**Midterm Project**

**Subject: Introduction to Machine Learning**

**Question 1 (5 points):**

(1) Present the steps in solving a classification problem or regression problem using machine learning methods.

(2) Perform the steps in question 1 on a classification or regression problem using different machine learning methods. Evaluate and compare the results of applying these methods;

(3) Find solutions to improve the accuracy of the models mentioned in question (2).

**Question 2 (5 points):**

(1) Let's study yourself and present what is the overfitting problem of machine learning models? Describe overfitting phenomena.

(2) Please present solutions to avoid overfitting, from general solutions to solutions for each specific machine learning method (including learned machine learning methods and you can be expanded to other learning methods).

(3) Present the problems mentioned in questions 1 and 2 through one or more real data sets (for classification and/and regression problems). Note that you must choose data sets and machine learning method where overfitting occurs.

Note:

* Do not use data sets that have been used in the classroom including: *iris, adult, abalone, house pricing, diabetes* data sets.
* All members have to submit solutions on the elearning system

Dự án Giữa kỳ

**môn: Nhập môn Học máy**

**Câu hỏi 1 (5đ):**

1. Trình bày về các bước trong việc giải quyết 1 bài toán classification hoặc bài toán regression sử dụng phương pháp học máy.
2. Thực hiện các bước trong câu 1 trên 1 bài toán classification hoặc regression sử dụng nhiều phương pháp học máy khác nhau. Đánh giá và So sánh kết quả áp dụng các phương pháp này;
3. Tìm cách cải tiến để nâng độ chính xác của các mô hình nêu ở câu (2).

**Câu hỏi 2 (5đ):**

1. Hãy tìm hiểu và trình bày về vấn đề overfiting của mô hình học máy là gì?
2. Hãy trình bày giải pháp để tránh overfiting, từ các giải pháp tổng quát đến các giải pháp cho từng phương pháp học máy cụ thể (bao gồm các phương pháp học máy đã học và có thể mở rộng ra các phương pháp học máy tự tìm hiểu).
3. Hãy trình bày các vấn đề nêu ở câu 1 và câu 2 thông qua một hoặc nhiều tập dữ liệu thật (cho bài toán classification hoặc/và regression). Lưu ý chọn tập dữ liệu và phương pháp học máy mà xảy ra hiện tượng overfiting.

Lưu ý:

* Không được sử dụng các bài toán, các tập dữ liệu đã được sử dụng trong qúa trình học trên lớp bao gồm: tập dữ liệu iris, adult, abalone, house pricing, diabetes.
* Tất cả các thành viên trong nhóm đều phải nộp bài.