ISE - 364/464: Introduction to Machine Learning
Homework Assignment 7

The goal of this assignment is to provide a series of problems that solidify knowledge in specific topics in deep learning and unsupervised learning. Specifically, multi-layer perceptrons, K-means clustering, Gaussian mixture models, as well as connecting ridge regression to Bayesian linear regression.

Grading: This assignment is due on Coursesite by E.O.D. 12/13/2024. All problems are worth the same number of points. If a problem has multiple parts, each of those parts will be worth equal amounts and will sum to the total number of points of the original problem (Example: If each problem is worth a single point, and problem 1 has 4 parts, each part will be worth 1/4th of a point). ISE - 364 students are only required to answer problems 1 through 4; however, you are allowed to answer the 5th graduate-level question (if done so correctly, you will receive extra credit in the amount that the 5th problem will be worth for the ISE - 464 students). ISE - 464 students are required to answer all 5 problems.

Submitting: Only electronic submissions on Coursesite are accepted.

1 Problems

1. (MLP for Multi-Target Regression) Consider a multi-layer perception model $h_\theta(x)$ with some general number of layers $L > 1$ and predefined activation functions $g^{[\ell]}$, where the parameters $\theta := \{W, b\}$ denote the collection of weight matrices and bias vectors for each layer of the network, i.e., $W := \left\{W^{[\ell]}\right\}_{\ell=1}^{L}$ and $b := \left\{b^{[\ell]}\right\}_{\ell=1}^{L}$ where $W^{[\ell]} \in \mathbb{R}^{n[\ell] \times n[\ell-1]}$ and $b^{[\ell]} \in \mathbb{R}^{n[\ell]}$ for all $\ell \in \{1, 2, ..., L\}$. Further, this neural network is utilized to predict a multi-target continuous variable, i.e., this is a regression problem where the target variable is a vector $y \in \mathbb{R}^K$ where K is the number of targets. The squared error loss function is suitable for training this network, and, for a single feature-target pair (x, y), is given by

$$J(\theta) := \frac{1}{2} \left\| h_\theta(x) - y \right\|_2^2 .$$

a) (0.8 points) Assuming that the activation function of the final layer is simply the identity function, i.e., $g^{[L]}(x) = I(x) = x$, derive expressions for the four following partial derivatives:

$$\frac{\partial J}{\partial a^{[L]}}, \quad \frac{\partial J}{\partial z^{[L]}}, \quad \frac{\partial J}{\partial W^{[L]}}, \quad \frac{\partial J}{\partial b^{[L]}}.$$

Also, write the dimensions of the resulting partial derivatives.

b) (0.2 points) Suppose that we did not want to use the identity function in the final layer. Could one utilize the softmax function for the multi-target regression problem? Justify your answer.

2. (Visualizing Feature Maps of an MLP) Consider a general MLP $h_\theta(x)$ (i.e., a general number of layers $L > 1$ and activation functions g) where the parameters $\theta := \{W, b\}$ denote the collection of weight matrices and bias vectors for each layer of the network, i.e., $W := \left\{W^{[\ell]}\right\}_{\ell=1}^{L}$ and $b := \left\{b^{[\ell]}\right\}_{\ell=1}^{L}$ where $W^{[\ell]} \in \mathbb{R}^{n[\ell] \times n[\ell-1]}$ and $b^{[\ell]} \in \mathbb{R}^{n[\ell]}$ for all $\ell \in \{1, 2, ..., L\}$. Derive the partial derivative $\frac{\partial z_c^{[L]}}{\partial a^{[0]}}$, where $z_c^{[L]} = \left(W^{[L]}\right)_c^\top a^{[L-1]} + b^{[L]}$, $\left(W^{[L]}\right)_c^\top$ is the c-th row of the weight matrix $W^{[L]}$, and where $a^{[0]}$ is the input vector.

Hint: To do this, you will need to first derive an expression for $\frac{\partial z_c^{[L]}}{\partial z^{[L-1]}}$ and then a general expression for $\frac{\partial z_c^{[L]}}{\partial z^{[\ell]}}$ for all $1 \leq \ell \leq L$. Lastly, you should be able to define $\frac{\partial z_c^{[L]}}{\partial a^{[0]}}$ in terms of $\frac{\partial z_c^{[L]}}{\partial z^{[1]}}$.)

3. (**K-Means Clustering**) Recall that for some dataset $\mathcal{D} = \left\{x^{(i)}\right\}_{i=1}^{m}$, training a K-means clustering model with $K \in \mathbb{N}$ clusters amounts to solving the problem

$$\min_{r,\mu} J(r, \mu) := \frac{1}{2} \sum_{i=1}^{m} \sum_{k=1}^{K} r_{i,k} \left\| x^{(i)} - \mu_k \right\|_2^2,$$

where $r_{i,k} \in \{0, 1\}$ are the cluster assignment variables (with values of 1 if the datapoint $x^{(i)}$ is assigned to cluster k, and 0 otherwise) and $\mu_k \in \mathbb{R}^n$ are the cluster centroids, for all $i \in \{1, 2, ..., m\}$ and $k \in \{1, 2, ..., K\}$.

a) Assume that each cluster has at least one datapoint assigned to it (i.e., $r_{i,k} = 1$ for at least one datapoint $x^{(i)}$ for all $k \in \{1, 2, ..., K\}$). Prove that the K-means clustering loss function J is positive definite in the centroid variables μ_k when the cluster assignment variables $r_{i,k}$ are held constant for all $i \in \{1, 2, ..., m\}$ and $k \in \{1, 2, ..., K\}$.

b) Derive the optimal solution for the centroids μ_k^* when the cluster assignment variables $r_{i,k}$ are held constant for all $i \in \{1, 2, ..., m\}$ and $k \in \{1, 2, ..., K\}$.

4. (**Gaussian Mixture Models**) Consider a Gaussian Mixture Model with $K \in \mathbb{N}$ clusters. Then, the incomplete-data log-likelihood function for a single observed datapoint $x \in \mathbb{R}^n$ corresponding to this model is given by

$$\ell(\theta; x) := \log \left\{ \sum_{k=1}^{K} \phi_k \mathcal{N}\left(x | \mu_k, \Sigma_k\right) \right\},$$

where the parameters are $\theta := \{\phi_1, \phi_2, ..., \phi_K, \mu_1, \mu_2, ..., \mu_K, \Sigma_1, \Sigma_2, ..., \Sigma_K\}$, where $\phi_k \in (0, 1)$ denotes the probability of the datapoint coming from the k-th Gaussian cluster, and where $\mu_k \in \mathbb{R}^n$ and $\Sigma_k \in \mathbb{R}^{n \times n}$ denote the mean and covariance matrix corresponding to the k-th Gaussian cluster, respectively, for all $k \in \{1, 2, ..., K\}$. While training a GMM when using the EM algorithm, one will need to compute the partial derivative of this log-likelihood with respect to μ_k. Prove that the partial derivative is

$$\frac{\partial \ell}{\partial \mu_k} := \frac{\phi_k \mathcal{N}\left(x | \mu_k, \Sigma_k\right)}{\sum_{j=1}^{K} \phi_j \mathcal{N}\left(x | \mu_j, \Sigma_j\right)} \Sigma_k^{-1} \left(x - \mu_k\right).$$

5. **(ISE-464 Graduate Students) (Bayesian and Ridge Regression)** Bayesian Linear Regression is a way of modeling a linear regression model $h_\theta(x) = \theta^\top x$ (with θ and x in $\mathbb{R}^n$, and the first entry of x is set to 1 to indicate the intercept) where the parameters are modeled as a multivariate Gaussian with mean 0 and covariance matrix $\tau^2 I$ (i.e., $\theta \sim \mathcal{N}\left(0, \tau^2 I\right)$ for some $\tau \in \mathbb{R}$ and where $I \in \mathbb{R}^{n\times n}$ denotes the identity matrix). Further, the target variable $y^{(i)} \in \mathbb{R}$ is still modeled as a Gaussian random variable with mean $\theta^\top x^{(i)}$ and variance σ (i.e., $y^{(i)} \sim \mathcal{N}\left(\theta^\top x^{(i)}, \sigma\right)$). Utilizing Bayes Theorem, one can formulate the Maximum A Posterior (MAP) estimate of the Bayesian linear regression model, denoted as θ^*_{MAP}, that maximizes the posterior distribution, as the problem

$$\theta^*_{MAP} := \underset{\theta}{\operatorname{argmax}}\, \mathbb{P}\left(\theta | \left\{\left(x^{(i)}, y^{(i)}\right)\right\}_{i=1}^m ; \tau, \sigma\right) = \underset{\theta}{\operatorname{argmax}}\, \mathbb{P}\left(\theta; \tau\right) \prod_{i=1}^m \mathbb{P}\left(y^{(i)} | x^{(i)}; \theta, \sigma\right),$$

where the prior distribution $\mathbb{P}\left(\theta; \tau\right)$ is given by

$$\mathbb{P}\left(\theta; \tau\right) = \frac{1}{(2\pi)^{n/2} |\tau^2 I|^{1/2}} e^{-\frac{1}{2}\theta^\top (\tau^2 I)^{-1}\theta},$$

and the posterior distribution $\mathbb{P}\left(y^{(i)} | x^{(i)}; \theta, \sigma\right)$ is given by

$$\mathbb{P}\left(y^{(i)} | x^{(i)}; \theta, \sigma\right) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2\sigma^2}(y^{(i)} - \theta^\top x^{(i)})^2}.$$

Prove that θ^*_{MAP} leads to the ℓ_2-regularized linear regression problem (ridge regression), i.e., prove that

$$\theta^*_{MAP} = \underset{\theta}{\operatorname{argmin}}\, \frac{1}{2} \sum_{i=1}^m \left(\theta^\top x^{(i)} - y^{(i)}\right)^2 + \frac{\lambda}{2} \|\theta\|_2^2,$$

for some $\lambda > 0$.