BÀI TẬP SỐ 1

Lớp chia nhóm, mỗi nhóm chọn một đề tài để thực hiện phương pháp Hồi quy tuyến tính và k-fold cross validation. Yêu cầu thực hiện của mỗi đề tài gồm:

1. Các nhóm đăng kí đề tài.

2. Mô tả đề tài: Tên đề tài, mô tả tóm tắt công việc thực hiện của đề tài.

3. Mô tả tập dữ liệu của bài toán: Dữ liệu gồm những chiều thông tin gì (mỗi vector dữ liệu có những thông tin gì), có bao nhiêu vector dữ liệu (ít nhất là 100 vector dữ liệu). Ví dụ dự đoán về bệnh tim của một người, thì mỗi người có 1 vector dữ liệu gồm các thông tin: tuổi, giới tính, nhịp tim, huyết áp, ... Thông tin cần dự đoán của bài toán là gì (mô tả dữ liệu đầu ra tương ứng với mỗi vector dữ liệu).

4. Chia tập dữ liệu thành 2 phần: 70% dùng để huấn luyện mô hình, 30% dùng để kiểm tra sự phù hợp của mô hình.

5. Mô tả ma trận dữ liệu huấn luyện (X), vector đầu ra (Y)

6. Dùng phương pháp Hồi quy tuyến tính để xây dựng mô hình cho bài toán với tập dữ liệu huấn luyện.

7. Dùng phương pháp Hồi quy tuyến tính và k-fold cross validation để xây dựng mô hình cho bài toán với tập dữ liệu huấn luyện.

   Bước 1: Chia toàn bộ tập dữ liệu huấn luyện thành k phần (phương pháp k-fold cross validation).

   Bước 2: Chọn ngẫu nhiên k-1 phần làm training data, 1 phần còn lại làm test data. Sử dụng phương pháp học máy đã lựa chọn trên tập training data và test data để xây dựng và đánh giá mô hình. Bước 2 này được làm k lần.

   Bước 3: Chọn mô hình có (train error + validation error) là nhỏ nhất.

   Ví dụ:

   Bước 1: Chia tập dữ liệu thành 3 tập A, B, C.

   Bước 2: Huấn luyện và đánh giá mô hình trên tập training data và test data.

   - Lần 1: training data: A, B; test data: C.

- Lần 2: training data: A, C; test data: B.
- Lần 3: training data: B, C; test data: A.

Trong 3 lần trên, lần nào có (train error + validation error) nhỏ nhất thì mô hình huấn luyện của lần đó được chọn làm mô hình dự đoán cho dữ liệu mới.

8. Dùng tập dữ liệu kiểm tra để so sánh kết quả dự đoán của mô hình với kết quả thực tế, để đánh giá sự phù hợp của mô hình. Tính tỷ lệ mẫu được dự đoán đúng trên tổng số mẫu.

9. Lý thuyết cần trình bày: Phương pháp Hồi quy tuyến tính, phương pháp k-fold cross validation. Mỗi phương pháp cần trình bày:

   + Input:

   + Output:

   + Method (cách thực hiện): Ý tưởng của phương pháp, cách thực hiện của phương pháp (cách xây dựng hàm mất mát, cách tìm hàm mất mát tối ưu, cách giải bài toán tối ưu, nghiệm của bài toán tối ưu), đánh giá phương pháp.

10. **Báo cáo làm trên file word theo mẫu GV**.

    Cần nộp GV: files code; files dữ liệu, file báo cáo và file các slide trình bày.

11. Thời gian nộp bài: hạn cuối 17h, thứ 2 (10/10)

12. Thời gian báo cáo bài tập: tiết học thứ 3 (11/10)