INTRODUCTION TO MACHINE LEARNING

LINEAR REGRESSION

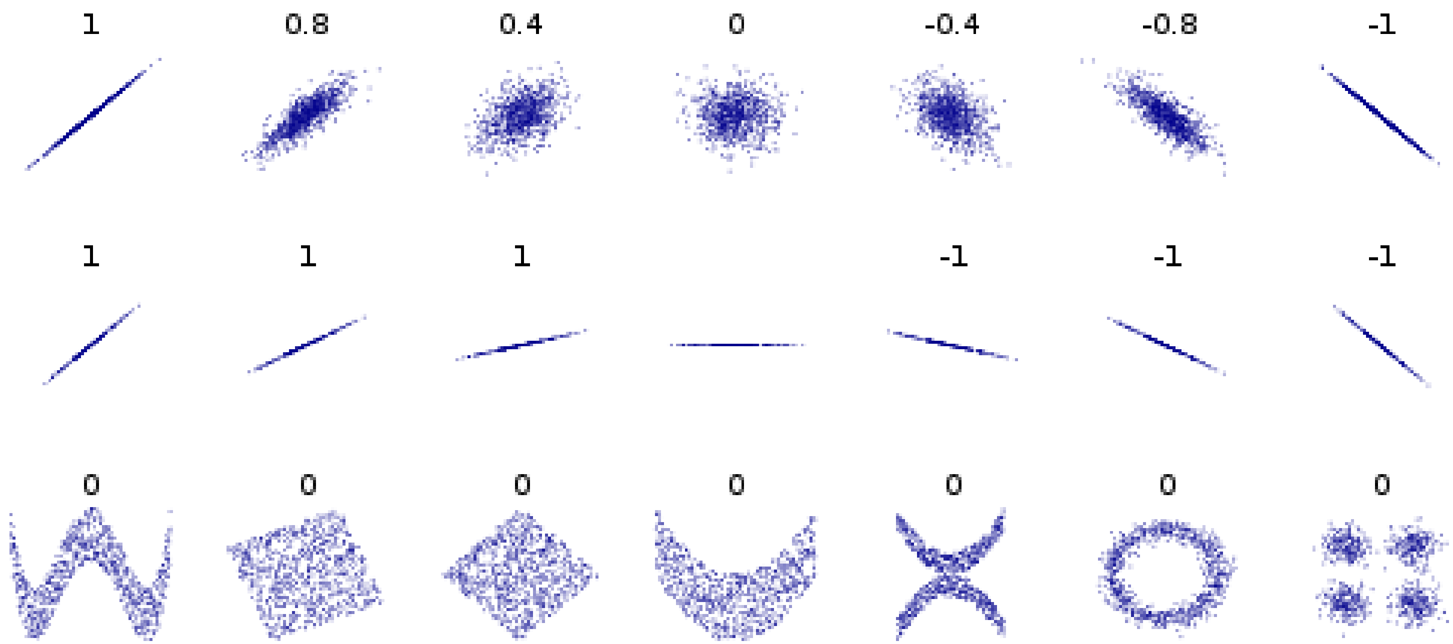
* Some contents are adapted from Dr. Hung Huang and Dr. Chengkai Li at UT Arlington

Mingon Kang, Ph.D.
Department of Computer Science @ UNLV

Correlation (r)

- Linear association between two variables
- Show how to determine both the nature and strength of relationship between two variables
- Correlation lies between +1 to -1
- Zero correlation indicates that there is no relationship between the variables
- Pearson correlation coefficient
  - most familiar measure of dependence between two quantities

Correlation (r)



Correlation (r)

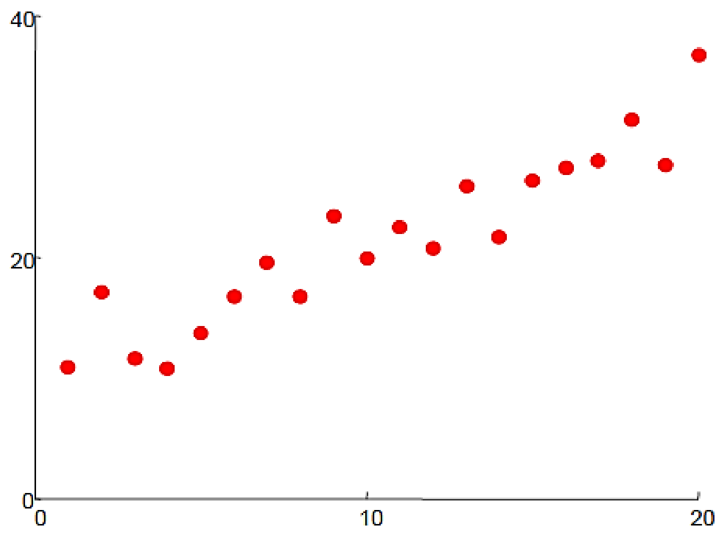
$$\rho_{X,Y} = \mathrm{corr}(X,Y) = \frac{\mathrm{cov}(X,Y)}{\sigma_X \sigma_Y} = \frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X \sigma_Y},$$

where *E* is the expected value operator, *cov(,)* means covariance, and *corr(,)* is a widely used alternative notation for the correlation coefficient

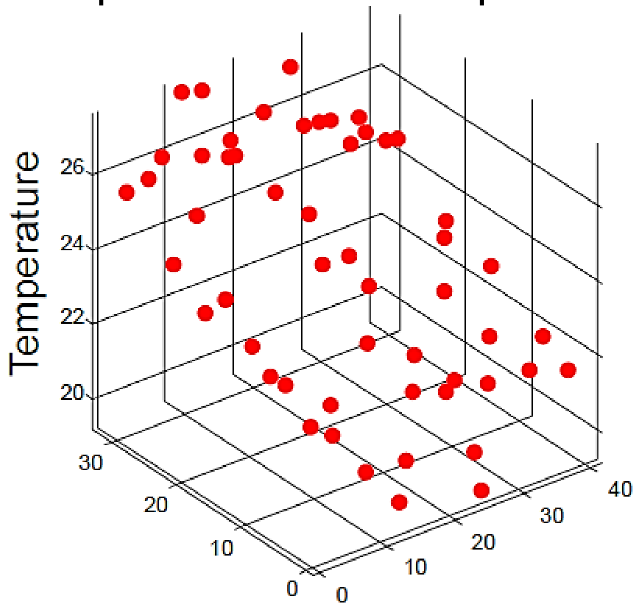
Reference: https://en.wikipedia.org/wiki/Correlation_and_dependence

Linear Regression

Samples with ONE independent variable



Samples with TWO independent variables

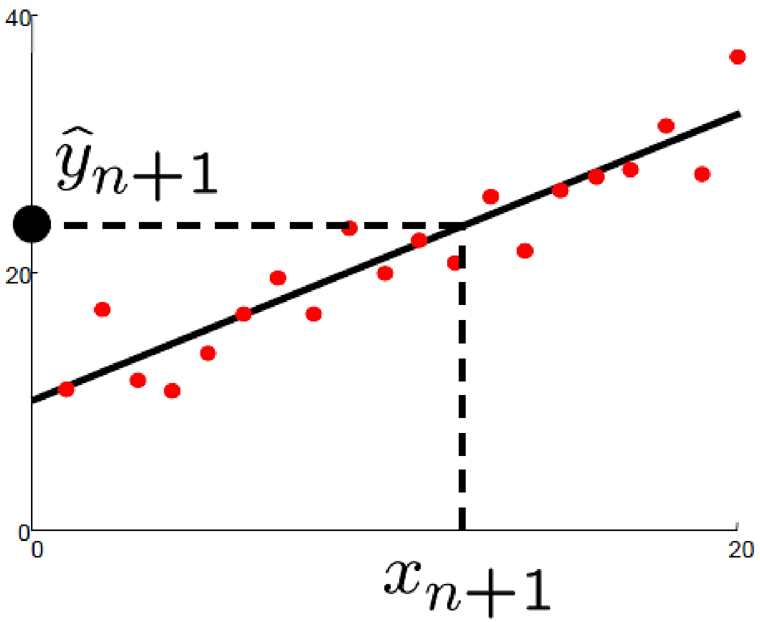


Given examples $(x_i, y_i)_{i=1 \ldots n}$

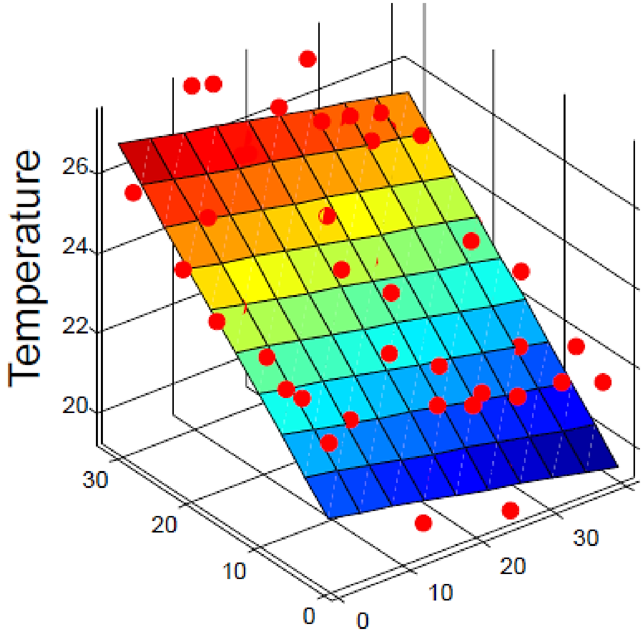
Predict y_{n+1} given a new point x_{n+1}

Linear Regression

Samples with ONE independent variable



Samples with TWO independent variables



Linear Regression

- How to represent the data as a vector/matrix
  - We assume a model:

$$\mathbf{y} = \mathrm{b}_0 + \mathbf{bX} + \epsilon,$$

where b_0 and $\mathbf{b}$ are *intercept* and *slope*, known as *coefficients* or *parameters*. ϵ is the error term (typically assumes that $\epsilon \sim N(\mu, \sigma^2)$

Linear Regression

- Simple linear regression
  - A single independent variable is used to predict
- Multiple linear regression
  - Two or more independent variables are used to predict

Linear Regression

- How to represent the data as a vector/matrix
  - Include bias constant (intercept) in the input vector
    - $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$, $\mathbf{y} \in \mathbb{R}^{n}$, $\mathbf{b} \in \mathbb{R}^{p+1}$, and $\mathbf{e} \in \mathbb{R}^{n}$

$$\mathbf{y} = \mathbf{X} \cdot \mathbf{b} + \mathbf{e}$$

$$\mathbf{X} = \{\mathbf{1}, \mathbf{x_1}, \mathbf{x_2}, \dots, \mathbf{x_p}\}, \mathbf{b} = \{b_0, b_1, b_2, \dots, b_p\}^{\mathrm{T}}$$
$$\mathbf{y} = \{y_1, y_2, \dots, y_n\}^{\mathrm{T}}, \mathbf{e} = \{e_1, e_2, \dots, e_n\}^{\mathrm{T}}$$

$\cdot$ is a dot product

equivalent to

$$\mathrm{y}_i = 1 * \mathrm{b}_0 + x_{i1}\mathrm{b}_1 + x_{i2}\mathrm{b}_2 + \cdots + x_{ip}\mathrm{b}_\mathrm{p} \ (1 \le i \le n)$$

Linear Regression

- Find the optimal coefficient vector **b** that makes the most similar observation

$$\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix} \begin{bmatrix} b_0 \\ \vdots \\ b_p \end{bmatrix} + \begin{bmatrix} e_1 \\ \vdots \\ e_n \end{bmatrix}$$

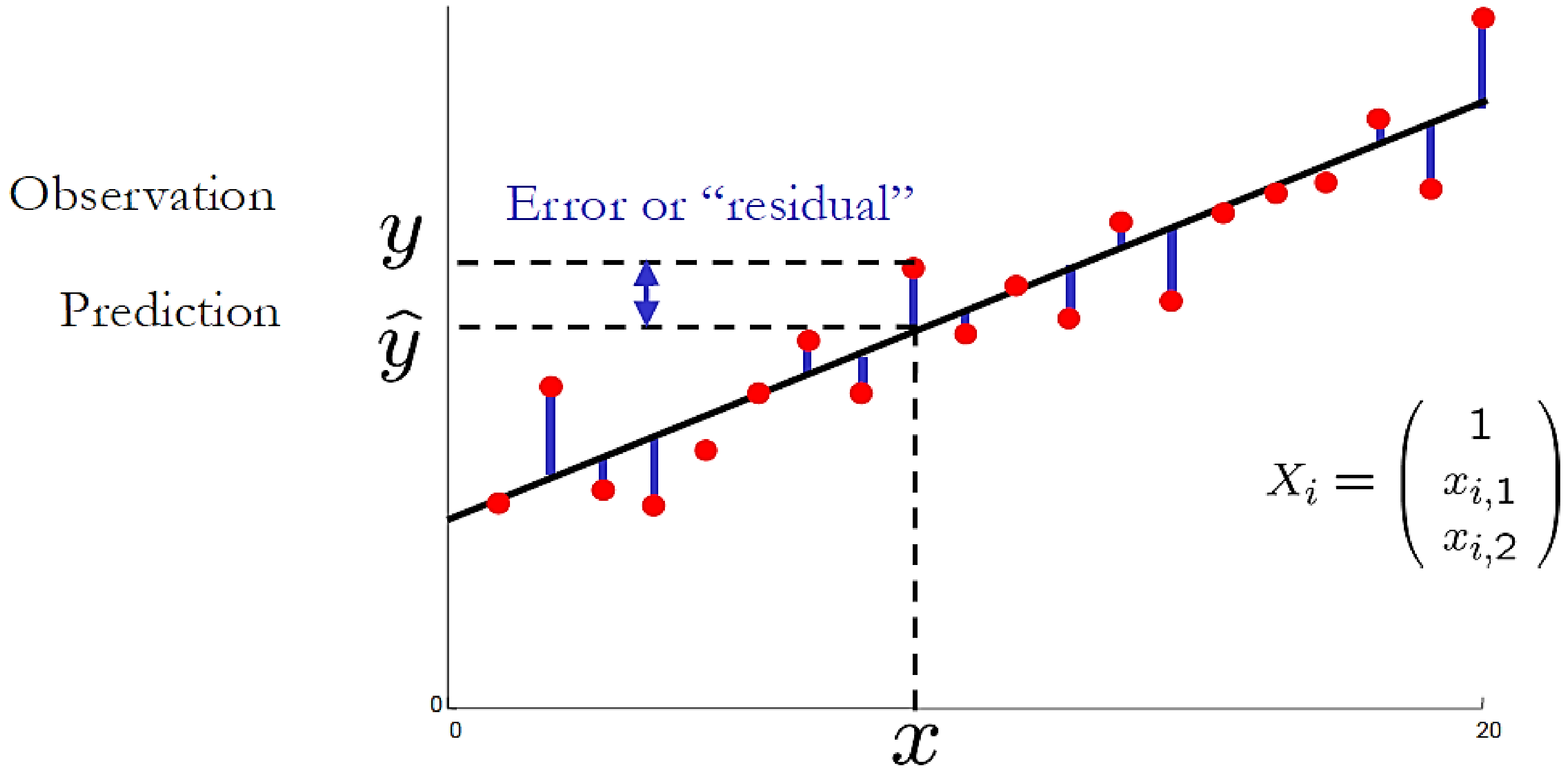
Ordinary Least Squares (OLS)

$$\mathbf{y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

- Estimate the unknown parameters ($\mathbf{b}$) in linear regression model
- Minimizing the sum of the squares of the differences between the observed responses and the predicted by a linear function

$$\text{Sum squared error} = \sum_{i=1}^{n} (y_i - \mathbf{x}_{i*}\mathbf{b})^2$$

Ordinary Least Squares (OLS)



Optimization

- Need to minimize the error

$$\min J(\mathbf{b}) = \sum_{i=1}^{n} (y_i - \mathbf{x}_{i,*}\mathbf{b})^2$$

- To obtain the optimal set of parameters (**b**), derivatives of the error w.r.t. each parameters must be zero.

Optimization

$$
\begin{aligned}
J = \mathbf{e}^{\mathrm{T}}\mathbf{e} &= (\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b}) \\
&= (\mathbf{y}' - \mathbf{b}'\mathbf{X}')(\mathbf{y} - \mathbf{X}\mathbf{b}) \\
&= \mathbf{y}'\mathbf{y} - \mathbf{y}'\mathbf{X}\mathbf{b} - \mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b} \\
&= \mathbf{y}'\mathbf{y} - 2\mathbf{b}'\mathbf{X}'\mathbf{y} + \mathbf{b}'\mathbf{X}'\mathbf{X}\mathbf{b}
\end{aligned}
$$

$$\frac{\partial \mathbf{e}'\mathbf{e}}{\partial \mathbf{b}} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b} = 0$$

$$(\mathbf{X}'\mathbf{X})\mathbf{b} = \mathbf{X}'\mathbf{y}$$

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$$

Matrix Cookbook: https://www.math.uwaterloo.ca/~hwolkowi/matrixcookbook.pdf

The Happiness Formula

- **H = (G + DH + C + 3R) / 6**
  - Happiness ("H") is equal to
    - your level of *Gratitude* ("G") +
    - the degree to which you are living consistent with your own personal *Definition of Happiness* ("DH") +
    - how much you *Contribute* to others ("C") +
    - your success in what I call the *3 R's* of happiness ("3R")

Ref: http://www.behappy101.com/happiness-formula.html

Linear regression with categorical variables

- We assumed that all variables are continuous variables
- Categorical variables:
  - Ordinal variables - Encode data with continuous values
    - Evaluation: Excellent (5), Very good (4), Good (3), Poor (2), Very poor (1)
  - Nominal variables – Use dummy variables
    - Department: Computer, Biology, Physics

	Computer	Biology	Physics
Computer	1	0	0
Biology	0	1	0
Physics	0	0	1

Linear regression for classification

- For binary classification
  - Encode class labels as $y = \{0, 1\}\ or\ \{-1, 1\}$
  - Apply OLS
  - Check which class the prediction is closer to
    - If class 1 is encoded to 1 and class 2 is -1.

$$\begin{aligned} &class\ 1 \quad if\ f(x) \geq 0 \\ &class\ 2 \quad if\ f(x) < 0 \end{aligned}$$

  - Linear models are NOT optimized for classification
- Logistic regression

Linear regression for classification

- ROC for classification

$$f(x) \underset{<}{\geq} \lambda$$

If $f(x)$ is less than λ, class 1. Otherwise class 2.

How can we know the optimal λ ?

- Let's revisit EVALUATION.

Linear regression for classification

- Multi-label classification
  - Encode classes label as:

	Computer	Biology	Physics
Computer	1	0	0
Biology	0	1	0
Physics	0	0	1

  - Perform linear regression multiple times for each class
  - Consider **y** and **b** as matrix

Assumptions in Linear regression

- Linearity of independent variable in the predictor
  - normally good approximation, especially for high-dimensional data
- Error has normal distribution, with mean zero and constant variance
  - important for tests
- Independent variables are independent from each other
  - Otherwise, it causes a **multicollinearity** problem; two or more predictor variables are highly correlated.
  - Should remove them

Think more!

	Feature 1	Feature 2	Feature 3	Feature 4
Coefficient	5.2	0.1	-6.6	0

- How can we interpret this model?
- What is the most useless feature?
  - Is it always useless to explain the dependent variable?
- What do negative coefficients represent?
- What is the most informative feature?

Different views between Statistics and CS

- In Statistics, description of the model is often more important.
  - Which variables are more informative and reliable to describe the responses? → p-values
  - How much information do the variables have?
- In Computer Science, the accuracy of prediction and classification is more important.
  - How well can we predict/classify?

Discussion

- What if data is imbalanced data?
- Why does OLS take squares instead of absolute values?