ELEG 5760 Machine Learning for Singal Processing Applications
Homework 1

General Guidelines:

- Please check the submission deadline on **Blackboard** and submit your solutions via **Blackboard**.
- Each homework's deadline has a grace period of 2 hours.
- Each student has one chance of late submission within 12 hours of the deadline.
- All other late submissions will be given 0 points with no exception.
- **Do not** close your browser of app before you have successfully uploaded your files. It is your own responsibility of keeping your file integrity.
- Round your results to 3 decimal places or keep the fractional numbers.

1. (30 points) Given a training dataset with two data samples, $x^{(1)} = [2,3]^T, y^{(1)} = 2, x^{(2)} = [-1,-1]^T, y^{(2)} = 3$, for softmax classification of 3 classes. Note that we append a constant $x_0^{(i)} = 1$ at the begining of each feature vector. Current parameters of the three classes are $\Theta_1 = [5,4,-2]^T, \Theta_2 = [-3,2,3]^T, \Theta_3 = [4,-2,3]^T$. We adopt a normalized version of the cross-entropy cost function (with a normalization factor) in this question.

$$L(\Theta) = -\frac{1}{m} \sum_{i=1}^{m} \sum_{l=1}^{k} \mathbf{1}(y^{(i)} = l) \log \left(\frac{\exp(\Theta_l^T x^{(i)})}{\sum_{j=1}^{k} \exp(\Theta_j^T x^{(i)})} \right)$$

   (a) (10 points) Find the cross-entropy cost function value for this training set.

   (b) (10 points) Find **negative** gradients of the parameters $\Theta_1, \Theta_2, \Theta_3$ with the training set.

   (c) (10 points) If a learning rate 0.1 is used, what would be parameters $\Theta_1, \Theta_2, \Theta_3$ after 1 iteration of gradient descent?

2. (20 points) In the lecture of logistic classification, we have stated that, when using the MSE loss for logistic classification, the gradient of θ_j of the classification hypothesis $g(\theta^T x)$ is calculated as

$$\frac{\partial}{\partial \theta_j} \text{MSE}(\theta) = (h(x) - y) x_j g'\left(\theta^T x\right),$$

   where g denotes the sigmoid function and g' denotes the derivative of the sigmoid function. Please show that how the formula is obtained.

3. (30 points) Design feature transformation of the data points (samples' feature vectors) to try to make the data linear separable in the transformed feature space.

   (a) (10 points) Consider the following 1D data points (1-dimensional feature vectors). Can you find a 1D feature transformation, i.e., $\phi : \mathbb{R} \to \mathbb{R}$, to make the data points linearly separable in the transformed feature space

Figure 1: Q1(a)

   (b) (10 points) You may not always need to map the data points into higher dimensional space to make the data linearly separable. Consider the following 2D data points (2-dimensional feature vectors). Can you find a 1D feature transformation, i.e., $\phi : \mathbb{R}^2 \to \mathbb{R}$, to make the data points linearly separable in the transformed feature space.

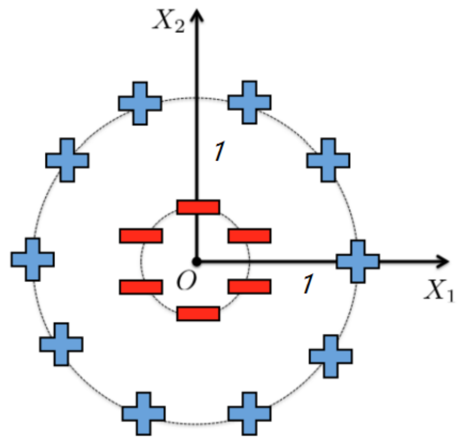

Figure 2: Q1(b)

(c) (10 points) Consider the following 2D data points (2-dimensional feature vectors). Can you find a 1D feature transformation, i.e., $\phi : \mathbb{R}^2 \to \mathbb{R}$, to make the data points linearly separable in the transformed feature space.

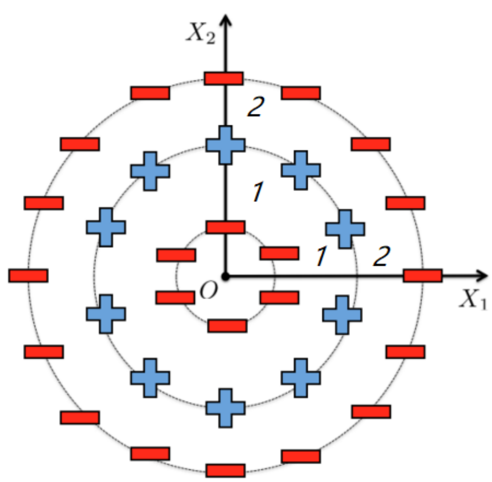

Figure 3: Q1(c)

4. (20 points) Prove or disprove the following functions are valid kernel functions. (Hint: You can try to find the corresponding feature transformation function to prove a case. You can provide a counter example to disprove a case.)

$$k(x, z) = (x^T z + 1)^2$$
$$k(x, z) = (x^T z - 1)^3$$