ĐẠI HỌC QUỐC GIA TP.HCM

**TRƯỜNG ĐH KHOA HỌC TỰ NHIÊN**



**CHUYÊN ĐỀ NÂNG CAO LĨNH VỰC KHOA HỌC DỮ LIỆU**

**Tên khóa chuyên đề: MACHINE LEARNING FOR DATA SCIENCE**

(MÁY HỌC CHO KHDL)

**Thời lượng: 23** giờ

**Giảng viên:** ThS. Huỳnh Thanh Sơn

1. **Mục tiêu khóa học**

Trang bị các kiến thức nền tảng về Machine Learning và các thuật toán cơ bản thường dùng, để học viên có thể áp dụng vào công việc phân tích, dự đoán, phân cụm dữ liệu. Các kiến thức này sẽ là nền móng để học viên nghiên cứu tiếp các thuật toán nâng cao hơn nếu có nhu cầu.

1. **Kết quả đầu ra**

Nắm vững kiến thức, và vận dụng được các thuật toán Máy học thông dụng vào các bài toán Khoa học dữ liệu.

1. **Nội dung chuyên đề**
	* Các khái niệm cơ bản trong Máy học
	* Các kỹ thuật máy học cơ bản để phân tích, dự đoán (Linear Regression, Decision Tree, Clustering, Logistic Regression, Support Vector Machine (SVM) …)
	* Case studies: Churn Prediction, Customer Segmentation, …
	* Có bài thực hành sau các buổi học
	* Yêu cầu đầu vào: Học viên đã có nền tảng Python cơ bản trước đó
2. **Nội dung chi tiết chuyên đề**

| **Buổi** | **Module** | **Topic** |
| --- | --- | --- |
| **1** | Hồi Quy Tuyến Tính (Linear Regression) | * Giới thiệu Máy học trong Khoa học Dữ liệu. Khái lược hồi quy tuyến tính.
* Các khái niệm cơ sở trong huấn luyện mô hình:
* Cách mô hình học từng bước nhờ thuật toán Gradient descent, hàm chi phí (Cost function) và tốc độ học (Learning rate).
* Hiểu vì sao mô hình có thể học quá kỹ hay chưa tới với khái niệm quá khớp (Overfitting) và thiếu khớp (Underfitting)
* Biết cách chia dữ liệu thành tập huấn luyện (Training set), kiểm định (validation set) và kiểm tra (test set) để đánh giá hiệu suất của mô hình.
 |
| **2** | Hồi Quy Logistic (Logistic Regression)  | * Nắm được khái niệm Hồi quy Logistic từ đó hiểu tại sao đây là mô hình nền tảng cho các bài toán phân loại.
* Hiểu phân loại nhị phân (Binary Classification) và ranh giới quyết định, từ đó thấy cách mô hình phân biệt giữa hai nhóm dữ liệu.
* Hàm sigmoid, công cụ biến đầu ra của mô hình thành xác suất để ra quyết định.
* Làm quen với hàm chi phí và tối ưu hóa trong hồi quy Logistic từ đó biết cách đo lường sai số và cải thiện mô hình.
* Mở rộng từ mô hình phân loại nhị phân sang mô hình phân loại đa lớp (multi-class classification).
* Vai trò của chuẩn hóa đặc trưng (Feature Scaling) trong huấn luyện mô hình.
 |
| **3** | Naive Bayes Classification | * Giới thiệu phân loại Naive Bayes (Naive Bayes Classification).
* Lý thuyết xác suất Bayes
* Giả định độc lập giữa các biến
* Naive Bayes với phân phối Gaussian.
 |
| **4** | Thuật Toán Gom Cụm (Clustering) | * Giới thiệu Học không giám sát và thuật toán gom cụm (Clustering).
* Khái niệm về Within-cluster variation để đánh giá mức độ gắn kết của dữ liệu bên trong một cụm.
* Thuật toán K-means Clustering.
* Phương pháp Elbow để chọn số cụm phù hợp cho dữ liệu.
 |
| **5** | Mô hình cây quyết định và rừng cây (Decision Tree & Random Forest) | * Làm quen với mô hình cây quyết định (Decision Tree).
* Các tiêu chí đánh giá đặc trưng trong cây quyết định:
* Entropy.
* Information Gain.
* Gain Ratio.
* Thuật toán ID3 để xây dựng mô hình cây quyết định
* Mô hình rừng cây (Random Forest) mở rộng từ mô hình cây quyết định.
 |
| **6** |
| **7** | Support Vector Machine (SVM) | * Làm quen với mô hình Support Vector Machine, cho bài toán phân loại.
* Tìm hiểu bài toán tối ưu hóa phía sau mô hình SVM.
* So sánh hai cách xây dựng mô hình SVM (Soft margin và Hard margin SVM).
* Giới thiệu SVM kernel trick như một mở rộng SVM cho những bài toán phức tạp, phi tuyến.
 |
| **8** | Case studies |  |

#