MONASH University

ETC3555

Statistical Machine Learning

The learning problem

25 July 2018

Outline

(Statistical) Machine learning

- The essence of machine learning
  - A pattern exists
  - We cannot pin it down mathematically
  - We have data on it
- Learning examples
  - Spam Detection
  - Product Recommendation
  - Credit Card Fraud Detection
  - Medical Diagnosis
- Other views of “learning form data”: Statistics, Data Mining, Data Science, etc.

Exercise I

Which of the following problems are best suited for Machine Learning?

1. Classifying numbers into primes and non-primes.
2. Detecting potential fraud in credit card charges.
3. Determining the time it would take a falling object to hit the ground.
4. Determining the optimal cycle for traffic lights in a busy intersection.

a) (2) and (4)

b) (1) and (2)

c) (1), (2), and (3)

d) (3)

Different learning problems

- Supervised learning
  - (input, correct output)
- Unsupervised learning
  - (input)
- Reinforcement learning
  - (input, some output, grade for this output)

Exercise II

For each of the following tasks, identify which type of learning is involved (supervised, unsupervised or reinforced) and the training data to be used. If a task can fit more that one type, explain how and describe the training data for each type.

- Recommending a book to a user in an online bookstore
- Playing tic-tac-toe
- Categorizing movies into different types
- Learning to play music

Outline

Components of learning

Components of learning

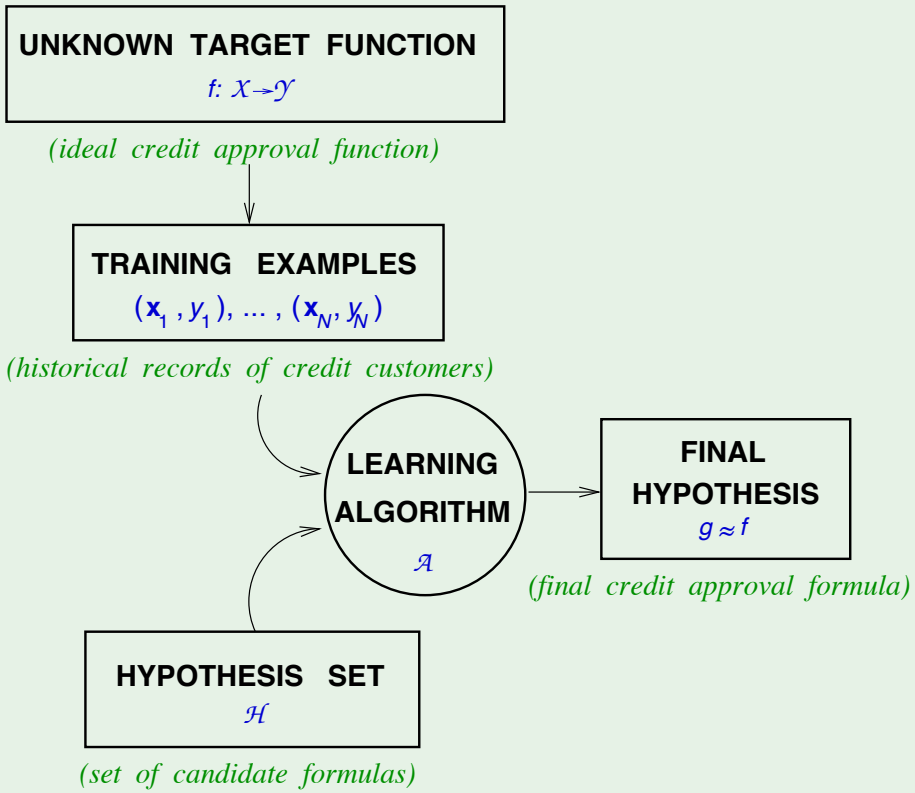
Formalization:

- Input: $\mathbf{x}$ (*customer application*)
- Output: y (*good/bad customer?*)
- Target function: $f : \mathcal{X} \to \mathcal{Y}$ (*ideal credit approval formula*)
- Data: $(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$ (*historical records*)

$\downarrow \quad \downarrow \quad \downarrow$

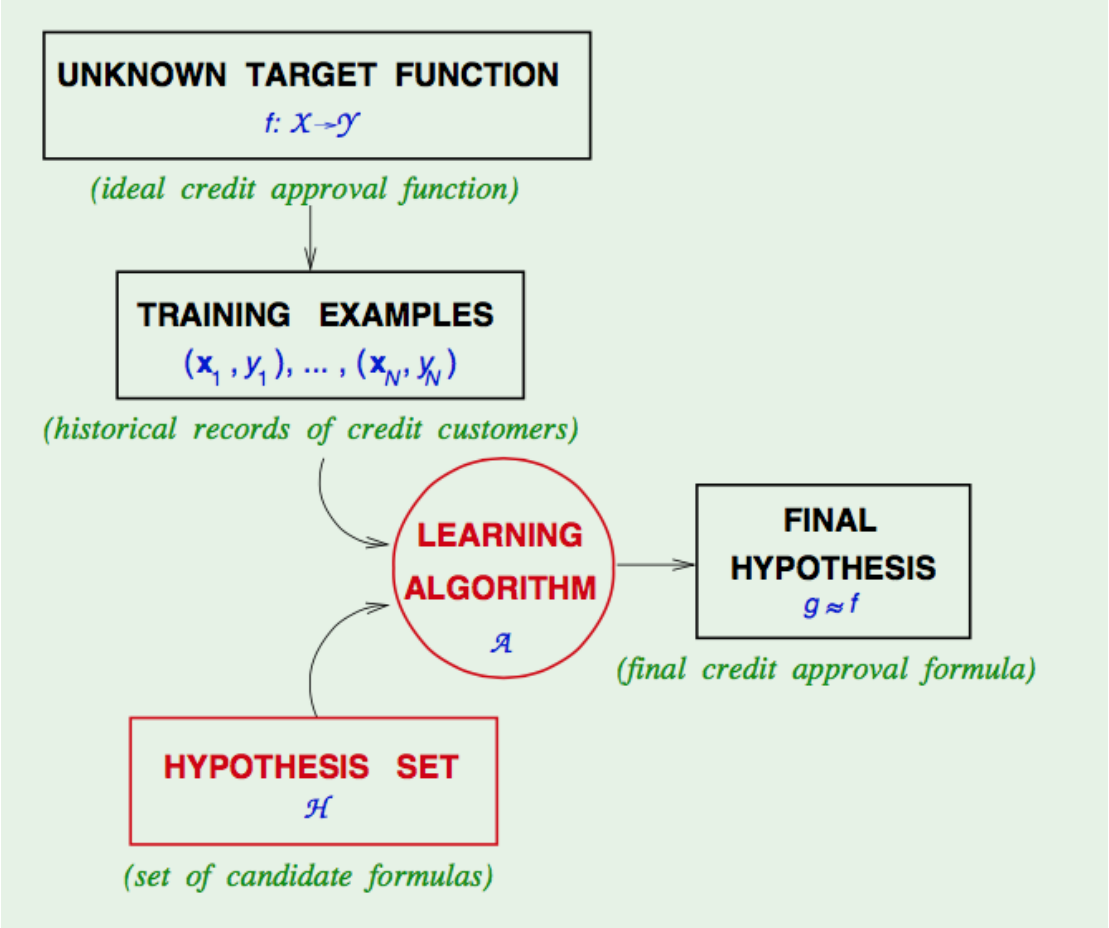
- Hypothesis: $g : \mathcal{X} \to \mathcal{Y}$ (*formula to be used*)

The learning process



The learning algorithm $\mathcal{A}$ picks $g \approx f$ from a hypothesis set $\mathcal{H}$ using training examples (data).

The learning model



"THe hypothesis set and learning algorithm are referred to informally as the *learning model*."

Outline

The perceptron hypothesis set

A simple hypothesis set - the 'perceptron'

For input $\mathbf{x} = (x_1, \cdots, x_d)$ 'attributes of a customer'

$$\text{Approve credit if} \quad \sum_{i=1}^{d} w_i x_i > \text{threshold},$$

$$\text{Deny credit if} \quad \sum_{i=1}^{d} w_i x_i < \text{threshold}.$$

This linear formula $h \in \mathcal{H}$ can be written as

$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i x_i\right) - \text{threshold}\right)$$

$$\text{sign}(s) = \begin{cases} -1 & \text{if } s < 0, \\ 1 & \text{if } s > 0. \end{cases}$$

For the moment, sign(0) is ignored (technicality).

The perceptron hypothesis set

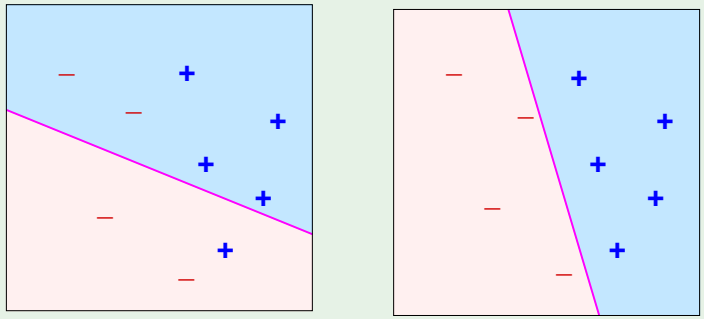
$$h(\mathbf{x}) = \text{sign}\left(\left(\sum_{i=1}^{d} w_i\, x_i\right) + \quad w_0 \quad\right)$$

Introduce an artificial coordinate $x_0 = 1$:

$$h(\mathbf{x}) = \text{sign}\left(\sum_{i=0}^{d} w_i\, x_i\right)$$

In vector form, the perceptron implements

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x})$$



'linearly separable' data

Note

The dot product of two vectors $\boldsymbol{w}$ and $\boldsymbol{x}$ is defined by $\boldsymbol{w} \cdot \boldsymbol{x} = \boldsymbol{w}^T \boldsymbol{x} = \|\boldsymbol{w}\|_2 \|\boldsymbol{x}\|_2 \cos(\theta)$ where $\cos(\theta)$ is the angle between $\boldsymbol{w}$ and $\boldsymbol{x}$.

- If the angle between $\boldsymbol{w}$ and $\boldsymbol{x}$ are less than 90 degrees, the dot product will be positive, as $\cos(\theta)$ will be positive.
- If the angle between $\boldsymbol{w}$ and $\boldsymbol{x}$ are greater than 90 degrees, the dot product will be negative, as $\cos(\theta)$ will be negative.
- If $\boldsymbol{w}$ and $\boldsymbol{x}$ are perpendicular (at 90 degrees to each other), the result of the dot product will be zero, because $\cos(\theta)$ will be zero.

The perceptron learning algorithm

A simple learning algorithm - PLA

The perceptron implements

$$h(\mathbf{x}) = \text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x})$$

Given the training set:

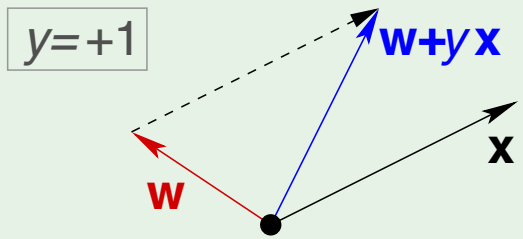
$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

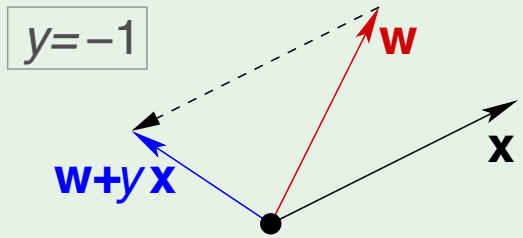
pick a misclassified point:

$$\text{sign}(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n) \neq y_n$$

and update the weight vector:

$$\boxed{\mathbf{w} \leftarrow \mathbf{w} + y_n\mathbf{x}_n}$$





The perceptron learning algorithm

Iterations of PLA

- One iteration of the PLA:

$$\mathbf{w} \leftarrow \mathbf{w} + y\mathbf{x}$$

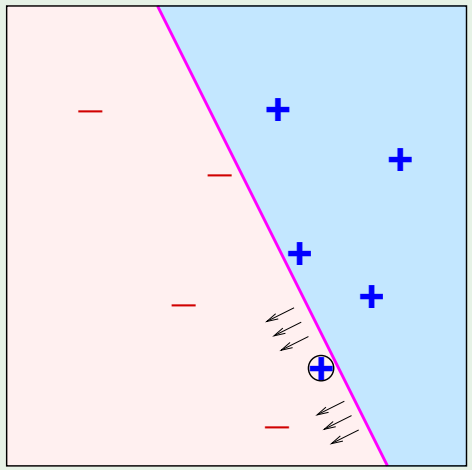
where $(\mathbf{x}, y)$ is a misclassified training point.

- At iteration $t = 1, 2, 3, \cdots$, pick a misclassified point from

$$(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \cdots, (\mathbf{x}_N, y_N)$$

and run a PLA iteration on it.

- That's it!



$\rightarrow$ Guarantee to converge if linearly separable data

Exercise III

Problem 1.2 Consider the perceptron in two dimensions: $h(\mathbf{x}) = \text{sign}(\mathbf{w}^{\mathrm{T}}\mathbf{x})$ where $\mathbf{w} = [w_0, w_1, w_2]^{\mathrm{T}}$ and $\mathbf{x} = [1, x_1, x_2]^{\mathrm{T}}$. Technically, $\mathbf{x}$ has three coordinates, but we call this perceptron two-dimensional because the first coordinate is fixed at 1.

(a) Show that the regions on the plane where $h(\mathbf{x}) = +1$ and $h(\mathbf{x}) = -1$ are separated by a line. If we express this line by the equation $x_2 = ax_1 + b$, what are the slope a and intercept b in terms of w_0, w_1, w_2?

(b) Draw a picture for the cases $\mathbf{w} = [1, 2, 3]^{\mathrm{T}}$ and $\mathbf{w} = -[1, 2, 3]^{\mathrm{T}}$.

In more than two dimensions, the $+1$ and -1 regions are separated by a *hyperplane*, the generalization of a line.

Solution

$h(\boldsymbol{x}) = +1 \implies \boldsymbol{w}^T\boldsymbol{x} > 0$, and $h(\boldsymbol{x}) = -1 \implies \boldsymbol{w}^T\boldsymbol{x} < 0$. These two regions are separated by the line $\boldsymbol{w}^T\boldsymbol{x} = 0$. This can also be written as $w_0 + w_1x_1 + w_2x_2 = 0$. In other words, $a = -\frac{w_1}{w_2}$ and $b = -\frac{w_0}{w_2}$.

Exercise IV

Consider the following dataset
$\boldsymbol{x}_1 = (3, 1), \boldsymbol{x}_2 = (1, -3), \boldsymbol{x}_3 = (-1, 3), \boldsymbol{x}_4 = (2.5, -1)$
and $y_1 = 1, y_2 = -1, y_3 = 1, y_4 = 1$.

1. Plot the data set in $[-1, 3] \times [-3, 3]$
2. Is the data linearly separable?
3. Run the perceptron algorithm with $\boldsymbol{w}(0) = (-3, 1, 1)^T$.

Exercise V

The weight update rule $w(t+1) \leftarrow w(t) + y(t)x(t)$ has the nice interpretation that it moves in the direction of classifying $x(t)$ correctly.

1. Show that $y(t)w^T(t)x(t) < 0$. [Hint: $x(t)$ is misclassified by $w(t)$]
2. Show that $y(t)w^T(t+1)x(t) > y(t)w^T(t)x(t)$. [Hint: Use the update rule].

Solution

1. Let $\hat{y}(t) = w^T(t)x(t)$. If $x(t)$ is misclassified by $w(t)$, then we have $\text{sign}(\hat{y}(t)) = 1$ and $y(t) = -1$, or $\text{sign}(\hat{y}(t)) = -1$ and $y(t) = 1$. This is equivalent to $y(t)\hat{y}(t) < 0$ or equivalently $y(t)w^T(t)x(t) < 0$.

2. $y(t)w^T(t+1)x(t) = y(t)w^T(t)x(t) + [y(t)]^2x^T(t)x(t) > y(t)w^T(t)x(t)$ since $x^T(t)x(t) > 0$. Note that the first coordinate of $x(t)$ is 1.