Chương 1.1 : Giới thiệu về học máy

Trịnh Anh Phúc [1]

[1]Bộ môn Khoa Học Máy Tính, Viện CNTT & TT, Trường Đại Học Bách Khoa Hà Nội

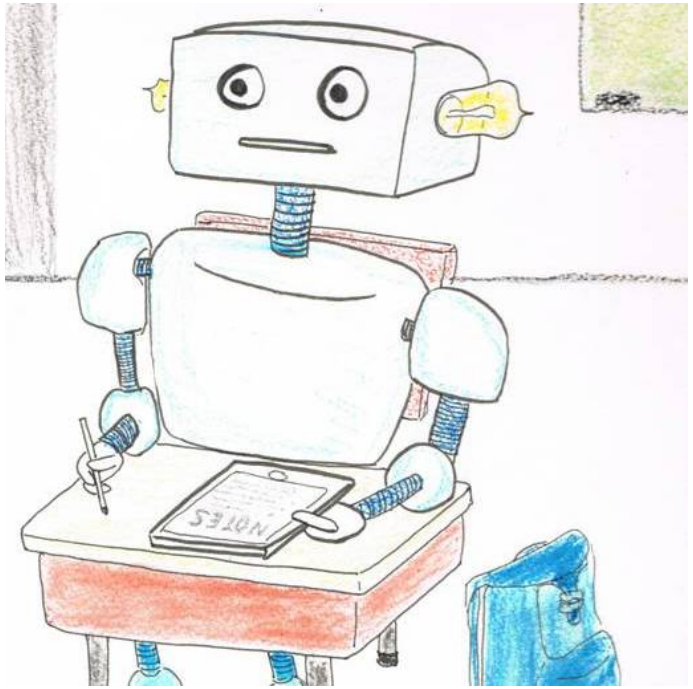
Ngày 9 tháng 3 năm 2020

Giới thiệu

Định nghĩa và các khái niệm cơ bản

Định nghĩa 1

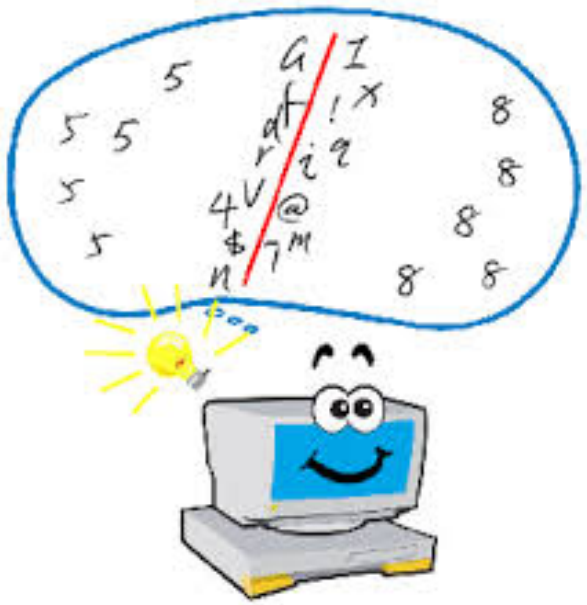
Học máy (machine learning) là một nhánh của trí tuệ nhân tạo liên quan đến việc xây dựng hoặc nghiên cứu các hệ thống có thể *học* từ dữ liệu.



Định nghĩa và các khái niệm cơ bản

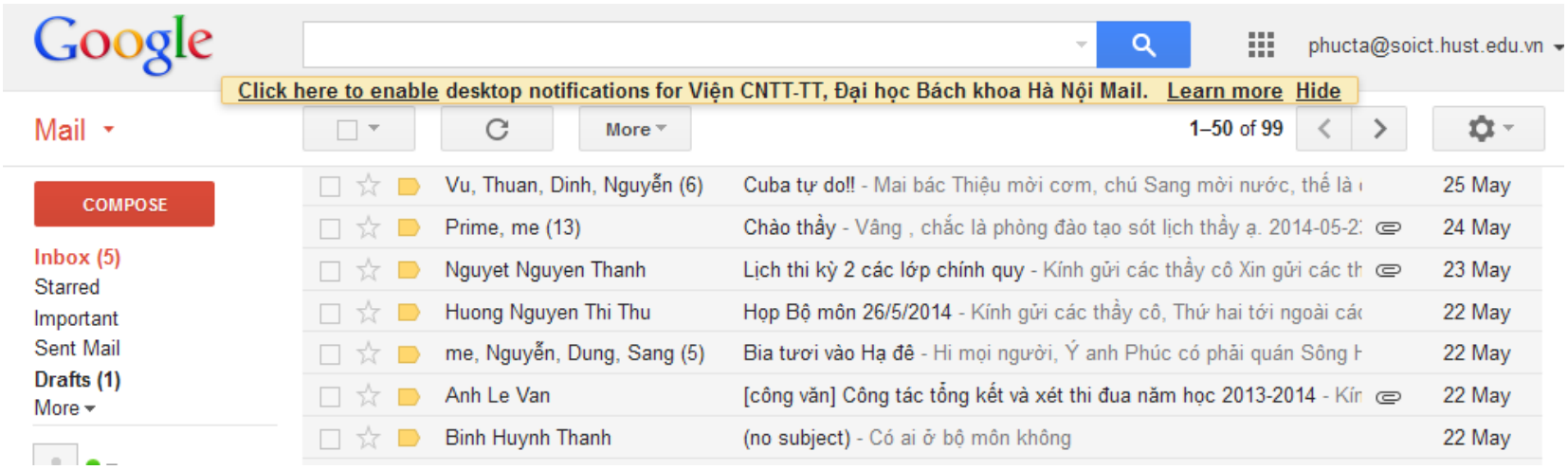
Định nghĩa 2

Học máy là một chương trình máy tính được gọi là học từ tập dữ liệu $\mathcal{D}$ để thực hiện một thao tác học $\mathcal{LT}$ - learning task - cho trước. Tập dữ liệu $\mathcal{D}$ thường được gọi là tập học - training set.



Định nghĩa và các khái niệm cơ bản

Ví dụ minh họa, ứng dụng phân loại thư điện tử



Định nghĩa và các khái niệm cơ bản

Vấn đề đặt ra

- Số lượng thư điện tử là quá lớn
- Có quá nhiều thư rác, quảng cáo nội dung không quan trọng làm đầy hòm thư
- Nhanh chóng xác định thư quan trọng liên quan công việc, gia đình
- Loại bỏ nhanh chóng thư rác, thư không quan trọng

Định nghĩa và các khái niệm cơ bản

Phân biệt tính chất của các thư điện tử

Một nhân viên văn phòng cần ứng dụng phân biệt các thư điện tử gửi đến hộp thư điện tử của mình. Yêu câu tách các thư này ra ba loại

- **Thư riêng tư** : thư liên quan công việc gia đình, thư vợ, chồng, con cái v.v.....
- **Thư công việc** : thư liên quan đến công việc công tác hàng ngày, thư lãnh đạo, thư văn phòng, v.v....
- **Thư rác** : quảng cáo, rao vặt, linh tinh v.v....

Dành cho trả lời câu hỏi

Định nghĩa và các khái niệm cơ bản

Phân biệt tính chất của các thư điện tử

Thông thường để tạo một ứng dụng học máy như trên cần một tập mẫu - tập học $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}$ với mọi $i = \overline{1, m}$, e.g. thư điện tử, được lấy mẫu một cách độc lập.

Định nghĩa và các khái niệm cơ bản

Phân biệt tính chất của các thư điện tử (tiếp)

Thông tin thuộc tính thư điện tử đc thu thập như sau

- Địa chỉ thư người gửi (sender) là người lạ x_1?
- Tiêu đề có nội dung lành mạnh hay không x_2?
- Có liên kết (hyperlink) đến nguồn tin khác x_3?
- Địa chỉ có đuôi là một cửa hàng trực tuyến mình đã mua hàng hay không x_4?
- Có từ khóa trong lĩnh vực văn phòng hay không x_5 ?

Do các thông tin thuộc tính đều là boolean nên tập giá trị nhị phân $\{0, 1\}$

Định nghĩa và các khái niệm cơ bản

Phân biệt tính chất của các thư điện tử (tiếp)

Có tất cả ba nhãn $\{0, 1, 2\}$- label - lần lược đc gán cho mỗi thư điện tử dùng để phân biệt ba loại thư

- **Thư công việc** - public
- **Thư riêng tư** - private
- **Thư rác** - spam

Định nghĩa và các khái niệm cơ bản

Nhân viên có một tập học $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_5, y_5)\}$ gồm 5 mẫu thư điện tử có thể trông như sau

STT	x	y
$(\mathbf{x}_1, y_1)$	$[x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0]^T$	0
$(\mathbf{x}_2, y_2)$	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0, x_5 = 1]^T$	1
$(\mathbf{x}_3, y_3)$	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0]^T$	1
$(\mathbf{x}_4, y_4)$	$[x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0]^T$	0
$(\mathbf{x}_5, y_5)$	$[x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 1]^T$	2

Định nghĩa và các khái niệm cơ bản

Xét mẫu thư $(\mathbf{x}_2, y_2)$ thì ta có được nội dung sau

$$\mathbf{x}_2 = [x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0, x_5 = 1]^T$$

- bức thư có địa chỉ không đến từ người lạ
- tiêu đề có nội dung lành mạnh
- có đường dẫn hyperlink
- địa chỉ gửi không phải từ cửa hàng mua hàng trực tuyến
- đồng thời có từ khóa trong lĩnh vực văn phòng

Nhãn gán cho bức thư này là

$$y_2 = 1$$

nghĩa là bức thư trên có nội dung riêng tư.

Định nghĩa và các khái niệm cơ bản

Thao tác học - $\mathcal{LT}$ có giám sát và không giám sát

Thao tác học - learning task được chia làm hai loại chính

- Loại thao tác học có giám sát - supervised learning
- Loại thao tác học không giám sát - unsupervised learning

Định nghĩa và các khái niệm cơ bản

Thao tác học có giám sát

Sở dĩ được gọi là có giám sát vì kèm theo mỗi dữ liệu, ta có một nhãn tương ứng $\mathcal{D} = \{(\mathbf{x}_1, y_1), \cdots, (\mathbf{x}_m, y_m)\}$. Ý nghĩa của từng cặp thành phần tập học

- Nhãn y là thành phần giám sát tương ứng mỗi dữ liệu $\mathbf{x}$
- Các cặp dữ liệu trong $\mathcal{D}$ là độc tập xác suất từng đôi
- Mỗi cặp $(\mathbf{x}, y)$ là phụ thuộc thuộc tính, ý nghĩa tương ứng yêu cầu của ứng dụng

Định nghĩa và các khái niệm cơ bản

Phân biệt thư điện tử

Bộ dữ liệu thể hiện rõ thao tác học có giám sát

- Các mẫu thư điện tử đều có gán nhãn tương ứng trong tập gồm ba giá trị rời rạc $\{0, 1, 2\}$
- Mỗi mẫu trong tập học $\mathcal{D}$ đều là một cặp dữ liệu, nhãn $(\mathbf{x}, y)$
- Có tất cả 6 dữ liệu độc lập trong tập học $\mathcal{D}$

Định nghĩa và các khái niệm cơ bản

Thao tác học không giám sát

Đặc điểm của thao tác học không giám sát là tập học không được gán nhãn $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$. Như vậy thành phần của tập học

- Không được gán nhãn y
- Chỉ gồm tập các dữ liệu $\mathbf{x}$ độc lập xác suất từng đôi
- Thường được dùng cho các ứng dụng trích chọn đặc trưng của tập dữ liệu $\mathcal{X}$

Định nghĩa và các khái niệm cơ bản

Phân biệt thư điện tử

Bộ dữ liệu trên cũng có thể chuyển tương ứng thành thao tác học không giám sát khi ta loại bỏ các nhãn y. Ứng dụng có thể là xem tần suất xuất hiện thư điện tử có nội dung ko lành mạnh.

STT	x
1	$[x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0]^T$
2	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0, x_5 = 1]^T$
3	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0]^T$
4	$[x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0]^T$
5	$[x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 1]^T$

Dành cho trả lời câu hỏi

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Định nghĩa về dự đoán - prediction

Là hành động tuyên bố ở thì tương lai dựa trên dữ liệu, sự kiện trong quá khứ.

Phân biệt thư điện tử

Nhân viên văn phòng muốn một chương trình cho phép dự đoán mọi thư điện tử, đến trong tương lai, thuộc một trong ba thể loại thư điện tử cho trước

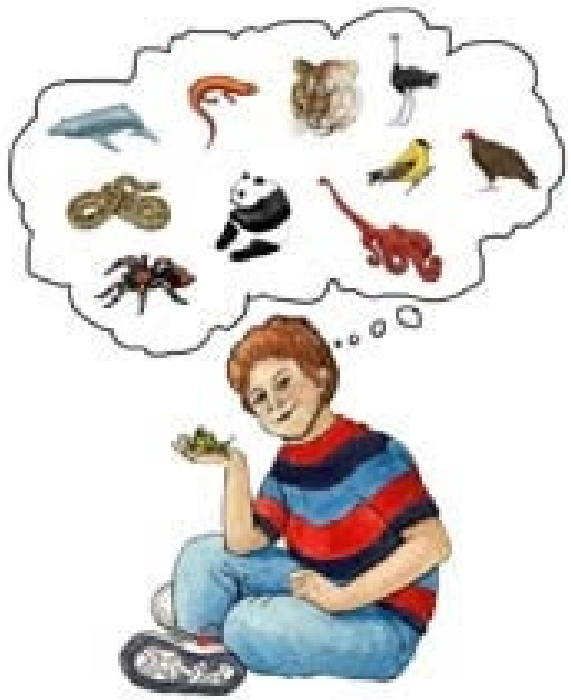
- Thư riêng tư
- Thư công việc
- Thư rác

Dựa trên tập dữ liệu học $\mathcal{D}$ đã sưu tập.

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Định nghĩa về hàm dự đoán - prediction function

Là một ánh xạ **không gian dữ liệu đầu vào** sang **không gian nhãn đầu ra** $f : \mathbf{x} \mapsto y$



Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Dự đoán thư điện tử

- **Không gian dữ liệu đầu vào** gồm các thư điện tử biểu diễn bằng vec tơ nhị phân 5 chiều $\mathbf{x} = [x_1, x_2, x_3, x_4, x_5]^T$
- **Không gian nhãn đầu ra** y gồm có 3 giá trị rời rạc $\{0, 1, 2\}$
- **Hàm dự đoán**

$$\hat{y} = f(\mathbf{x}; \mathbf{w}) = \begin{cases} 0 \text{ nếu } \mathbf{x} \cdot \mathbf{w} < 1 \\ 1 \text{ nếu } 1 \leq \mathbf{x} \cdot \mathbf{w} < 2 \\ 2 \text{ nếu } 2 \leq \mathbf{x} \cdot \mathbf{w} \end{cases}$$

trong đó $\mathbf{w} = [w_1 = 1, w_2 = 0, w_3 = 0, w_4 = 0, w_5 = 1]^T$

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Dự đoán thư điện tử (tiếp)

Áp dụng công thức hàm dự đoán tương ứng tập dữ liệu $\mathcal{D}$, kết quả có được như sau

STT	$\mathbf{x}$	y	$\hat{y}$
$(\mathbf{x}_1, y_1)$	$[x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0]^T$	0	0
$(\mathbf{x}_2, y_2)$	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0, x_5 = 1]^T$	1	1
$(\mathbf{x}_3, y_3)$	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0]^T$	1	0
$(\mathbf{x}_4, y_4)$	$[x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0]^T$	0	1
$(\mathbf{x}_5, y_5)$	$[x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 1]^T$	2	2

Dự đoán sai hai mẫu số thứ tự 3 và 4 với tham số
$\mathbf{w} = [w_1 = 1, w_2 = 0, w_3 = 0, w_4 = 0, w_5 = 1]^T$

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Dự đoán thư điện tử (tiếp)

Có các nhận xét chung như sau

- Hàm dự đoán có kết quả đúng ở mức trung bình ≈ 60 %
- Tham số $\mathbf{w}$ thay đổi dẫn đến kết quả dự đoán sẽ thay đổi

$\Rightarrow$ Nhân viên văn phòng không hài lòng, cô ấy muốn một hàm dự đoán hay tham số $\mathbf{w}$ đoán đúng hơn 60% cơ :)

Dành cho trả lời câu hỏi

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Hàm mục tiêu - target function

Là hàm phụ thuộc vào tham số **w** và tập dữ liệu học $\mathcal{D}$ nhằm đo kết quả dự đoán của hàm dự đoán trên tập dữ liệu học $\mathcal{D}$. Hàm này có giá trị biến thiên có ý nghĩa như sau

- *nhỏ* khi kết quả dự đoán *tốt*
- *lớn* khi kết quả dự đoán *tồi*

trên tập dữ liệu học $\mathcal{D}$

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Các hàm mục tiêu tiêu biểu

- Hàm hay dùng cho bài toán phân loại - hinge loss

$$hl(\mathbf{x}_i, y_i) = \begin{cases} 0 \text{ nếu } f(\mathbf{x}_i; \mathbf{w}) = y_i \\ 1 \text{ không đúng} \end{cases}$$

Dùng cho mọi cặp dữ liệu trong tập học $\mathcal{D}$, hàm mục tiêu sau

$$\mathcal{L}(\mathcal{D}) = \sum_{i=1}^{m} hl(\mathbf{x}_i, y_i)$$

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Dự đoán thư điện tử (tiếp)

Áp dụng công thức trên tập dữ liệu $\mathcal{D}$ thư điện tử, kết quả có được $\mathcal{L}(\mathcal{D}) = \sum_{i=1}^{5} hl(\mathbf{x}_i, y_i) = 2$

STT	$\mathbf{x}$	y	$\hat{y}$
$(\mathbf{x}_1, y_1)$	$[x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0]^T$	0	0
$(\mathbf{x}_2, y_2)$	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0, x_5 = 1]^T$	1	1
$(\mathbf{x}_3, y_3)$	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0]^T$	1	0
$(\mathbf{x}_4, y_4)$	$[x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0]^T$	0	1
$(\mathbf{x}_5, y_5)$	$[x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 1]^T$	2	2

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Các hàm mục tiêu tiêu biểu (tiếp)

- Hàm hay dùng cho bài toán hồi quy - square loss

$$sl(\mathbf{x}_i, y_i) = (f(\mathbf{x}_i; \mathbf{w}) - y_i)^2$$

Tổng cho mọi cặp dữ liệu trong tập học $\mathcal{D}$, chúng ta có hàm mục tiêu sau

$$\mathcal{L}(\mathcal{D}) = \sum_{i=1}^{m} sl(\mathbf{x}_i, y_i) = \sum_{i=1}^{m} (f(\mathbf{x}_i) - y_i)^2$$

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Dự đoán thư điện tử (tiếp)

Áp dụng công thức trên tập dữ liệu $\mathcal{D}$ thư điện tử, kết quả có được $\mathcal{L}(\mathcal{D}) = \sum_{i=1}^{5} sl(\mathbf{x}_i, y_i) = 1^2 + 1^2 = 2$

STT	$\mathbf{x}$	y	$\hat{y}$
$(\mathbf{x}_1, y_1)$	$[x_1 = 0, x_2 = 0, x_3 = 0, x_4 = 1, x_5 = 0]^T$	0	0
$(\mathbf{x}_2, y_2)$	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 0, x_5 = 1]^T$	1	1
$(\mathbf{x}_3, y_3)$	$[x_1 = 0, x_2 = 1, x_3 = 1, x_4 = 1, x_5 = 0]^T$	1	0
$(\mathbf{x}_4, y_4)$	$[x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 0]^T$	0	1
$(\mathbf{x}_5, y_5)$	$[x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 1]^T$	2	2

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Các hàm mục tiêu tiêu biểu (tiếp)

- Hàm hay dùng cho bài toán không giám sát - loglikelihood

$$lg(\mathbf{x}_i) = -\log P(f(\mathbf{x}_i; \mathbf{w}))$$

Tổng cho mọi dữ liệu thuộc $\mathcal{X} = \{\mathbf{x}_1, \cdots, \mathbf{x}_m\}$, chúng ta có hàm mục tiêu sau

$$\mathcal{L}(\mathcal{X}) = \sum_{i=1}^{m} lg(\mathbf{x}_i) = -\sum_{i=1}^{m} \log P(f(\mathbf{x}_i; \mathbf{w}))$$

Dành cho trả lời câu hỏi

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Bài toán học máy

Qui về bài toán tìm tham số $\mathbf{w}$ nhằm tối thiểu hóa hàm mục tiêu $\mathcal{L}(\mathcal{D})$ hay $\mathcal{L}(\mathcal{X})$.

$$\mathbf{w} = \arg\min \mathcal{L}(\mathcal{D})$$

Phụ thuộc vào các yếu tố

- Hàm dự đoán
- Hàm mục tiêu
- Giải thuật tối ưu tương ứng

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Ước lượng tham số **w** trong ví dụ thư điện tử

Sử dụng hàm mục tiêu hinge loss, ta có đẳng thức

$$\begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 1 & 1 & 0 & 1 \\ 0 & 1 & 1 & 1 & 0 \\ 1 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} w_1 \\ w_2 \\ w_3 \\ w_4 \\ w_5 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \\ 1 \\ 0 \\ 2 \end{pmatrix}$$

giải ra, ta có

$$\mathbf{w}^* = [w_1 = 2, w_2 = 3, w_3 = -2, w_4 = 0, w_5 = 0]^T$$

Hàm dự đoán, Hàm mục tiêu, Ước lượng tham số

Ước lượng tham số **w** trong ví dụ thư điện tử (tiếp)

Áp dụng tham số $\mathbf{w}^*$ ta có $\mathcal{L}(\mathcal{D}) = 0$ hay độ chính xác lên đến 100%. Cô nhân viên văn phòng hài lòng

Tập kiểm tra và vấn đề học "tủ" và học chưa đủ

Định nghĩa về tập kiểm tra - test set

Là tập các mẫu dữ liệu được lấy độc lập với tập học nhằm kiểm tra chất lượng của hàm dự đoán hay toàn bộ quá trình học

Tập kiểm tra và vấn đề học "tủ" và học chưa đủ

Kiểm tra chất lượng ứng dụng thư điện tử

Do tính cẩn thận, cô nhân viên văn phòng cũng tạo ra một tập kiểm tra $\mathcal{T}$ gồm các thư điện tử khác như sau

STT	x	y
1	$[x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 1]^T$	0
2	$[x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 1, x_5 = 1]^T$	1
3	$[x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 1]^T$	1
4	$[x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0]^T$	2
5	$[x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 0]^T$	2

Dùng tham số $\mathbf{w}^* = [w_1 = 2, w_2 = 3, w_3 = -2, w_4 = 0, w_5 = 0]^T$ bị sai đến 75%

Tập kiểm tra và vấn đề học "tủ" và học chưa đủ

Kiểm tra chất lượng ứng dụng thư điện tử (tiếp)

Cô nhân viên dùng tham số tối ưu $\mathbf{w}^*$ do nó làm tối thiểu hàm mục tiêu $\mathcal{L}(\mathcal{D}) = 0$

STT	x	y	$\hat{y}$
1	$[x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 1]^T$	0	0
2	$[x_1 = 0, x_2 = 1, x_3 = 0, x_4 = 1, x_5 = 1]^T$	1	2
3	$[x_1 = 0, x_2 = 0, x_3 = 1, x_4 = 1, x_5 = 1]^T$	1	0
4	$[x_1 = 1, x_2 = 0, x_3 = 1, x_4 = 0, x_5 = 0]^T$	2	0
5	$[x_1 = 1, x_2 = 0, x_3 = 0, x_4 = 0, x_5 = 0]^T$	2	2

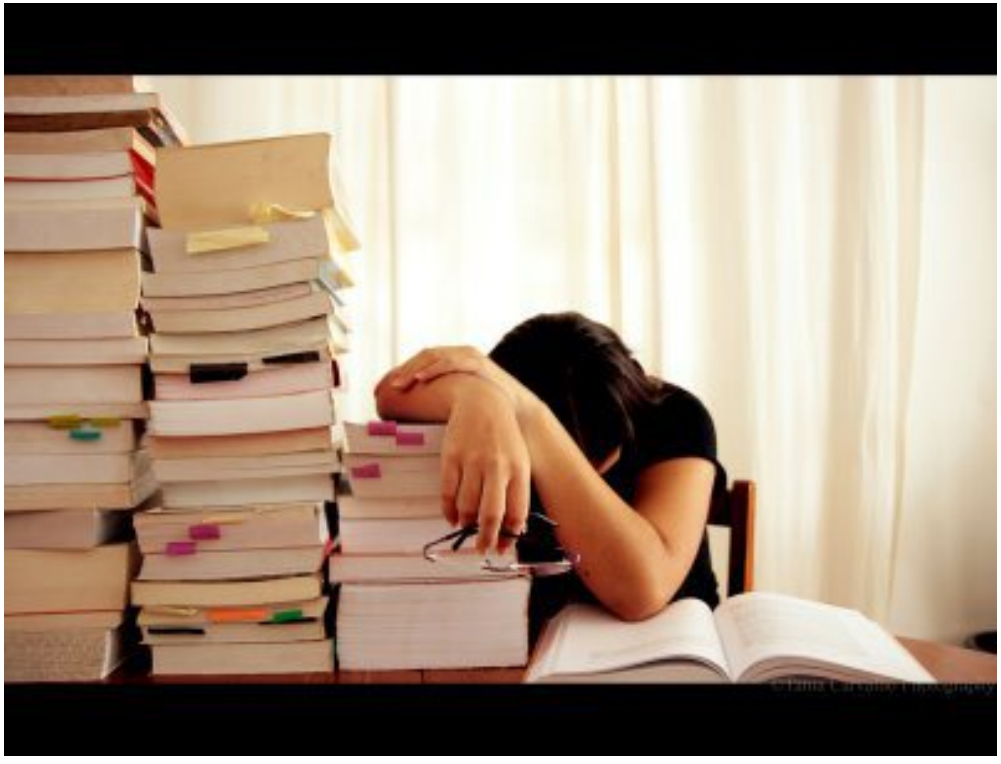
Tuy nhiên khi dùng tham số
$\mathbf{w}^* = [w_1 = 2, w_2 = 3, w_3 = -2, w_4 = 0, w_5 = 0]^T$ bị sai đến 60%

Tập kiểm tra và vấn đề học "tủ" và học chưa đủ

Tập kiểm tra và vấn đề học "tủ" và học chưa đủ

Định nghĩa về hiện tượng học "tủ" - overfitting

Là hiện tượng ước lượng tham số **w** cho kết quả tốt ở tập học nhưng cho kết quả tồi ở tập kiểm tra



Tập kiểm tra và vấn đề học "tủ" và học chưa đủ

Định nghĩa về hiện tượng học chưa đủ - underfitting

Trái ngược hiện tượng học "tủ", ta thậm chí không tìm được tham số **w** cho kết quả tốt ở tập học.

Dành cho trả lời câu hỏi

Tổng kết

- Các khái niệm cơ bản về tập dữ liệu học
- Ví dụ minh họa đi kèm
- Hàm dự đoán, hàm mục tiêu, ước lượng tham số
- Kiểm tra chất lượng ứng dụng học máy, vấn đề học "tủ"
- Hướng mô hình thay vì hướng bài toán
  1. Cây quyết định
  2. Máy học vec tơ
  3. Mạng nơ ron
  4. Mô hình trộn
  5. Mô hình markov ẩn

Bài tập về nhà

- Tải ngẫu nhiên một vài tập dữ liệu học máy từ website : UCI Machine Learning Repository (University of California, Irvine) `http://archive.ics.uci.edu/ml/`
- Giải thích ý nghĩa của thao tác học, phân loại thao tác học, trình bày rõ đặc điểm của tập học, tập test, các thuộc tính, số lượng mẫu, kiểu bài toán