Learning From Data

Yaser S. Abu-Mostafa
California Institute of Technology

Lecture 9: The Linear Model II

Sponsored by Caltech's Provost Office, E&AS Division, and IST • Tuesday, May 1, 2012

Where we are

- Linear classification $\checkmark$
- Linear regression $\checkmark$
- Logistic regression
- Nonlinear transforms

Logistic regression - Outline

- The model

- Error measure

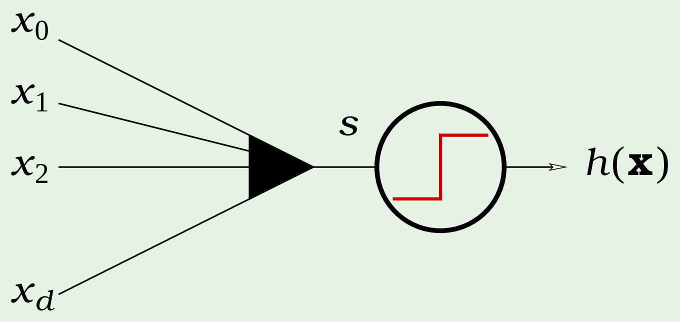
- Learning algorithm

A third linear model

$$s = \sum_{i=0}^{d} w_i x_i$$

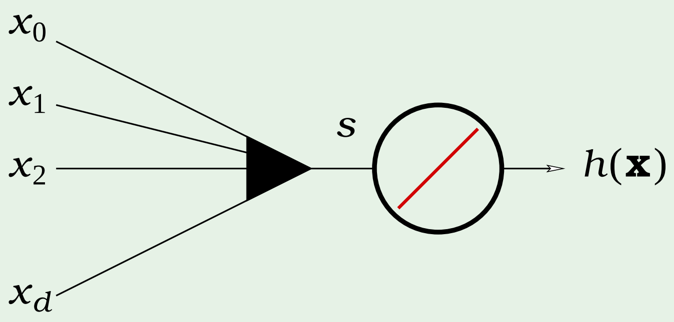
linear classification

$$h(\mathbf{x}) = \text{sign}(s)$$



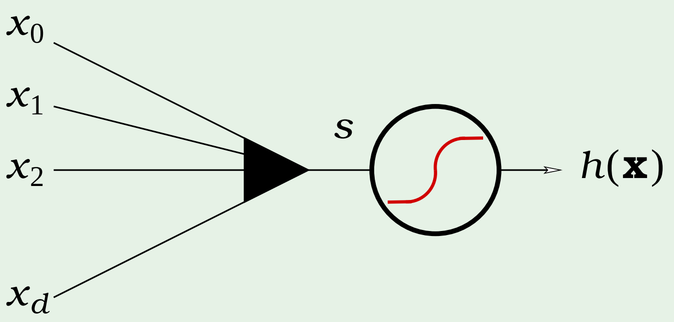
linear regression

$$h(\mathbf{x}) = s$$



logistic regression

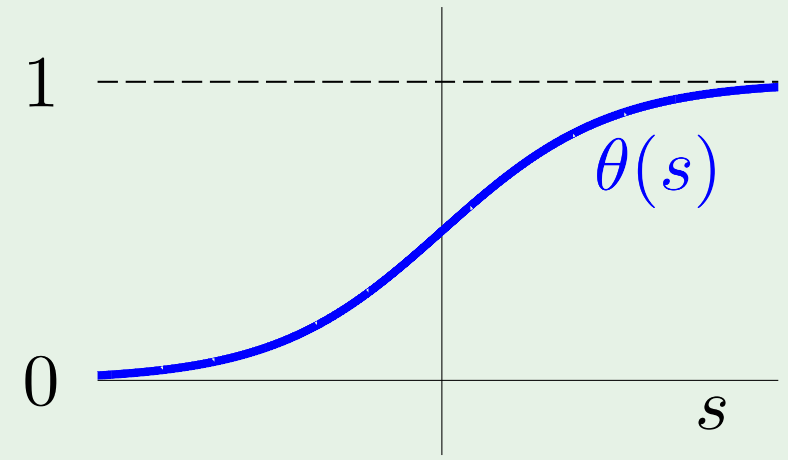
$$h(\mathbf{x}) = \theta(s)$$



The logistic function θ

The formula:

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



soft threshold: uncertainty

sigmoid: flattened out 's'

Probability interpretation

$h(\mathbf{x}) = \theta(s)$ is interpreted as a probability

Example. Prediction of heart attacks

Input $\mathbf{x}$: cholesterol level, age, weight, etc.

$\theta(s)$: probability of a heart attack

The signal $s = \mathbf{w}^{\intercal}\mathbf{x}$ "risk score"

Genuine probability

Data $(\mathbf{x}, y)$ with binary y, generated by a noisy target:

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

The target $f : \mathbb{R}^d \to [0, 1]$ is the probability

Learn $g(\mathbf{x}) = \theta(\mathbf{w}^\intercal \mathbf{x}) \approx f(\mathbf{x})$

Error measure

For each $(\mathbf{x}, y)$, y is generated by probability $f(\mathbf{x})$

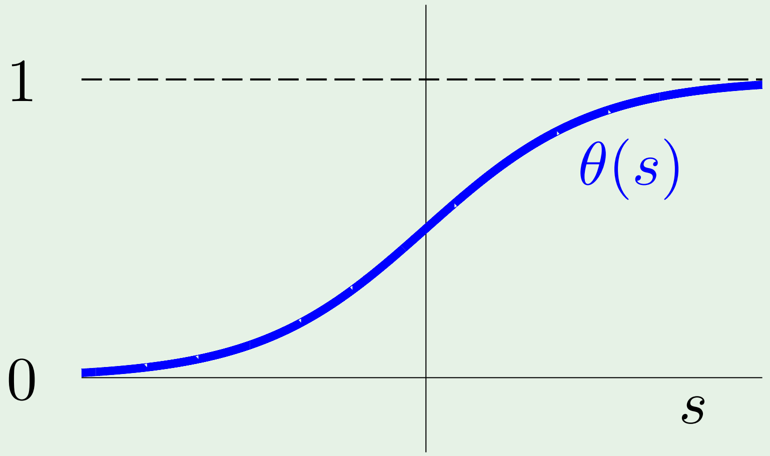
Plausible error measure based on **likelihood**:

If $h = f$, how likely to get y from $\mathbf{x}$?

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

Formula for likelihood

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



Substitute $h(\mathbf{x}) = \theta(\mathbf{w}^{\intercal}\mathbf{x})$, noting $\theta(-s) = 1 - \theta(s)$

θ(s) = $e^s/(1+e^s)$
so, θ(–s) = $e^{-s}/(1+e^{-s})$
=$1/e^s x(1/(1+e^{-s}))$
=$1/(e^s+1)$
=1- $e^s/(1+e^s)$ = 1 - θ(s)
hence, θ(–s) = 1 - θ(s)

$$P(y \mid \mathbf{x}) = \theta(y\ \mathbf{w}^{\intercal}\mathbf{x})$$

Likelihood of $\mathcal{D} = (\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_N, y_N)$ is

$$\prod_{n=1}^{N} P(y_n \mid \mathbf{x}_n) = \prod_{n=1}^{N} \theta(y_n \mathbf{w}^{\intercal}\mathbf{x}_n)$$

Maximizing the likelihood

Minimize $$-\frac{1}{N}\ln\left(\prod_{n=1}^{N}\theta(y_n\,\mathbf{w}^{\mathsf{T}}\,\mathbf{x}_n)\right)$$

Apply rule: ln(MxN) = lnM + lnN
then, $\ln(M^{-1}) = -\ln(M)$

$$=\frac{1}{N}\sum_{n=1}^{N}\ln\left(\frac{1}{\theta(y_n\,\mathbf{w}^{\mathsf{T}}\,\mathbf{x}_n)}\right)$$

Divide numerator and denominator of θ(s) by e^s

$$\left[\theta(s)=\frac{1}{1+e^{-s}}\right]$$

$$E_{\text{in}}(\mathbf{w})=\frac{1}{N}\sum_{n=1}^{N}\underbrace{\ln\left(1+e^{-y_n\mathbf{w}^{\mathsf{T}}\mathbf{x}_n}\right)}_{\mathrm{e}\left(h(\mathbf{x}_n),y_n\right)}$$

Intuitively, $\ln(1+\exp(-y_n W^T x_n))$ should work.
if $W^T x_n$ is very positive and y_n is 1 then it gives exp(high negative) ≅ 0 and then ln(1) ≅ approx 0 which is expected.

"cross-entropy" error

Logistic regression - Outline

- The model
- Error measure
- Learning algorithm

How to minimize E_{in}

For logistic regression,

Cross Entropy Error

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \ln\left(1 + e^{-y_n \mathbf{w}^{\mathsf{T}}\mathbf{x}_n}\right) \qquad \longleftarrow \text{ iterative solution}$$

its derivation is easy: $e^{-ywx}/(1+e^{-ywx})*(-yx)$ since: derv(lnx) = 1/x
$= (1/e^{ywx})*(1+ 1/e^{ywx})*(-yx)$
$=(1/e^{ywx})*((e^{ywx}+ 1)/e^{ywx})*(-yx)$
$= -yx/(1+e^{ywx})$

Compare to linear regression:

squared error

$$E_{\text{in}}(\mathbf{w}) = \frac{1}{N}\sum_{n=1}^{N} \left(\mathbf{w}^{\mathsf{T}}\mathbf{x}_n - y_n\right)^2 \qquad \longleftarrow \text{ closed-form solution}$$

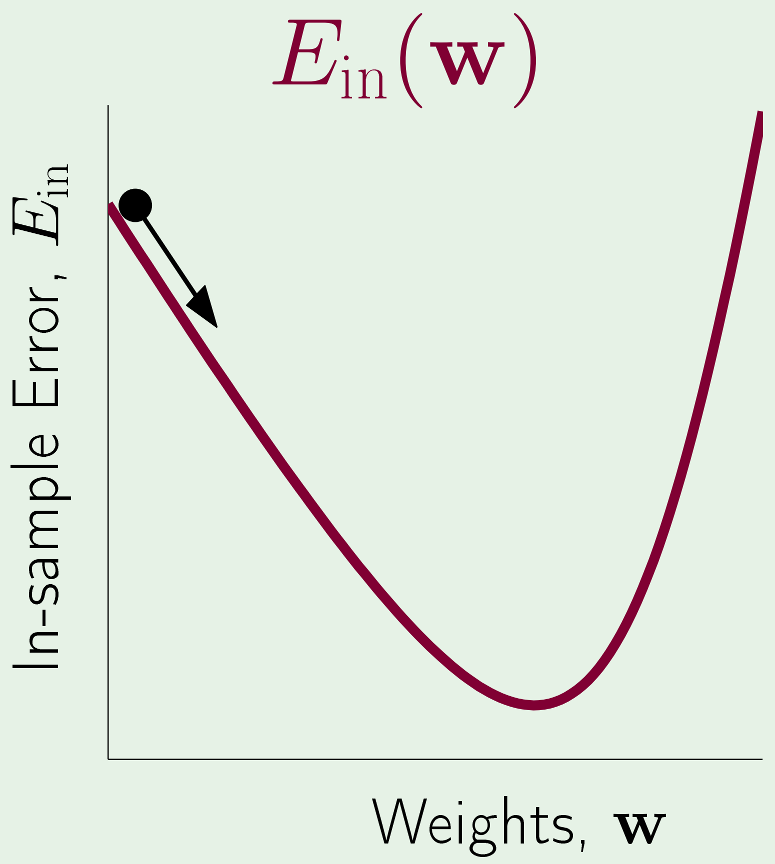
Iterative method: gradient descent

General method for nonlinear optimization

Start at $\mathbf{w}(0)$; take a step along steepest slope

Fixed step size: $\quad \mathbf{w}(1) \;=\; \mathbf{w}(0) + \eta\,\hat{\mathbf{v}}$

What is the direction $\hat{\mathbf{v}}$?



For multi-dimension V has infinite directions. Here only 2D is shown.

Formula for the direction $\hat{\mathbf{v}}$

Taylor Series:
Approximate f(x)

$f(x) \cong f(a) \;+\; (x-a)f'(a)/1! \;+\; (x-a)^2 f''(a)/2! \;+\; \ldots$
or

$f(x+a) \cong f(a) \;+\; (x)f'(a)/1! \;+\; (x)^2 f''(a)/2! \;+\; \ldots$
for Error Ein or E:

$E(\eta v+w0) \cong E(w0) \;+\; (\eta v)E'(w0)/1! \;+\; (\eta v)^2 E''(w0)/2! \;+\; \ldots$

$E(\eta v+w0) - E(w0) \cong E(w0) - E(w0) \;+\; (\eta v)E'(w0)/1! \;+\; (\eta v)^2 E''(w0)/2! \;+\; \ldots$

$\cong (\eta v)E'(w0) \;+\; (\eta v)^2 E''(w0)/2! \;+\; \ldots$

$\therefore \Delta E \cong (\eta v)\nabla E(w0) \;+\; O(\eta^2)$

$$\Delta E_{\text{in}} = E_{\text{in}}(\,\mathbf{w}(0) + \eta\hat{\mathbf{v}}\,) - E_{\text{in}}(\mathbf{w}(0))$$

$$= \eta \nabla E_{\text{in}}(\mathbf{w}(0))^{\mathrm{T}}\hat{\mathbf{v}} + O(\eta^2)$$

$>= \eta(v)\nabla E_{in}$

$>= \eta(\nabla E_{in}/||E_{in}||)\ \nabla E_{in}$ $\quad >= \eta||E_{in}||$ Nooo... it is always > 0.

$$\geq -\eta \|\nabla E_{\text{in}}(\mathbf{w}(0))\|$$

We need ΔE_{in} to be less than 0.

$\boxed{A.A^T=||A||x||A||\cos 0 = ||A||^2}$

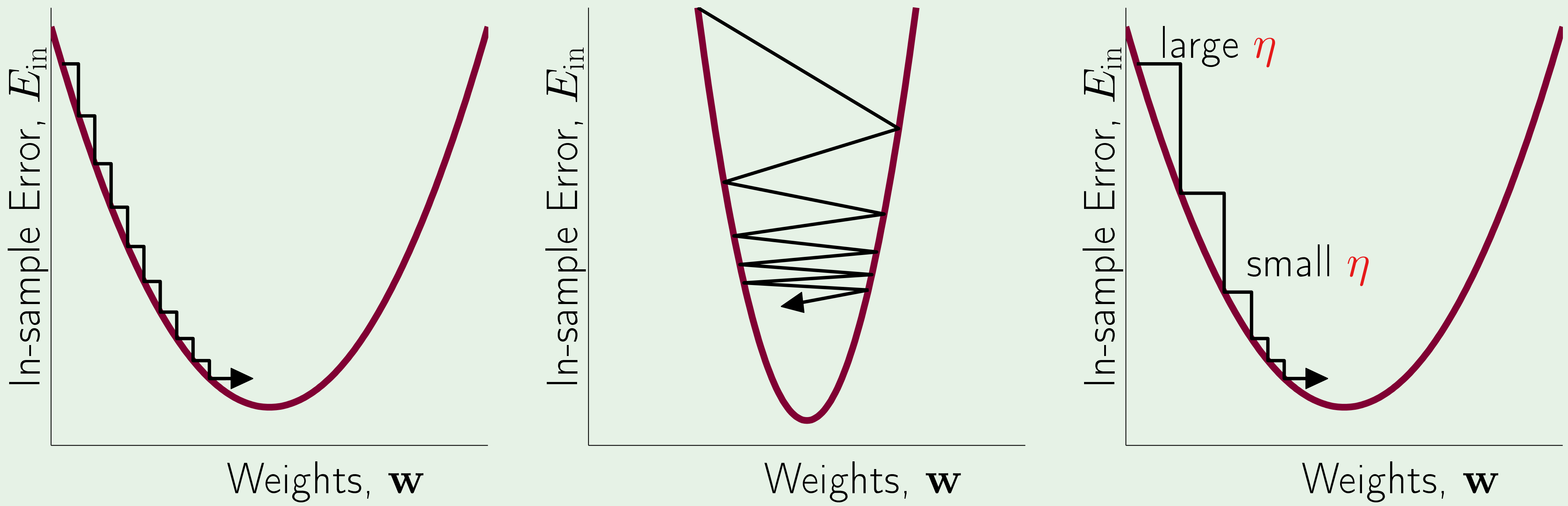
$\therefore A*A^T/||A|| = ||A||$

Since $\hat{\mathbf{v}}$ is a unit vector,

$$\hat{\mathbf{v}} = -\frac{\nabla E_{\text{in}}(\mathbf{w}(0))}{\|\nabla E_{\text{in}}(\mathbf{w}(0))\|}$$

Fixed-size step?

How η affects the algorithm:



η too small | η too large | variable η – just right

η should increase with the slope

Easy implementation

since eta is prop size of ∇E
η = const*||∇E||
$\therefore$ ΔW = -η ∇E/||∇E||
= -const*||∇E|| * ∇E / ||∇E||
= -const ∇E
lets call this constant another eta
$\therefore$ ΔW = -η ∇E

Instead of

$$\Delta \mathbf{w} = \eta\, \hat{\mathbf{v}}$$

$$= -\eta \frac{\nabla E_{\text{in}}(\mathbf{w}(0))}{\|\nabla E_{\text{in}}(\mathbf{w}(0))\|}$$

Have

$$\Delta \mathbf{w} = -\,\eta\ \nabla E_{\text{in}}(\mathbf{w}(0))$$

Fixed **learning rate** η

Logistic regression algorithm

Initialize the weights at $t = 0$ to $\mathbf{w}(0)$

for $t = 0, 1, 2, \ldots$ **do**

  Compute the gradient

$$\nabla E_{\text{in}} = -\frac{1}{N} \sum_{n=1}^{N} \frac{y_n \mathbf{x}_n}{1 + e^{y_n \mathbf{w}^{\intercal}(t) \mathbf{x}_n}}$$

  Update the weights: $\mathbf{w}(t+1) = \mathbf{w}(t) - \eta \nabla E_{\text{in}}$

  Iterate to the next step until it is time to stop

Return the final weights $\mathbf{w}$

Summary of Linear Models