

TRƯỜNG ĐẠI HỌC CNTT&TT
KHOA CÔNG NGHỆ THÔNG TIN

Đàm Thanh Phương
Hà Thị Thanh
Trần Quang Quý

BÀI GIẢNG
HỌC MÁY

NĂM HỌC 2022-2023

LƯU HÀNH NỘI BỘ

Mục lục

Chương I TỔNG QUAN

| | | |
|----------|--|----|
| 1 | Các khái niệm cơ bản | 16 |
| 1.1 | Khái niệm học máy | 16 |
| 1.2 | Dữ liệu | 17 |
| 1.3 | Các bài toán cơ bản trong học máy | 18 |
| 1.4 | Các thuật toán học máy | 20 |
| 1.5 | Hàm mất mát và tham số mô hình | 21 |
| 1.6 | Những thách thức chính của học máy | 22 |
| 1.7 | Quy trình thực hiện một dự án học máy | 24 |
| 1.8 | Câu hỏi ôn tập bài 1 | 29 |
| 2 | Các kỹ thuật xây dựng đặc trưng | 30 |
| 2.1 | Giới thiệu | 30 |
| 2.2 | Mô hình chung cho các bài toán học máy | 31 |
| 2.3 | Một số kỹ thuật trích chọn đặc trưng | 33 |
| 2.4 | Chuẩn hoá vector đặc trưng | 35 |
| 2.5 | Câu hỏi ôn tập bài 2 | 36 |

| | |
|-------------------------------|----|
| 2.6 BÀI TẬP CUỐI CHƯƠNG | 36 |
|-------------------------------|----|

Chương II CÁC THUẬT TOÁN HỌC CÓ GIÁM SÁT

| | |
|--|-----------|
| 3 Hồi quy tuyến tính | 40 |
| 3.1 Giới thiệu | 40 |
| 3.2 Xây dựng và tối ưu hàm mất mát | 41 |
| 3.3 Ví dụ trên Python | 43 |
| 3.4 Thảo luận | 46 |
| 3.5 Câu hỏi ôn tập bài 3 | 48 |
| 4 K lân cận | 49 |
| 4.1 Giới thiệu | 49 |
| 4.2 Phân tích toán học | 50 |
| 4.3 Ví dụ trên cơ sở dữ liệu Iris | 51 |
| 4.4 Thảo luận | 54 |
| 4.5 Câu hỏi ôn tập bài 4 | 56 |

| | | |
|----------|---|----|
| 5 | Bộ phân loại naive Bayes | 57 |
| 5.1 | Bộ phân loại naive Bayes | 57 |
| 5.2 | Các phân phối thường dùng trong NBC | 59 |
| 5.3 | Ví dụ | 60 |
| 5.4 | Thảo luận | 68 |
| 5.5 | Câu hỏi ôn tập bài 5 | 69 |
| 6 | Hạ Gradient | 70 |
| 6.1 | Giới thiệu | 70 |
| 6.2 | Hạ gradient cho hàm một biến | 71 |
| 6.3 | Hạ gradient cho hàm nhiều biến | 76 |
| 6.4 | Hạ gradient với momentum | 79 |
| 6.5 | Nesterov accelerated gradient | 82 |
| 6.6 | Hạ gradient ngẫu nhiên | 83 |
| 6.7 | Thảo luận | 85 |
| 6.8 | Câu hỏi ôn tập bài 6 | 86 |
| 7 | Thuật toán học perceptron | 88 |
| 7.1 | Giới thiệu | 88 |
| 7.2 | Thuật toán học perceptron | 89 |
| 7.3 | Ví dụ và minh hoạ trên Python | 92 |
| 7.4 | Mô hình mạng neuron đầu tiên | 93 |
| 7.5 | Thảo Luận | 95 |
| 7.6 | Câu hỏi ôn tập bài 7 | 97 |

| | |
|---|-----|
| 8 Hồi quy logistic | 98 |
| 8.1 Giới thiệu | 98 |
| 8.2 Hàm mất mát và phương pháp tối ưu | 100 |
| 8.3 Triển khai thuật toán trên Python | 103 |
| 8.4 Tính chất của hồi quy logistic | 106 |
| 8.5 Bài toán phân biệt hai chữ số viết tay | 108 |
| 8.6 Bài toán phân loại đa lớp | 109 |
| 8.7 Thảo luận | 111 |
| 8.8 Câu hỏi ôn tập bài 8 | 113 |
| 9 Hồi quy softmax | 115 |
| 9.1 Giới thiệu | 115 |
| 9.2 Hàm softmax | 116 |
| 9.3 Hàm mất mát và phương pháp tối ưu | 119 |
| 9.4 Ví dụ trên Python | 124 |
| 9.5 Thảo luận | 127 |
| 9.6 Câu hỏi ôn tập bài 9 | 127 |
| 10 Máy vector hỗ trợ | 129 |
| 10.1 Giới thiệu | 129 |
| 10.2 Xây dựng bài toán tối ưu cho máy vector hỗ trợ | 131 |
| 10.3 Bài toán đối ngẫu của máy vector hỗ trợ | 133 |
| 10.4 Lập trình tìm nghiệm cho máy vector hỗ trợ | 136 |
| 10.5 Tóm tắt | 138 |
| 10.6 Câu hỏi ôn tập bài 10 | 139 |

| | |
|--------------------------|-----|
| 10.7 BÀI TẬP CUỐI CHƯƠNG | 140 |
|--------------------------|-----|

Chương III CÁC THUẬT TOÁN HỌC KHÔNG GIÁM SÁT

| | |
|--|-----|
| 11 Phân cụm K-means | 144 |
| 11.1 Giới thiệu | 144 |
| 11.2 Phân tích toán học | 145 |
| 11.3 Ví dụ trên Python | 148 |
| 11.4 Phân cụm chữ số viết tay | 152 |
| 11.5 Tách vật thể trong ảnh | 155 |
| 11.6 Nén ảnh | 156 |
| 11.7 Thảo luận | 157 |
| 11.8 Câu hỏi ôn tập bài 11 | 160 |
| 12 Phân tích thành phần chính | 161 |
| 12.1 Phân tích thành phần chính | 161 |
| 12.2 Các bước thực hiện phân tích thành phần chính | 166 |
| 12.3 Liên hệ với phân tích giá trị suy biến | 167 |
| 12.4 Làm thế nào để chọn số chiều của dữ liệu mới | 169 |
| 12.5 Lưu ý về tính toán phân tích thành phần chính | 169 |
| 12.6 Một số ứng dụng | 170 |
| 12.7 Thảo luận | 174 |
| 12.8 Câu hỏi ôn tập bài 12 | 174 |
| 12.9 BÀI TẬP CUỐI CHƯƠNG | 175 |

Chương IV HỆ THỐNG GỢI Ý

| | |
|--|-----|
| 13 Hệ thống gợi ý dựa trên nội dung | 178 |
| 13.1 Giới thiệu | 178 |
| 13.2 Ma trận tiện ích | 179 |
| 13.3 Hệ thống dựa trên nội dung | 181 |
| 13.4 Bài toán MovieLens 100k | 184 |
| 13.5 Thảo luận | 188 |
| 13.6 Câu hỏi ôn tập bài 13 | 188 |
| 14 Lọc cộng tác lân cận | 190 |
| 14.1 Giới thiệu | 190 |
| 14.2 Lọc cộng tác theo người dùng | 191 |
| 14.3 Lọc cộng tác sản phẩm | 196 |
| 14.4 Lập trình trên Python | 198 |
| 14.5 Thảo luận | 201 |
| 14.6 Câu hỏi ôn tập bài 14 | 201 |
| 15 Lọc cộng tác phân tích ma trận | 203 |
| 15.1 Giới thiệu | 203 |
| 15.2 Xây dựng và tối ưu hàm mất mát | 205 |
| 15.3 Lập trình Python | 207 |
| 15.4 Thảo luận | 210 |
| 15.5 Câu hỏi ôn tập bài 15 | 210 |
| 15.6 Bài tập cuối chương | 211 |

Chương V CÁC BÀI TẬP THỰC HÀNH

| | | |
|----------|---|-----|
| 1 | Làm quen môi trường | 214 |
| 1.1 | Giới thiệu google colab | 214 |
| 1.2 | Các nguồn dữ liệu huấn luyện miễn phí | 214 |
| 1.3 | Các lệnh nhập xuất dữ liệu | 215 |
| 1.4 | Mô hình | 216 |
| 2 | Dự án học máy từ đầu đến cuối | 217 |
| 2.1 | Thu thập, khám phá trực quan dữ liệu | 217 |
| 2.2 | Chọn, huấn luyện, tinh chỉnh mô hình | 218 |
| 3 | Huấn luyện mô hình hồi quy tuyến tính | 219 |
| 3.1 | Hồi quy tuyến tính | 219 |
| 3.2 | Chọn, huấn luyện, tinh chỉnh mô hình | 220 |
| 4 | Huấn luyện mô hình hồi quy logistic, softmax | 222 |
| 4.1 | Ước lượng xác suất | 222 |
| 4.2 | Huấn luyện và hàm chi phí | 222 |
| 4.3 | Ranh giới quyết định | 222 |
| 4.4 | Hồi quy softmax | 223 |
| 5 | Huấn luyện mô hình SVM | 224 |
| 5.1 | Phân loại SVM tuyến tính | 224 |
| 5.2 | Phân loại SVM phi tuyến | 224 |
| 5.3 | Hồi quy SVM | 225 |

| | |
|---|-----|
| 6 Thực hành phân cụm Kmean | 226 |
| Tài liệu tham khảo | 227 |
| Tài liệu tham khảo | 227 |

DANH SÁCH HÌNH ẢNH

| | | |
|------|--|----|
| 3.1 | Minh hoạ dữ liệu và đường thẳng xấp xỉ tìm được bởi hồi quy tuyến tính | 45 |
| 3.2 | (a) Hồi quy đa thức bậc ba (b) Hồi quy tuyến tính nhạy cảm với nhiễu. | 46 |
| 6.1 | Khảo sát sự biến thiên của một đa thức bậc hai. | 71 |
| 6.6 | Nghiệm của bài toán hồi quy tuyến tính (đường thẳng màu đen) tìm được bằng thư viện scikit-learn. | 77 |
| 6.7 | Đường đi nghiệm của hồi quy tuyến tính với các tốc độ học khác nhau. | 78 |
| 6.12 | Đường đi của nghiệm cho bài toán hồi quy tuyến tính với hai phương pháp Hạ gradient khác nhau. NAG cho nghiệm mượt hơn và nhanh hơn. | 83 |
| 6.13 | Ví dụ về giá trị hàm mất mát sau mỗi vòng lặp khi sử dụng mini-batch Hạ gradient. Hàm mất mát dao động sau mỗi lần cập nhật nhưng nhìn chung giảm dần và có xu hướng hội tụ. | 85 |
| 7.1 | Bài toán phân loại nhị phân trong không gian hai chiều. | 89 |
| 7.2 | Các đường thẳng trong không gian hai chiều | 90 |
| 7.4 | Biểu diễn perceptron và hồi quy tuyến tính dưới dạng mạng neuron. | 94 |
| 7.5 | Cấu trúc của một neuron thần kinh sinh học. | 95 |
| 7.6 | PLA có thể có nhiều nghiệm. | 96 |