Machine Learning

Master Degree in Computer Science - AIS Curriculum

Lezione 6 - The Linear Model II

Marco Piangerelli
marco.piangerelli@unicam.it

08 Novembre 2018

The Linear Model

- Linear Classification

The Linear Model

- Linear Classification
- Linear Regression

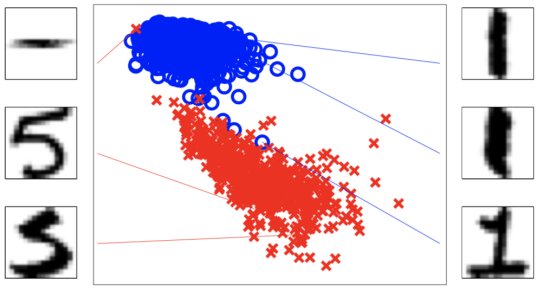
The Linear Model

- Linear Classification
- Linear Regression

The Linear Model

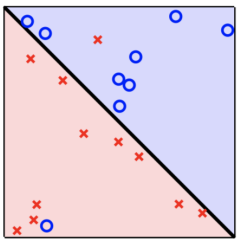
- Linear Classification $\rightarrow$ binary classification function
- Linear Regression $\rightarrow$ real-valued function

Good Features are Important

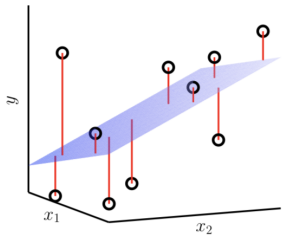


Before looking at the data, we can **reason** that symmetry and intensity should be good features **based on our knowledge of the problem**.

Algorithms



Linear Classification.
Pocket algorithm can tolerate errors
Simple and efficient



Linear Regression.
Single step learning:

$$\mathbf{w} = \mathrm{X}^{\dagger}\mathbf{y} = (\mathrm{X}^{\mathrm{T}}\mathrm{X})^{-1}\mathrm{X}^{\mathrm{T}}\mathbf{y}$$

Very efficient $O(Nd^2)$ *exact* algorithm.

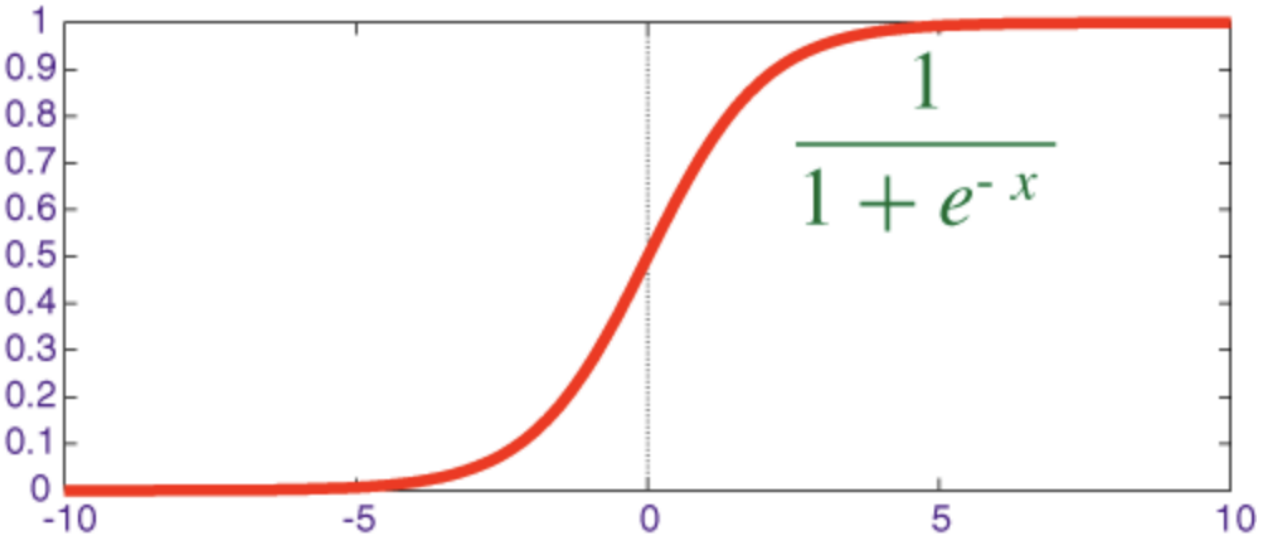
The Linear Model

- Linear Classification $\rightarrow$ binary classification function
- Linear Regression $\rightarrow$ real-valued function
- Logistic Regression $\rightarrow$ real-valued function in the sense of a Probability

The Linear Model

- Linear Classification $\rightarrow$ binary classification function
- Linear Regression $\rightarrow$ real-valued function
- Logistic Regression $\rightarrow$ real-valued function in the sense of a Probability

The Dataset



The Dataset

Will someone have a heart attack over the next year?

age	62 years
gender	male
blood sugar	120 mg/dL40,000
HDL	50
LDL	120
Mass	190 lbs
Height	$5' \ 10''$
...	...

Classification: Yes/No

Logistic Regression: Likelihood of heart attack logistic regression $\equiv y \in [0, 1]$

The Dataset

Will someone have a heart attack over the next year?

age	62 years
gender	male
blood sugar	120 mg/dL40,000
HDL	50
LDL	120
Mass	190 lbs
Height	$5' \ 10''$
...	...

Classification: Yes/No

Logistic Regression: Likelihood of heart attack

logistic regression $\equiv y \in [0, 1]$

$$h(\mathbf{x}) = \theta\left(\sum_{i=0}^{d} w_i x_i\right) = \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

Logistic regression

Will someone have a heart attack over the next year?

age	62 years
gender	male
blood sugar	120 mg/dL40,000
HDL	50
LDL	120
Mass	190 lbs
Height	$5' \ 10''$
...	...

Classification: Yes/No

Logistic Regression: Likelihood of heart attack

logistic regression $\equiv y \in [0, 1]$

$$h(\mathbf{x}) = \theta\left(\sum_{i=0}^{d} w_i x_i\right) = \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$

The Logistic regression

Will someone have a heart attack over the next year?

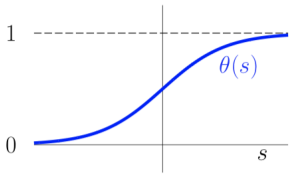
age	62 years
gender	male
blood sugar	120 mg/dL40,000
HDL	50
LDL	120
Mass	190 lbs
Height	5′ 10″
...	...

Classification: Yes/No

Logistic Regression: Likelihood of heart attack

logistic regression $\equiv y \in [0, 1]$

$$h(\mathbf{x}) = \theta\left(\sum_{i=0}^{d} w_i x_i\right) = \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})$$



$$\theta(s) = \frac{e^s}{1+e^s} = \frac{1}{1+e^{-s}}.$$

$$\theta(-s) = \frac{e^{-s}}{1+e^{-s}} = \frac{1}{1+e^s} = 1 - \theta(s).$$

The Logistic regression

$$\mathcal{D} = (\mathbf{x}_1, y_1 = \pm 1), \cdots, (\mathbf{x}_N, y_N = \pm 1)$$

$\mathbf{x}_n$	$\leftarrow$ a person's health information
$y_n = \pm 1$	$\leftarrow$ **did** they have a heart attack or not

We cannot measure a *probability*.

We can only see the occurence of an event and try to *infer* a probability.

The Logistic regression

$$f(\mathbf{x}) = \mathbb{P}[y = +1 \mid \mathbf{x}].$$

The data is generated from a *noisy* target function:

$$P(y \mid \mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{for } y = +1; \\ 1 - f(\mathbf{x}) & \text{for } y = -1. \end{cases}$$

The Logistic regression

'fitting' the data means finding a good h

h is good if:

$$\begin{cases} h(\mathbf{x}_n) \approx 1 & \text{whenever } y_n = +1; \\ h(\mathbf{x}_n) \approx 0 & \text{whenever } y_n = -1. \end{cases}$$

A simple error measure that captures this:

$$E_{\text{in}}(h) = \frac{1}{N} \sum_{n=1}^{N} \left(h(\mathbf{x}_n) - \tfrac{1}{2}(1 + y_n) \right)^2 .$$

Not very conveniente hard to minimize

The Cross Entropy Error

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + e^{-y_n \cdot \mathbf{w}^{\mathrm{T}}\mathbf{x}})$$

It looks complicated and ugly $(\ln, e^{(\cdot)}, \ldots)$,

But,

– it is based on an intuitive probabilistic interpretation of h.

– it is very convenient and mathematically friendly ('easy' to minimize).

The probabilistic Interpretation

Suppose that $h(\mathbf{x}) = \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})$ closely captures $\mathbb{P}[+1|\mathbf{x}]$:

$$P(y \mid \mathbf{x}) = \begin{cases} \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x}) & \text{for } y = +1; \\ 1 - \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x}) & \text{for } y = -1. \end{cases}$$

The probabilistic Interpretation

So, if $h(\mathbf{x}) = \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})$ closely captures $\mathbb{P}[+1|\mathbf{x}]$:

$$P(y \mid \mathbf{x}) = \begin{cases} \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x}) & \text{for } y = +1; \\ \theta(-\mathbf{w}^{\mathrm{T}}\mathbf{x}) & \text{for } y = -1. \end{cases}$$

The probabilistic Interpretation

So, if $h(\mathbf{x}) = \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x})$ closely captures $\mathbb{P}[+1|\mathbf{x}]$:

$$P(y \mid \mathbf{x}) = \begin{cases} \theta(\mathbf{w}^{\mathrm{T}}\mathbf{x}) & \text{for } y = +1; \\ \theta(-\mathbf{w}^{\mathrm{T}}\mathbf{x}) & \text{for } y = -1. \end{cases}$$

. . . or, more compactly,

$$P(y \mid \mathbf{x}) = \theta(y \cdot \mathbf{w}^{\mathrm{T}}\mathbf{x})$$

The Likelihood

$$P(y \mid \mathbf{x}) = \theta(y \cdot \mathbf{w}^{\mathrm{T}}\mathbf{x})$$

Recall: $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ are independently generated

Likelihood:
The probability of getting the $y_1, \ldots, y_N$ in $\mathcal{D}$ from the corresponding $\mathbf{x}_1, \ldots, \mathbf{x}_N$:

$$P(y_1, \ldots, y_N \mid \mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{n=1}^{N} P(y_n \mid \mathbf{x}_n).$$

The likelihood measures the probability that the data were generated if f were h.

Maximizing the Likelihood

$$
\begin{aligned}
& \max && \textstyle\prod_{n=1}^{N} P(y_n \mid \mathbf{x}_n) \\
\Leftrightarrow\ & \max && \ln\left(\textstyle\prod_{n=1}^{N} P(y_n \mid \mathbf{x}_n)\right) \\
\equiv\ & \max && \textstyle\sum_{n=1}^{N} \ln P(y_n \mid \mathbf{x}_n) \\
\Leftrightarrow\ & \min && -\frac{1}{N}\textstyle\sum_{n=1}^{N} \ln P(y_n \mid \mathbf{x}_n) \\
\equiv\ & \min && \frac{1}{N}\textstyle\sum_{n=1}^{N} \ln \frac{1}{P(y_n|\mathbf{x}_n)} \\
\equiv\ & \min && \frac{1}{N}\textstyle\sum_{n=1}^{N} \ln \frac{1}{\theta(y_n \cdot \mathbf{w}^{\mathrm{T}}\mathbf{x}_n)} \qquad \leftarrow \text{we specialize to our ``model'' here} \\
\equiv\ & \min && \frac{1}{N}\textstyle\sum_{n=1}^{N} \ln(1 + e^{-y_n \cdot \mathbf{w}^{\mathrm{T}}\mathbf{x}_n})
\end{aligned}
$$

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln(1 + e^{-y_n \cdot \mathbf{w}^{\mathrm{T}}\mathbf{x}_n})$$

How?

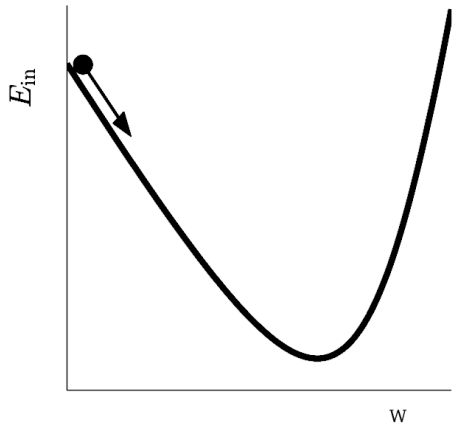
Classification – PLA/Pocket (iterative)

Regression – pseudoinverse (analytic), from solving $\nabla_{\mathbf{w}} E_{\text{in}}(\mathbf{w}) = \mathbf{0}$.

Logistic Regression – analytic won't work.

Numerically/iteratively set $\nabla_{\mathbf{w}} E_{\text{in}}(\mathbf{w}) \to \mathbf{0}$.

Gradient Descendent



$E_{\text{in}}(\mathbf{w})$ is a **convex function** of $\mathbf{w}$.

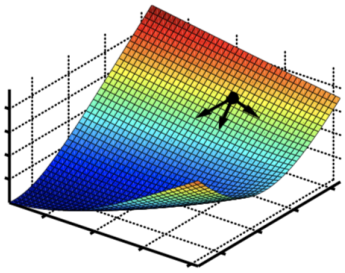
Gradient Descendent

Assume you are at weights $\mathbf{w}(t)$ and you take a step of size η in the direction $\hat{\mathbf{v}}$.

$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta\hat{\mathbf{v}}$$

We get to pick $\hat{\mathbf{v}}$ ← what's the best direction to take the step?

Pick $\hat{\mathbf{v}}$ to make $E_{\text{in}}(\mathbf{w}(t+1))$ as small as possible.



Gradient Descendent

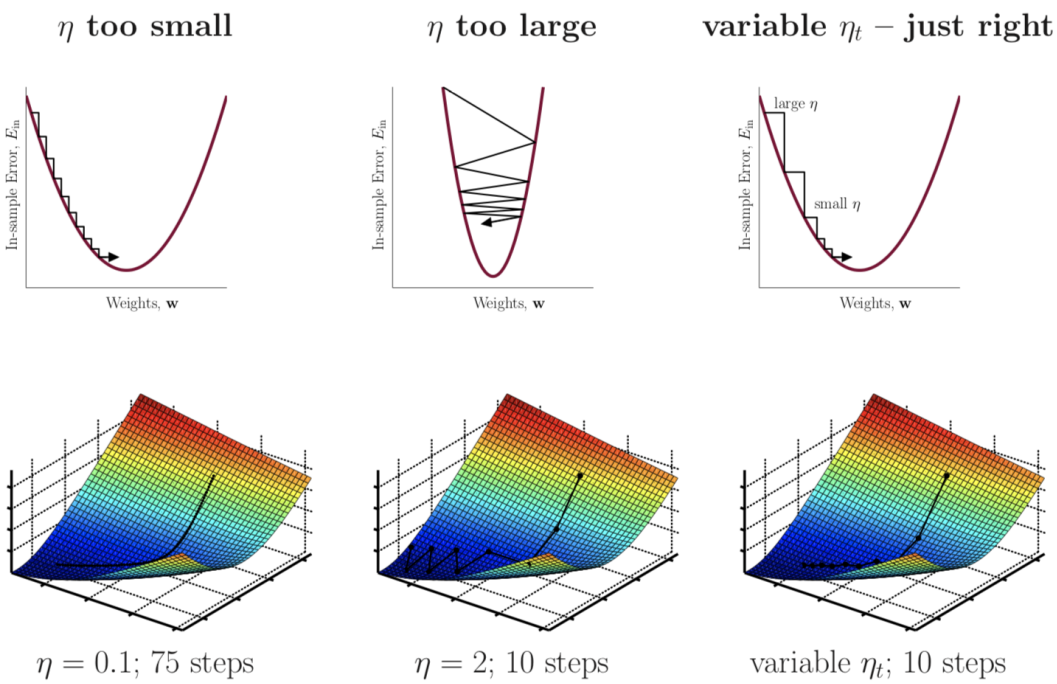
Approximating the change in E_{in}

$$\begin{aligned}
\Delta E_{\text{in}} &= E_{\text{in}}(\mathbf{w}(t+1)) - E_{\text{in}}(\mathbf{w}(t)) \\
&= E_{\text{in}}(\mathbf{w}(t) + \eta\hat{\mathbf{v}}) - E_{\text{in}}(\mathbf{w}(t)) \\
&= \eta\,\nabla E_{\text{in}}(\mathbf{w}(t))^{\mathrm{T}}\hat{\mathbf{v}} + O(\eta^2) \qquad \text{Taylor's approximation} \\
&\overbrace{\qquad\qquad}^{\text{minimized at } \hat{\mathbf{v}} = -\frac{\nabla E_{\text{in}}(\mathbf{w}(t))}{\|\nabla E_{\text{in}}(\mathbf{w}(t))\|}} \\
&\gtrapprox -\eta \|\nabla E_{\text{in}}(\mathbf{w}(t))\|
\end{aligned}$$

The best (steepest) direction to move is the negative gradient:

$$\hat{\mathbf{v}} = -\frac{\nabla E_{\text{in}}(\mathbf{w}(t))}{\|\nabla E_{\text{in}}(\mathbf{w}(t))\|}$$

Gradient Descendent



Gradient Descendent & Logistic Function

$$\eta_t = \eta \cdot \| \nabla E_{\text{in}}(\mathbf{w}(t)) \|$$

$\| \nabla E_{\text{in}}(\mathbf{w}(t)) \| \to 0$ when closer to the minimum.

$$\begin{aligned} \hat{\mathbf{v}} &= -\eta_t \cdot \frac{\nabla E_{\text{in}}(\mathbf{w}(t))}{\| \nabla E_{\text{in}}(\mathbf{w}(t)) \|} \\ &= -\eta \cdot \cancel{\| \nabla E_{\text{in}}(\mathbf{w}(t)) \|} \cdot \frac{\nabla E_{\text{in}}(\mathbf{w}(t))}{\cancel{\| \nabla E_{\text{in}}(\mathbf{w}(t)) \|}} \end{aligned}$$

$$\hat{\mathbf{v}} = -\eta \cdot \nabla E_{\text{in}}(\mathbf{w}(t))$$

1: Initialize at step $t = 0$ to $\mathbf{w}(0)$.
2: **for** $t = 0, 1, 2, \ldots$ **do**
3: Compute the gradient
$$\mathbf{g}_t = \nabla E_{\text{in}}(\mathbf{w}(t)).$$
4: Move in the direction $\mathbf{v}_t = -\mathbf{g}_t$.
5: Update the weights:
$$\mathbf{w}(t+1) = \mathbf{w}(t) + \eta \mathbf{v}_t.$$
6: Iterate 'until it is time to stop'.
7: **end for**
8: Return the final weights.

Gradient descent can minimize any smooth function, for example

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N} \sum_{n=1}^{N} \ln(1 + e^{-y_n \cdot \mathbf{w}^{\mathrm{T}} \mathbf{x}})$$