# Machine Learning (AI511)

## Assignment-1 (Total Marks: 10)

## (Due Date: 17-Sept-2023)

---

**Instructions:**

1. Please make sure that your solution is written in your words.
2. Please feel free to ask any questions in class or via LMS.
3. For Coding Problems:
   (a) The same dataset (football.csv) will be used for all the coding problems.
   (b) Each student has to submit a zip containing separate jupyter notebooks. The notebooks should contain all your observations, results, approaches, etc (whatever you feel necessary for us to evaluate your solutions/code/understanding). Name your zip as ¡RollNo¿.zip. Name all notebooks as Qx.ipynb
   (c) Your understanding and approach will be evaluated along with your results. *There will be a short viva for the assignment.*

---

**Theoretical Problems:**

1. Let $\mathbf{x}$ and $\mathbf{z}$ be two independent random vectors with a joint pdf $p(\mathbf{x}, \mathbf{z})$. Show that the mean of their sum $\mathbf{y} = \mathbf{x} + \mathbf{z}$ is given by the sum of the means of each of the variable separately. Similarly, show that the covariance matrix of $\mathbf{y}$ is given by the sum of the covariance matrices of $\mathbf{x}$ and $\mathbf{z}$.

2. Consider a sequence of $N$ coin tosses. The observation for the $n$th toss is $y_n$. Find the maximum likelihood (ML) estimate of probability of head $\theta$. If you believe that the coin is biased, how would you model your belief while learning $\theta$. In particular, how would you model if i) the coin is biased; ii) coin is fair; and iii) you have no prior knowledge about $\theta$. (*Hint:* Being binary, the observation $y_n$ can be modelled using the Binomial distribution. Frame the problem such that your prior belief and the likelihood form conjugate pair. Play with parameters of the model for your prior belief to realise the above three cases.)

3. Consider a linear model of the form

$$y_n = w_0 + \sum_{i=1}^{D} w_i x_{n,i}$$

together with a sum-of-squares error function of the form

$$E_D(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n - t_n)^2 .$$

Now suppose that Gaussian noise $\epsilon_{n,i} \sim \mathcal{N}(0, \sigma^2)$ is added independently to each of the input variables $x_{n,i}$. By making use of $\mathbb{E}[\epsilon_{n,i}] = 0$ and $\mathbf{E}[\epsilon_{n,i}\epsilon_{n,j}] = \delta_{i,j}\sigma^2$ with $\mathbb{E}[\cdot]$ being the expectation operation and $\delta_{i,j} = 1$ for $i = j$ and 0 otherwise, show that minimizing $E_D$ averaged over the noise distribution is equivalent to minimizing the sum-of-squares error for noise-free input variables with the addition of a weight-decay regularization term, in which the bias parameter $w_0$ is omitted from the regularizer.

4. For the generative data model consider in the class: $\mathbf{t} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$ with the elements of $\boldsymbol{\epsilon}$ being IID with distribution function $\mathcal{N}(0, \sigma^2)$.

   (a) Give the expression of generating distribution.

   (b) Use the above distribution to derive the correlation matrix $\mathbb{E}[\mathbf{t}\mathbf{t}^T]$.

5. **Sequential Learning:** Consider data consisting of $N$ scaler-valued observations $x_1, x_2, \ldots, x_N$. Assume each observation is drawn independently from a Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$. Show that the maximum likelihood (ML) solution for $\sigma^2$ is obtained as

$$\sigma^2_{(N)} = \sigma^2_{(N-1)} + \frac{1}{N}\left\{(x_N - \mu)^2 - \sigma^2_{(N-1)}\right\}, \tag{1}$$

where

$$\sigma^2_{(N-1)} = \frac{1}{N-1}\sum_{n=1}^{N}(x_N - \mu)^2. \tag{2}$$

This result has a nice interpretation, as follows. After observing $N-1$ data points, we have estimated $\sigma^2$ by $\sigma^2_{(N-1)}$. We now observe data point $x_N$, and we obtain our revised estimate $\sigma^2_{(N)}$ by moving the old estimate a small amount, proportional to $1/N$, in the direction of the 'error signal' $\frac{1}{N}\left\{(x_N - \mu)^2 - \sigma^2_{(N-1)}\right\}$. Note that, as $N$ increases, the contribution from successive data points gets smaller.

**Coding Problems:**

1. **Linear regression and Regularization:**

(a) Predict the "Overall" (target attribute: "overall") rating of the players using Linear regression report the Mean Absolute Error(MAE), Mean Square Error(MSE), R2 score.

(b) Compare the performance of linear regression, Ridge regression, and Lasso regression models. Perform the hyperparameters tuning and observe how they affect the model's bias-variance trade-off, investigate the impact of the Lasso regularisation parameter on this feature selection process.

Note: Please carry out the necessary data preprocessing and test-train split as $20:80\%$. The use of the scikit-learn library is allowed for this question. For (b), include necessary metrics like MSE, MAE, R2 Score for performance analysis and necessary plots (Ex: Scatter plots/line plots) for hyperparameters tuning.

2. **Logistic Regression:** You are given a dataset named football.csv containing information about football players. Your task is to build a machine learning model to classify whether a player's contribution type is more inclined towards being type 1 or 0, where 1 indicates players with contributions in the attacking half of the football field and 0 indicates players with contributions in the defending half of the field. The classification column is "contribution_type".

   (a) Train a Logistic Regression model using the training data. Implement logistic regression from scratch. You're NOT ALLOWED to use sklearn for this question.

   (b) Make predictions on the test data using the trained model.

   (c) Calculate the F1 score, accuracy score, and confusion matrix to evaluate the model's performance.

3. **K-Means - Clustering of Football Clubs:** You are given a dataset containing football player information. The objective is to cluster different football clubs based on various attributes.

   (a) Your initial step should be extracting the club information from the player dataset. It will involve computing the "average player" of each club. One way is to group the dataset by "club_name_id" and calculate the mean values for all relevant features. You are encouraged to explore different ideas.

   (b) Now, use K-Means to cluster the football clubs. You can determine the criteria for clustering by considering various features. For example - First try clustering using all

features, after that try again using only features representing player stats or financial attributes.

(c) To find the optimal number of clusters, K, you can use the elbow method.

4. **Random Forest:** For this question also, you have to use the same football dataset. The aim is to use the Random Forest model to do classification and regression both.

   (a) Classification: Your target column is 'contribution_type'. Process the data as you want, modify/drop any columns that you want, and play around with the hyperparameters. Try to understand and observe the difference in results. Try different losses (or quality criterion) – 'gini', 'entropy', 'log_loss'. After training the model, report test accuracy and f1 score.

   (b) Regression: Your target column is 'overall'. Again, you are free to process the dataset and encouraged to try different hyperparameters. Use MSE and MAE one by one to train the models, and report test MSE and MAE for both models.

   Note: You can use sklearn library to get Random Forest implementations.