameters are $\theta = [\mathbf{w}, \phi, b]$ where $\mathbf{w} = [w_1, ..., w_K]$ and $\phi = [\phi_1, ..., \phi_K]$. The number of periodic components $K$ and their periods $\rho_k$ are fixed.

**a.** (*5 pts*) Consider the case where $K = 2$, $\rho = [50, 25]$ and $\theta = [0.5, 0.75, -0.25, 0.75, 1.0]$. Plot $f_\theta(x)$ from 0 to 100 using these parameters.

**Example Solution:**




**b.** (*5 pts*) Suppose we have a data set $\mathcal{D}$ containing just two observations $(20, 0)$ and $(40, 2.5)$. Compute the empirical risk for this data set using $K = 2$, $\rho = [50, 25]$ and $\theta = [0.5, 0.75, -0.25, 0.75, 1.0]$.

**Example Solution:** The empirical risk for this data set using $K = 2$, $\rho = [50, 25]$ and $\theta = [0.5, 0.75, -0.25, 0.75, 1.0]$ is $1.9765$.

**c.** (*5 pts*) Derive the gradient of the empirical risk function for this model under the squared prediction loss and assuming $K = 2$. Clearly indicate the components that correspond to $\mathbf{w}$, $\phi$ and $b$. Show your work.

**Example Solution:** The empirical risk function under the squared prediction loss is $R(f, \mathcal{D}) = \frac{1}{N} \sum_{n=1}^{N} (y_n - f_\theta(x_n))^2$. We begin to find the gradient of the empirical risk function under the squared prediction loss as follows

$$\nabla R(f, \mathcal{D}) = \nabla \frac{1}{N} \sum_{n=1}^{N} (y_n - f_\theta(x_n))^2 \tag{7}$$

$$= \frac{1}{N} \sum_{n=1}^{N} \nabla (y_n - f_\theta(x_n))^2 \qquad \text{(moving the gradient inside the summation)} \tag{8}$$

$$= \frac{1}{N} \sum_{n=1}^{N} 2(y_n - f_\theta(x_n))(-1) \nabla f_\theta(x_n) \qquad \text{(using the chain rule)} \tag{9}$$

$$= 2\frac{1}{N} \sum_{n=1}^{N} (f_\theta(x_n) - y_n) \nabla f_\theta(x_n) \qquad \text{(simplifying)} \tag{10}$$

Now we would like to get the individual components of $\nabla f_\theta(x_n)$ corresponding to $\mathbf{w}$, $\phi$, and $b$ to substitute into the expression above. We first find the partial derivative of $f_\theta(x_n)$ with respect to $w_i$:

$$\frac{\partial}{\partial w_i} f_\theta(x_n) = \frac{\partial}{\partial w_i} \left( b + \sum_{k=1}^{K} w_k \cos\left( \frac{2\pi}{\rho_k} x - \phi_k \right) \right) \tag{11}$$

$$= \frac{\partial}{\partial w_i} \left( w_i \cos\left( \frac{2\pi}{\rho_i} x - \phi_i \right) \right) \tag{12}$$

$$= \cos\left( \frac{2\pi}{\rho_i} x - \phi_i \right) \tag{13}$$

Next, we find the partial derivative of $f_\theta(x_n)$ with respect to $\phi_i$.

$$\frac{\partial}{\partial \phi_i} f_\theta(x_n) = \frac{\partial}{\partial \phi_i} \left( b + \sum_{k=1}^{K} w_k \cos\left( \frac{2\pi}{\rho_k} x - \phi_k \right) \right) \tag{14}$$

$$= \frac{\partial}{\partial \phi_i} \left( w_i \cos\left( \frac{2\pi}{\rho_i} x - \phi_i \right) \right) \tag{15}$$

$$= w_i (-1) \sin\left( \frac{2\pi}{\rho_i} x - \phi_i \right) (-1) \tag{16}$$

$$= w_i \sin\left( \frac{2\pi}{\rho_i} x - \phi_i \right) \tag{17}$$

Finally, we find the partial derivative of the gradient of $f_\theta(x_n)$ with respect to $b$.

$$\frac{\partial}{\partial b} f_\theta(x_n) = \frac{\partial}{\partial b} \left( b + \sum_{k=1}^{K} w_k \cos\left( \frac{2\pi}{\rho_k} x - \phi_k \right) \right) \tag{18}$$

$$= \frac{\partial}{\partial b} b = 1 \tag{19}$$

This yields a final gradient vector for the case of $K$ components that is structured as follows:

$$\nabla R(f, \mathcal{D}) = \frac{2}{N} \sum_{n=1}^{N} (f_\theta(x_n) - y_n) \left[ \frac{\partial}{\partial w_1} f_\theta(x_n), ..., \frac{\partial}{\partial w_K} f_\theta(x_n), \frac{\partial}{\partial \phi_1} f_\theta(x_n), ..., \frac{\partial}{\partial \phi_1} f_\theta(x_n), 1 \right]^T$$

**d.** (*5 pts*)Suppose we have a data set $\mathcal{D}$ containing just two observations $(20, 0)$ and $(40, 2.5)$. Compute the gradient of the empirical risk for this data set using $K = 2$, $\rho = [50, 25]$ and $\theta = [0.5, 0.75, -0.25, 0.75, 1.0]$.Clearly indicate the components that correspond to $\mathbf{w}$, $\phi$ and $b$.

**Example Solution:** The gradient of the empirical risk for this data set using $K = 2$, $\rho = [50, 25]$ and $\theta = [0.5, 0.75, -0.25, 0.75, 1.0]$ is $[-1.26009, 1.86893, 0.87534, -0.32878, -1.757]$ where $[-1.26009, 1.86893]$ corresponds to $\mathbf{w}$, $[0.87534, -0.32878]$ corresponds to $\phi$, and $-1.757$ corresponds to $b$.