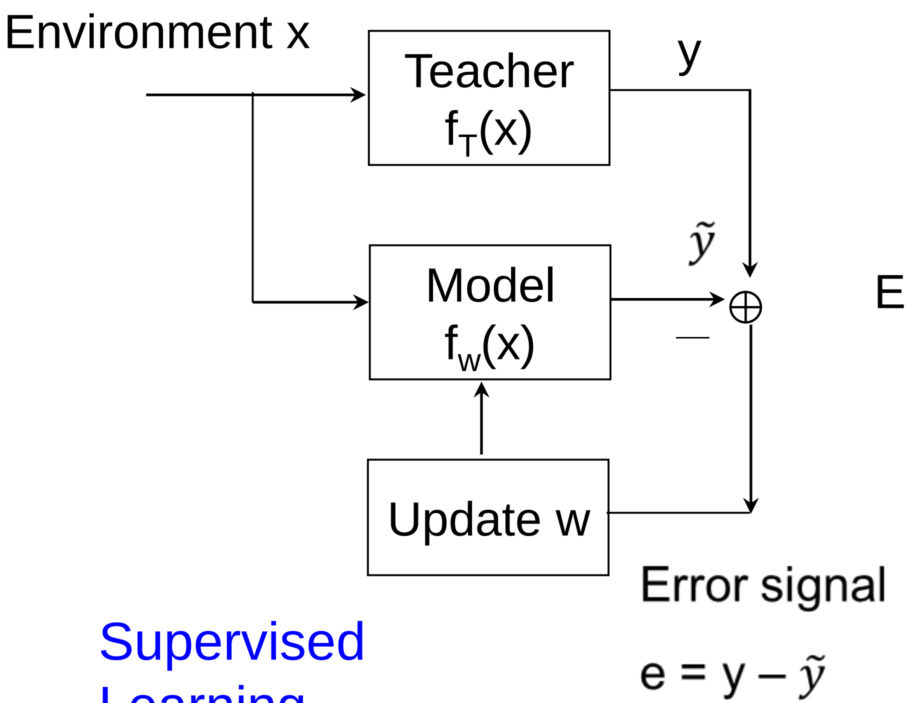
Regression

林伯慎 Bor-shen Lin
bslin@cs.ntust.edu.tw
http:www.cs.ntust.edu.tw/~bslin

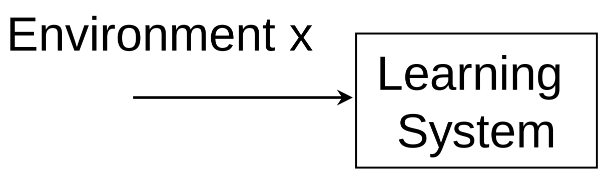
Agenda

- Fundamental of Learning
- Basic Concepts of Regression
- Linear Regression
- Nonlinear regression
- Analysis

Learning Systems

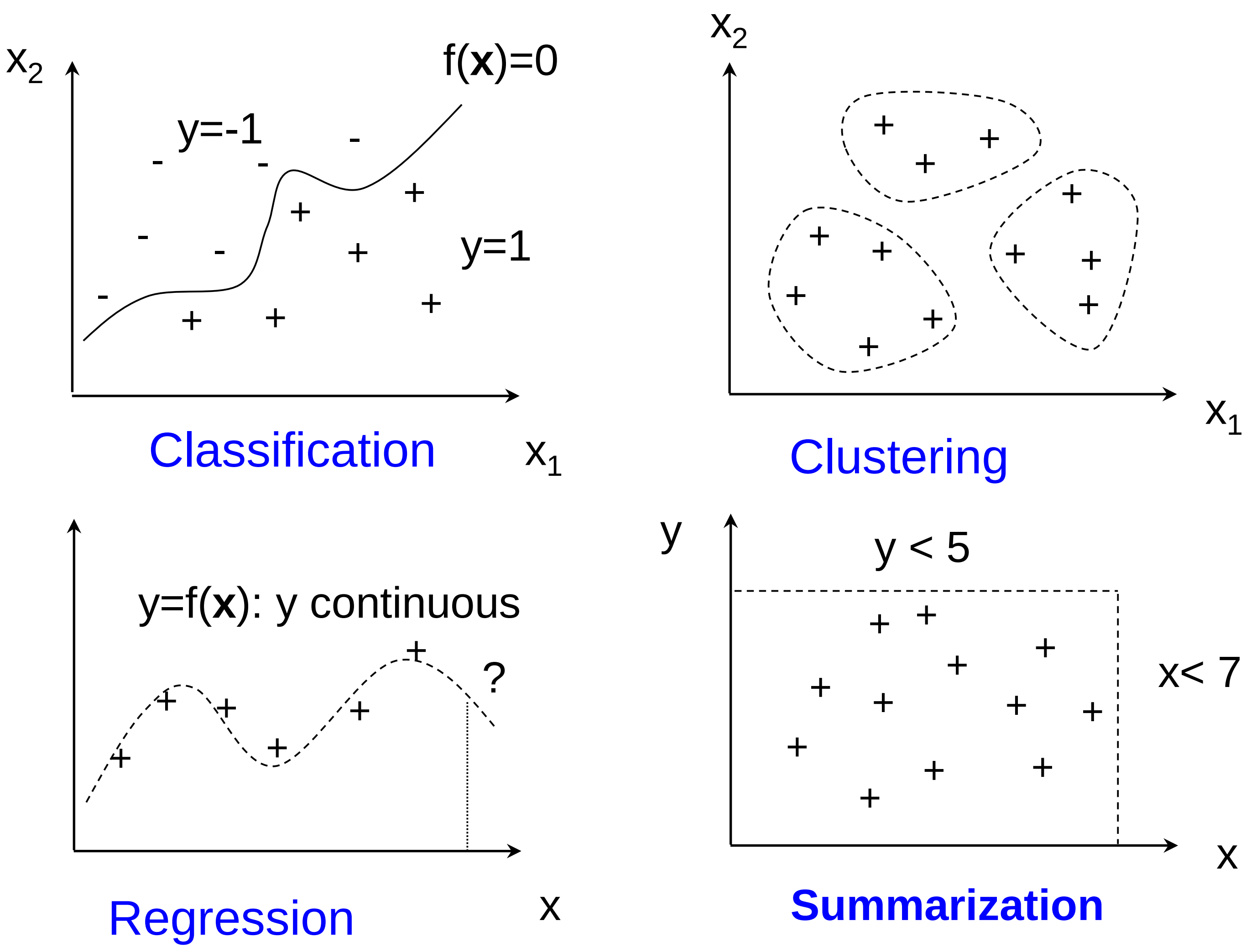


Supervised Learning



Unsupervised Learning

Visualization of Learning Problem



UNSUPERVISED LEARNING

- Stochastic process
  - Find parameters from observations (coin/dice tossing)
  - Discrete/Continuous distribution
  - GMM
- Clustering
  - Find clusters for given training samples

Supervised Learning

- $y = f(x)$
- Regression
  - y is continuous
  - $f_w(\mathbf{x})$ is used as the estimation of y
- Classification
  - y is discrete
  - $f_w(\mathbf{x}) = 0$ as the decision boundary (surface)

    $f_w(\mathbf{x}) = 0$ for $\mathbf{x}$ on the boundary (surface)

    $f_w(\mathbf{x}) > 0$ and $f_w(\mathbf{x}) < 0$ on either side
  - Tests of hypothesis

Regression Analysis

- Objective
  - Determine the best model that can relate the output variable Y to various input variables $X_1, X_2, \ldots, X_n$.
  - X_i's : explanatory (or independent) variables
  - Y: response (or dependent) variable
- Why
  - The output is expensive to measure, but the inputs are not
  - The values of the inputs are known before the output is known → prediction
  - Controlling the input values, we can predict the behavior of corresponding outputs
  - There might be a causal link between the inputs and the output

Regression

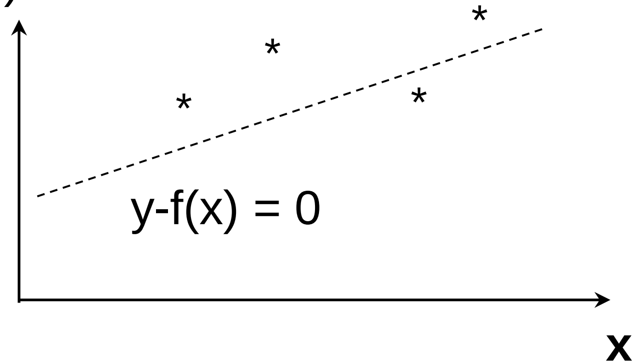
- $\mathbf{x}$: continuous variable(s)
- y: scalar, continuous variables
- Training set $T = \{ (\mathbf{x_i}, y_i) \}$
- Regression model $f_\mathbf{w}$: $y = f_\mathbf{w}(\mathbf{x})$
  - $\mathbf{w}$: a set of parameters
  - e.g. $f_\mathbf{w}(x)=ax^3+bx^2+cx+d \rightarrow \mathbf{w} = (a, b, c, d)$
- Prediction error

  $e(\mathbf{w}) \equiv E_T[(y - f_\mathbf{w}(\mathbf{x}))^2] = \Sigma_i(y_i - f_\mathbf{w}(\mathbf{x_i}))^2$
- Regression: optimization
  - Find $\mathbf{w}^*$ such that $e(\mathbf{w}^*)$ is minimal

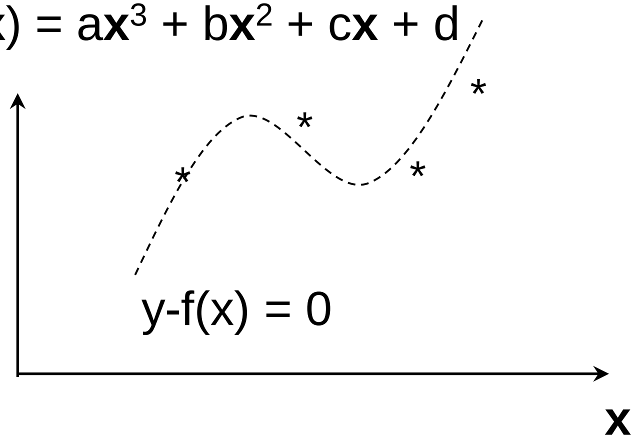
    i.e. $\mathbf{w}^* = \mathrm{argmin}_w\, e(\mathbf{w})$

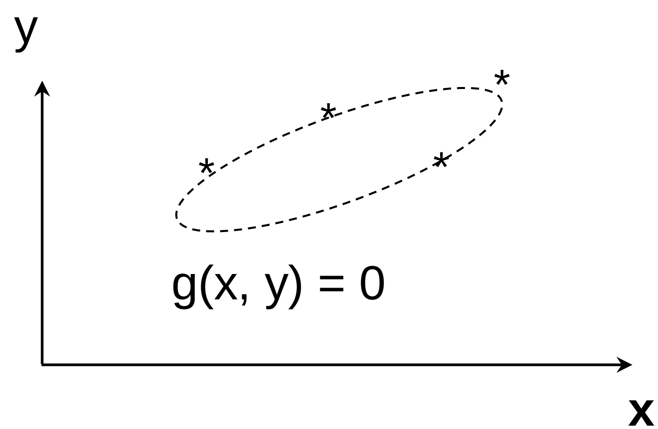
Regression Model

$y = f(\mathbf{x}) = a\mathbf{x} + b$



$y = f(\mathbf{x}) = a\mathbf{x}^3 + b\mathbf{x}^2 + c\mathbf{x} + d$





Linear Regression: Input as Scalar

- Training data: $\{ (x_i, y_i) \}$
- Single input variable x: $y = f(x) = \alpha + \beta x$
- To minimize the sum of square errors

$$e(\alpha, \beta) = \Sigma_i \, e_i^2 = \Sigma_i(y_i\text{-}f(x_i))^2 = \Sigma_i(y_i - \alpha - \beta x_i)^2$$

$$\rightarrow \quad \frac{\partial E}{\partial \alpha} = -2\Sigma_i(y_i\text{-}\alpha\text{-}\beta x_i) = 0$$

$$\frac{\partial E}{\partial \beta} = -2\Sigma_i(y_i\text{-}\alpha\text{-}\beta x_i)x_i = 0$$

$$\rightarrow \quad n\alpha + \beta\Sigma_i x_i = \Sigma_i y_i$$

$$\alpha\Sigma_i x_i + \beta\Sigma_i x_i^2 = \Sigma_i x_i y_i$$

$$\rightarrow \quad \beta^* = [\Sigma_i(x_i - \mu_x)(y_i - \mu_y)] \, / \, \Sigma_i(x_i - \mu_x)^2$$

$$\alpha^* = \mu_y - \beta\mu_x$$

Example

X	Y
1	3
8	9
11	11
4	5
3	2

- $\mu_x = 5, \mu_y = 6$
- $\beta = 1.04$

  $\alpha = 0.8$
- $y = 0.8 + 1.04x$

Linear Regression: Input as Multiple-dimensional Vector

- $y = f(\mathbf{x}) = \alpha+\beta_1 x_1+\beta_2 x_2+\ldots+\beta_n x_n = \boldsymbol{\beta}'\mathbf{X}$

  $\boldsymbol{\beta} \equiv [\alpha, \beta_1, \ldots, \beta_n]'$ $\quad (n+1) \times 1$

  $\mathbf{X} = [1, x_1, x_2, \ldots, x_n]'$ augmented vector
- Training data $\{ (\mathbf{x_j}, y_j) \}$, $j = 1,2,\ldots, m$
  - $\mathbf{x}_j = [x_{j1}, x_{j2}, \ldots, x_{jn}]'$ $\quad n \times 1$
- $\tilde{y}_j = \alpha+\beta_1 x_{j1}+\beta_2 x_{j2}+\ldots+\beta_n x_{jn} \equiv \boldsymbol{\beta} \cdot \mathbf{X_j}$
  - $\mathbf{X_j} \equiv [1, x_{j1}, x_{j2}, \ldots, x_{jn}]'$ $\quad (n+1) \times 1$
  - $\mathbf{y} \equiv [y_1, y_2, \ldots y_m]'$ $\quad m \times 1$
  - $\underline{\mathbf{X}} \equiv [\mathbf{X_1}, \mathbf{X_2}, \ldots, \mathbf{X_m}]'$ $\quad m \times (n+1)$
- To minimize $e(\boldsymbol{\beta}) = (\mathbf{y} - \underline{\mathbf{X}}\boldsymbol{\beta})'(\mathbf{y} - \underline{\mathbf{X}}\boldsymbol{\beta})$

  $\rightarrow \boldsymbol{\beta}^* = (\underline{\mathbf{X}}'\,\underline{\mathbf{X}})^{-1}(\underline{\mathbf{X}}'\,\mathbf{y})$

Nonlinear Regression

- Select the proper transformation of input variables or their combinations
- $Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_1 X_3 + \beta_4 \cdot X_2 X_3$

  $X_4 = X_1 X_3,\ X_5 = X_2 X_3$

  $\rightarrow Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_4 + \beta_4 \cdot X_5$
- $Y = \alpha + \beta_1 \cdot X + \beta_2 \cdot X^2 + \beta_3 \cdot X^3$

  $X_1 = X,\ X_2 = X^2,\ X_3 = X^3$

  $\rightarrow Y = \alpha + \beta_1 \cdot X_1 + \beta_2 \cdot X_2 + \beta_3 \cdot X_3$

Nonlinear Regression

Function	Transformation	Form of simple linear regression
$y = \alpha e^{\beta x}$	$y' = \log(y)$	y' vs. x
$y = \alpha x^{\beta}$	$y' = \log(y)$, $x' = \log(x)$	y' vs. x'
$y = \alpha + \beta(1/x)$	$x' = 1/x$	y vs. x'
$y = x/(\alpha + \beta x)$	$y' = 1/y$, $x' = 1/x$	y' vs. x'

Identify Relevant Input Variables

- Sequential search approach
  - Adding or deleting variables until some overall criterion is satisfied or optimized
- Combinatorial approach
  - All possible combinations
- Criteria
  - Correlation Analysis
  - Analysis of Variance

Correlation Analysis for I/O Variables

- Correlation coefficient r

  $r \equiv S_{xy} / (S_{xx} \cdot S_{yy})^{1/2}$

  $S_{xx} \equiv \Sigma_i(x_i - \mu_x)^2$

  $S_{yy} \equiv \Sigma_i(y_i - \mu_y)^2$

  $S_{xy} \equiv \Sigma_i(x_i - \mu_x)(y_i - \mu_y)$
- $r > 0$: x, y positively correlated

  (studying vs. score)
- $r < 0$: x, y negatively correlated

  (playing vs. score)
- $r = 0$: x, y uncorrelated

  (time for dinner vs. score)

Analysis of Variance (ANOVA)

- Variance $S^2 \equiv \Sigma_i(y_i - f(\mathbf{x}_i))^2 / (m-1)$
  - $f(\mathbf{x_i}) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_n x_{in}$
- F-ratio (F-statistic test) $F = S^2_{new}/S^2_{old}$
  - Variance: estimation error, F: change of error
  - If the variance is not increased a lot by eliminating some variable → the variable might be less important
- Multivariate Analysis (multiple output variables)

$\mathbf{y} = f(\mathbf{x}) = \boldsymbol{\beta}_\mathbf{0} + \boldsymbol{\beta}_\mathbf{1} x_{i1} + \boldsymbol{\beta}_\mathbf{2} x_{i2} + \ldots + \boldsymbol{\beta}_\mathbf{n} x_{in}$

$R \equiv \Sigma_i(\mathbf{y_j} - \mathbf{y_j}')(\mathbf{y_j} - \mathbf{y_j}')^T$

Example of ANOVA

Case	Set of inputs	S^2	F
1	x_1, x_2, x_3	3.56	
2	x_1, x_2	3.98	1.12
3	x_1, x_3	6.22	1.75
4	x_2, x_3	8.34	2.34
5	x_1	9.02	2.27
6	X_2	9.89	2.48

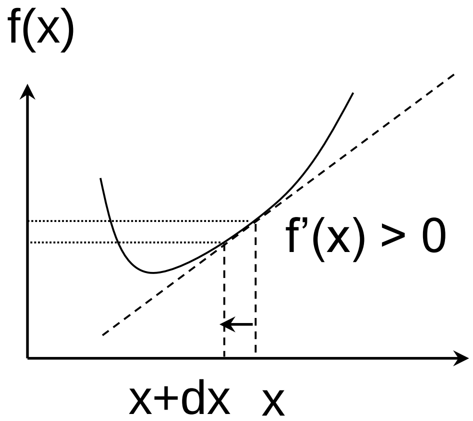
Concept of GPD

- Minimization f(x)
- df = $\frac{\partial f}{\partial x} \cdot dx$

  if dx = -ε $\cdot \frac{\partial f}{\partial x}$

  → df < 0

  x updated along the negative of derivative

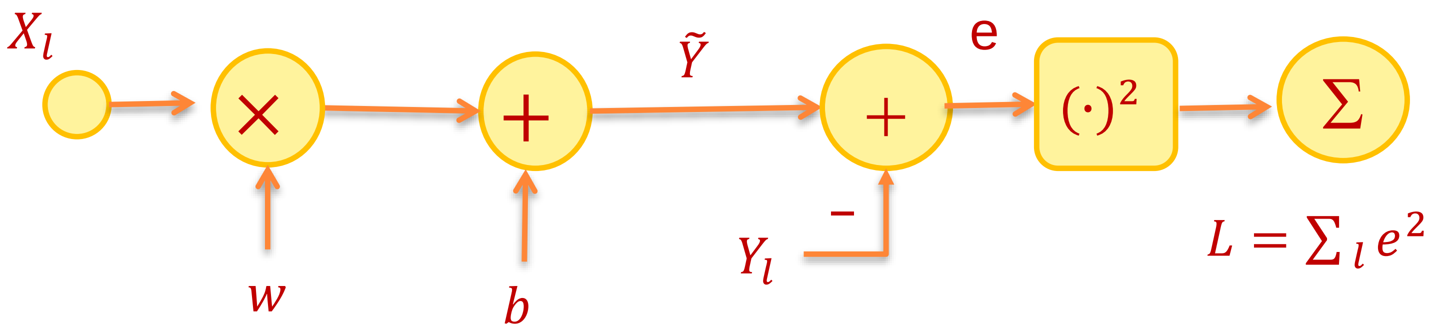


- Reach local minimum
- Select initial x randomly

  → compute dx

  → x' = x + dx
- Converging → df is close to 0

Solve Linear Regression by GD



- 目標函數$L(w, b) = \sum_l (wX_l + b - Y_l)^2$
- $\Delta w = -\varepsilon \frac{\partial L}{\partial w}, \Delta b = -\varepsilon \frac{\partial L}{\partial b}$
- $\frac{\partial L}{\partial w} = \frac{\partial L}{\partial e}\frac{\partial e}{\partial \tilde{Y}}\frac{\partial \tilde{Y}}{\partial w} = (2e)(+1)X$
- $\frac{\partial L}{\partial b} = \frac{\partial L}{\partial e}\frac{\partial e}{\partial \tilde{Y}}\frac{\partial \tilde{Y}}{\partial b} = (2e)(+1)$

GENERALIZATION

- $y = f(\mathbf{x}) \rightarrow g(\mathbf{x}, y) = f(\mathbf{x}) - y = 0$
- Can be generalized as $g(\underline{\mathbf{z}}) = g(\mathbf{x}, y)$ where y is not a function of **x**
  - $\underline{\mathbf{z}}$ contains all input/output variables
  - e.g. $g(x, y) = \frac{(x-x_0)^2}{a^2} + \frac{(y-y_0)^2}{b^2} - 1$
  - $g(\underline{\mathbf{z}}) = 0$ is a curve in 2-D space
  - $g(\underline{\mathbf{z}}) = 0$ is a surface in 3-D space (or higher)
- Parameters of $g(\underline{\mathbf{x}})$ can also be optimized through gradient descent

Solve by GD

- Minimization of square error

  $e(\underline{\mathbf{w}}, \underline{\mathbf{z}}) = g_{\underline{\mathbf{w}}}(\underline{\mathbf{z}})^2$

  - $\underline{\mathbf{w}}$: a set of parameters
  - $\underline{\mathbf{z}}$: input/output variables

- $dw_i = -\varepsilon \frac{\partial e}{\partial w_i} = -\varepsilon(\frac{\partial e}{\partial g})(\frac{\partial g}{\partial w_i})$

  $= -\, 2\varepsilon\, g_{\underline{\mathbf{w}}}(\underline{\mathbf{z}})(\frac{\partial g}{\partial w_i})$

- $d\underline{\mathbf{w}} = -\varepsilon' \cdot g_{\underline{\mathbf{w}}}(\underline{\mathbf{z}}) \cdot \bigtriangledown_{\underline{\mathbf{w}}} g$

  $\bigtriangledown_{\underline{\mathbf{w}}} g \equiv (\frac{\partial g}{\partial w_1}, \frac{\partial g}{\partial w_2}, \ldots, \frac{\partial g}{\partial w_m})$

  $\equiv (g_1(x), g_2(x), \ldots, g_m(x))$

EXAMPLE

- $y = ax^3 + bx^2 + cx + d$, with data $\{ (x_i, y_i) \}$
- Rewritten as

  $g(\underline{\mathbf{z}}) = 1 + w_0 y + w_1 x + w_2 x^2 + w_3 x^3 = 0$

  $\underline{\mathbf{z}} = (x, y)$: x input, y output
- $g_0(\underline{\mathbf{z}}) = \frac{\partial g}{\partial w_0} = y$, $\quad g_1(\underline{\mathbf{z}}) = \frac{\partial g}{\partial w_1} = x$

  $g_2(\underline{\mathbf{z}}) = \frac{\partial g}{\partial w_2} = x^2$, $\quad g_3(\underline{\mathbf{z}}) = \frac{\partial g}{\partial w_3} = x^3$

  $\nabla_{\underline{\mathbf{w}}} g = [y, x, x^2, x^3]$
- $\mathbf{dw} = -\, \varepsilon\, g(\underline{\mathbf{x}})\, \nabla_{\underline{\mathbf{w}}} g$
- $dw_i = -\, \varepsilon\, g(\underline{\mathbf{x}}) \frac{\partial g}{\partial w_i} = -\, \varepsilon\, g(\underline{\mathbf{x}}) g_i(\underline{\mathbf{x}})$