Course Section 2

Xiao-Wei CAO
acdoge.cao@gmail.com

Chat Freely

- Last Week
  - Too many MATH ☹
  - Traditional Computer Vision
  - Signals and Systems ☹
  - Programming Practice ☺
  - Reading List (not paper now)
- Other things...
  - Lecture / Section / Tutorial
  - What’s your expectation
  - Adaptive adjustment

“学不动啦！！！”

Andrew Ng (My favor teacher)
Google Brain / Baidu Research
Co-founder of Coursera
Founder of Landing AI and deeplearning.ai

What happened in this week

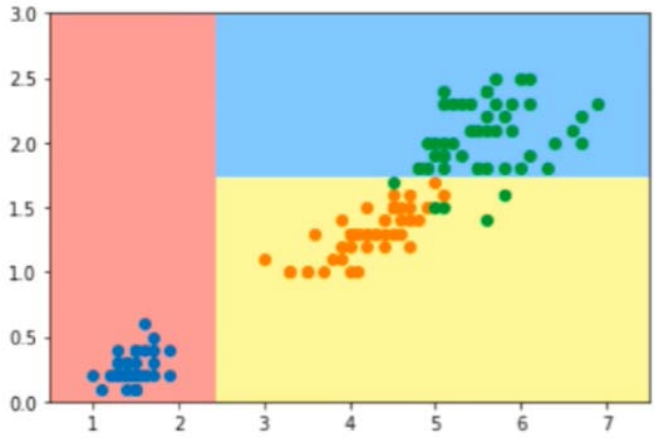
- 旷视6号员工范浩强：高二开始实习，“兼职”读姚班，25岁在CVPR斩获第四个世界第一 – 量子位

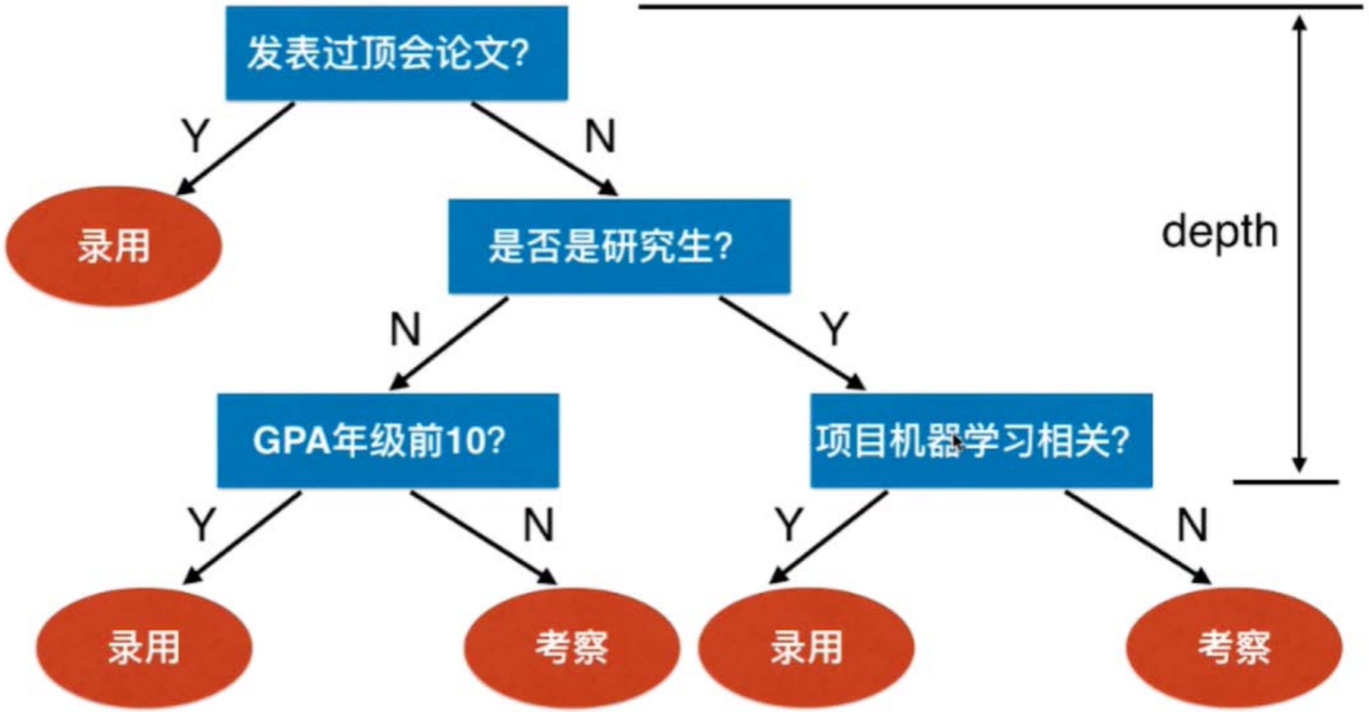
Contents

- Quick Review
  - Decision Tree
  - Linear Models
- Quiz Discussion
- Supplement
  - More Linear Models
  - Machine Learning Again
- Warm Up for Next Week

Decision Tree

- Example: ML Engineer HR
- Branch - Feature
- Leaf Node – Decision
- Decision Boundary





via @Yubo Liu (liuyubobobo)

Decision Tree – How to Build

- Information Entropy

$$H = -\sum_{i=1}^{k} p_i \log(p_i)$$

$$\left\{\frac{1}{3},\frac{1}{3},\frac{1}{3}\right\}\ H = -\frac{1}{3}\log\left(\frac{1}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)-\frac{1}{3}\log\left(\frac{1}{3}\right)=1.0986$$

$$\left\{\frac{1}{10},\frac{2}{10},\frac{7}{10}\right\}\ H = -\frac{1}{10}\log\left(\frac{1}{10}\right)-\frac{2}{10}\log\left(\frac{2}{10}\right)-\frac{7}{10}\log\left(\frac{7}{10}\right)=0.8108$$

$$\{1,0,0\}\ H = -1\cdot\log(1)=0$$

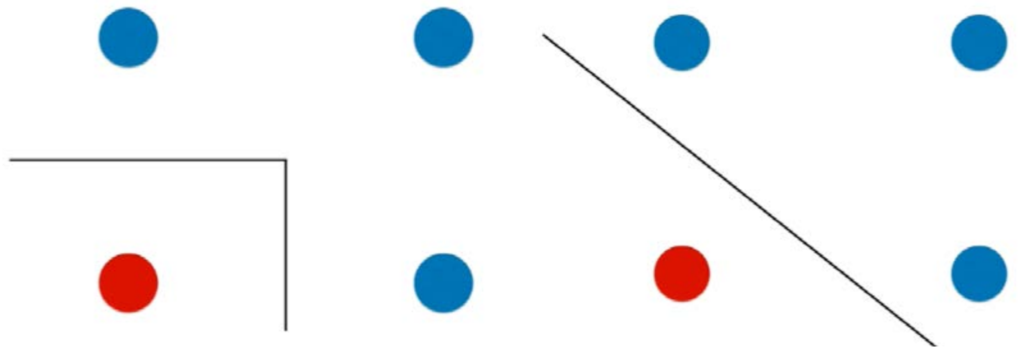
- System determinism becomes stronger – H ↓
- Gini Coefficient (sklearn default) - CART

$$G = 1-\sum_{i=1}^{k} p_i^2$$

Simulate by coding

Decision Tree – Pruning

- m samples, n features
- Complexity:
  - prediction: $O(\log m)$
  - training: $O(n * m * \log m)$
- Overfitting (similar to KNN)



More Decision Tree

- Tree > Forest
  - Ensemble Methods
- Random Forest

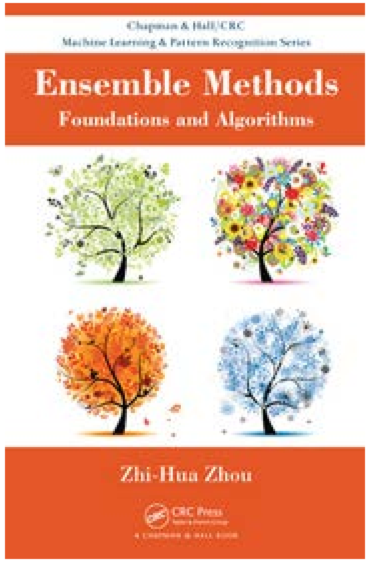
周志华 中文简历 Brief CV

Zhi-Hua Zhou can be pronounced simply as [Jihua Joe]

Professor, Computer Science and Artificial Intelligence, Nanjing University, China

Fellow of the ACM, AAAI, AAAS, IEEE, IAPR, IET/IEE, CCF, CAAI

周志华 著.机器学习,北京:清华大学出版社,2016. (ISBN 978-7-302-42328-7)

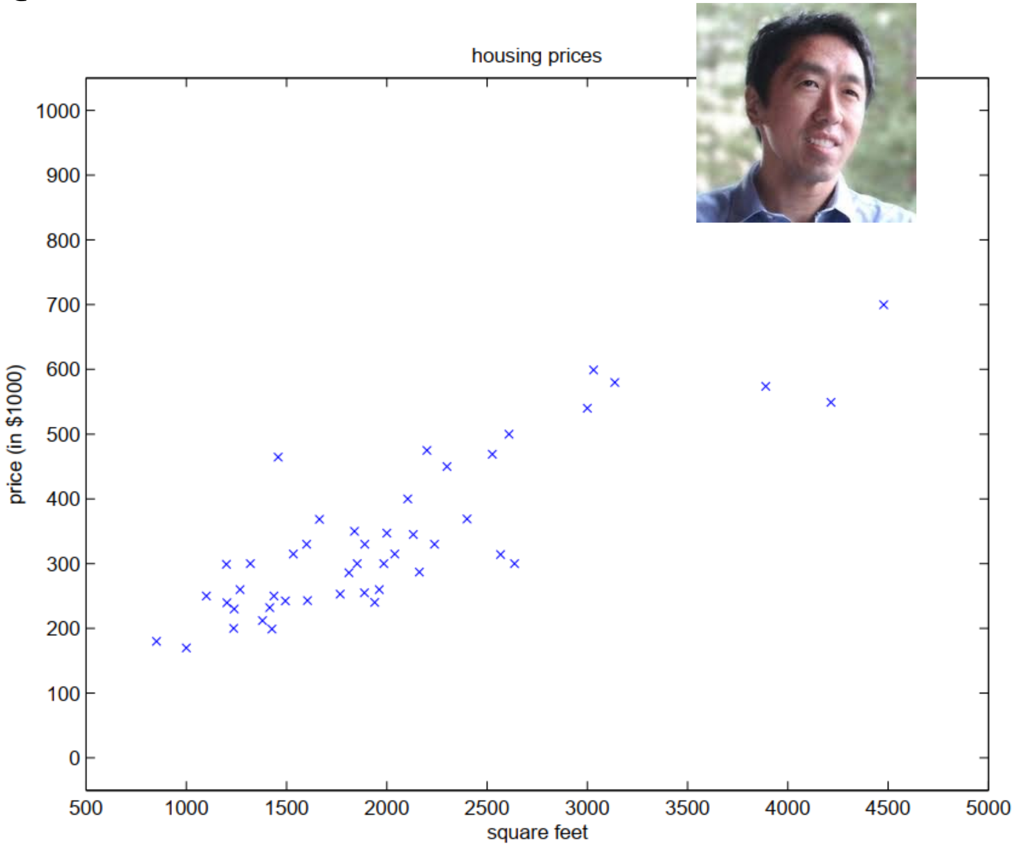


Z.-H. Zhou. Ensemble Methods: Foundations and Algorithms, Boca Raton, FL: Chapman & Hall/CRC, 2012. (ISBN 978-1-439-830031)

Linear Models

- Easy to understand and implement
  - Least Squares Regression
- Nonlinear Model Basis
- Interpretable (Why it works)
- Machine Learning Concepts

We'll talk more later.



http://cs229.stanford.edu/

Contents

MNIST Dataset

- Official Webpage
- Benchmark
- Download & Load

train-images-idx3-ubyte.gz: training set images (9912422 bytes)
train-labels-idx1-ubyte.gz: training set labels (28881 bytes)
t10k-images-idx3-ubyte.gz: test set images (1648877 bytes)
t10k-labels-idx1-ubyte.gz: test set labels (4542 bytes)

http://yann.lecun.com/exdb/mnist/

Dataset loading utilities

6. Dataset loading utilities

- 6.1. General dataset API
- ▶ 6.2. Toy datasets
- ▶ 6.3. Real world datasets
- ▶ 6.4. Generated datasets
- ▶ 6.5. Loading other datasets

6. Dataset loading utilities

The `sklearn.datasets` package embeds some small toy datasets as introduced in the Getting Started section.

This package also features helpers to fetch larger datasets commonly used by the machine learning community to benchmark algorithms on data that comes from the 'real world'.

To evaluate the impact of the scale of the dataset (`n_samples` and `n_features`) while controlling the statistical properties of the data (typically the correlation and informativeness of the features), it is also possible to generate synthetic data.

https://scikit-learn.org/stable/user_guide.html

Data preprocessing

- Import libraries
- Read data
- Checking for missing values
- Checking for categorical data
- Standardize the data
- PCA transformation
- Data splitting

Learn from Scikit-Learn Docs

- Ordinary Least Squares(OLS)
  - Ridge / Lasso Regression
- Decision Trees
  - ID3, C4.5, C5.0 and CART
- Support Vector Machines

1.1. Generalized Linear Models

https://scikit-learn.org/stable/supervised_learning.html

Try reading in English

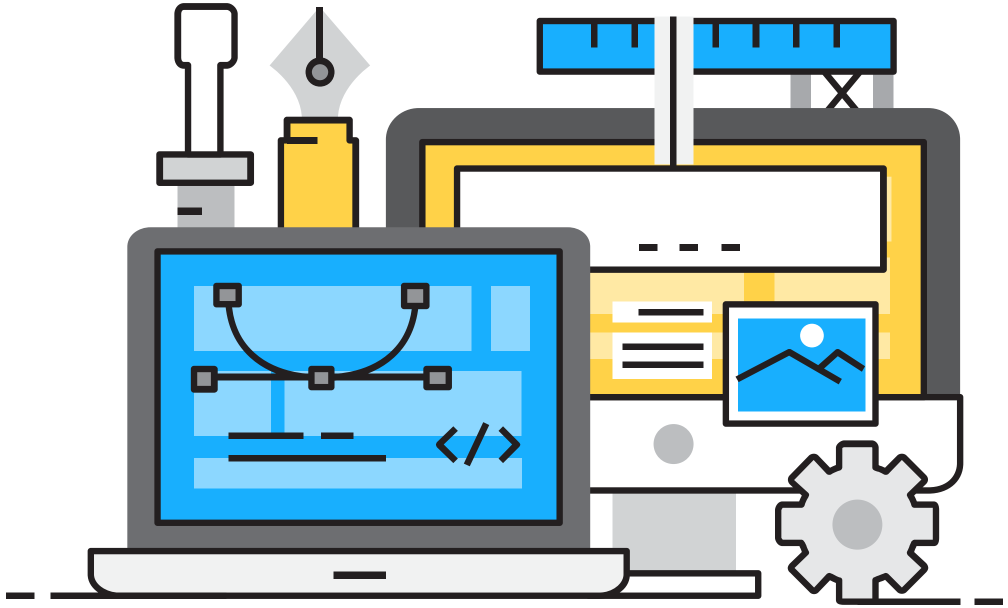
- Just few weeks
- Machine Translation
  - Google Translation

Reinventing the wheel

- Using exist tools:
  - grasp basic concept
- Learn from the source code:
  - understand details
  - avoid complacence
- Why must know details?
  - better to select (analysis)
  - better to create (inspiration)
  - learning how to learn (meta)

abstract ← concrete
- analogy
- example
- explanation

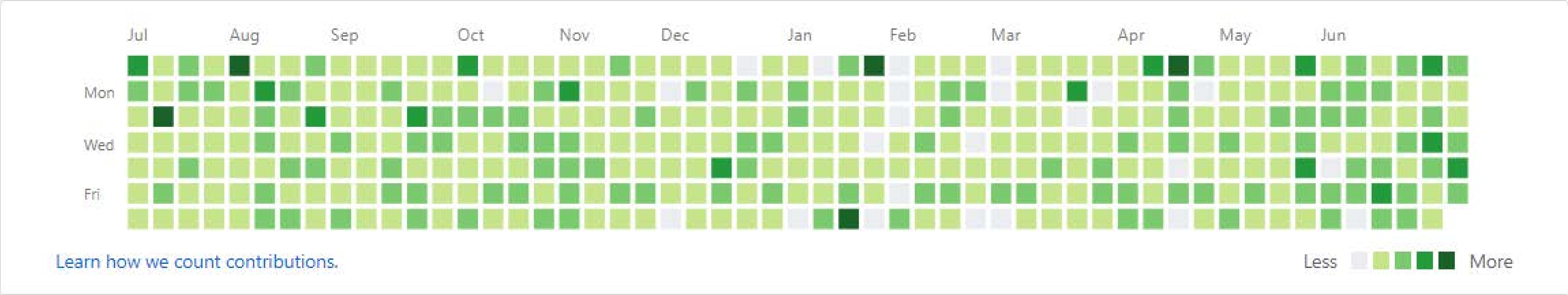
abstract
↓
concrete
↓
abstract
↓



Example: Ruan Yifeng

- Beginner Level
  - Stay Focused, Keep Shipping.
  - Build Early, Build Always.
  - Improve yourself,
  - Write solid/simple/stupid code.

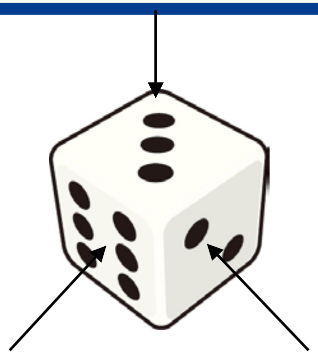
1,240 contributions in the last year



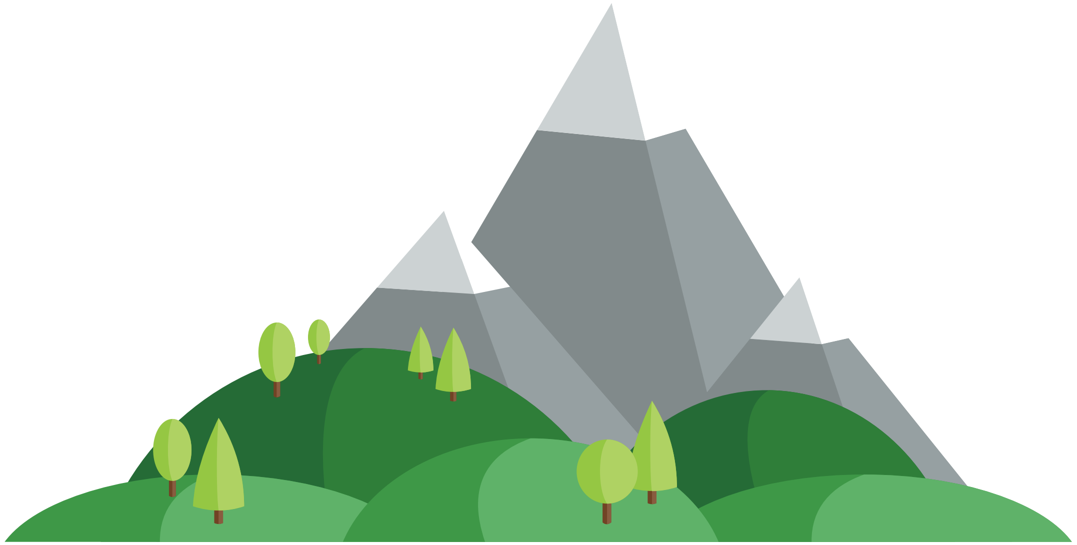
https://www.ruanyifeng.com/blog/

But... Learn to Question

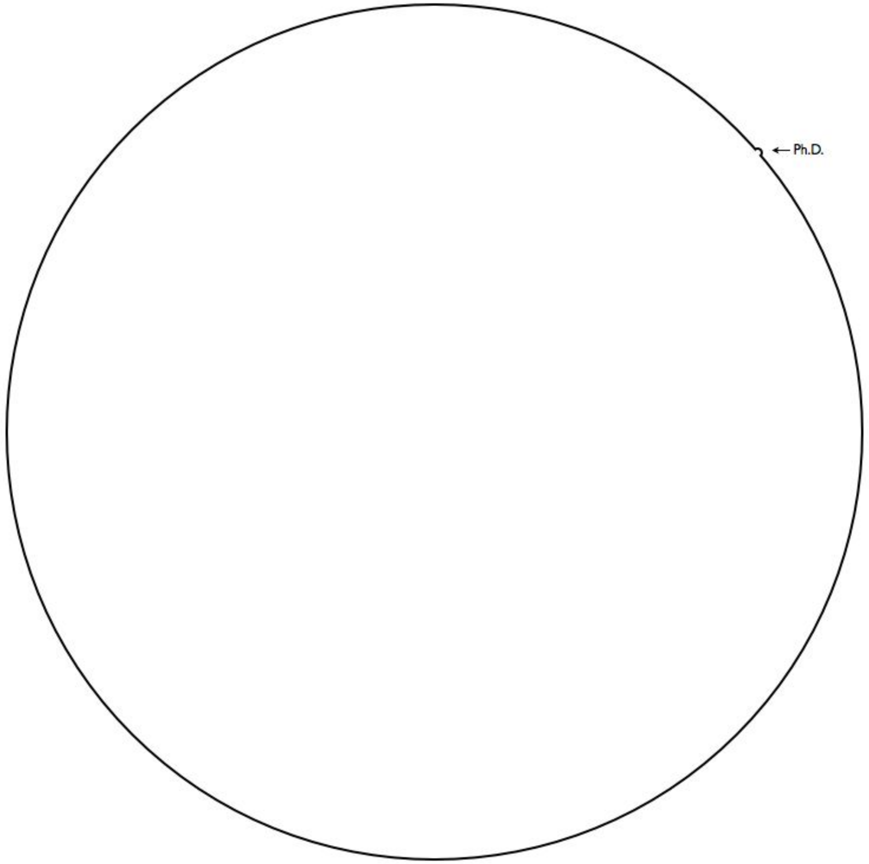
- Question and Challenge
  - Knowledge ≠ Truth
  - Independent thinking
  - Different perspectives
  - Find consensus
- Example :
  - Is what I'm saying now right
- Philosophy:
  - Liar paradox



3 or 6 or 2 ???
(1 4 5 ???)



The illustrated guide to a Ph.D.



http://matt.might.net/articles/phd-school-in-pictures/

Break

5 mins

Contents

Linear Models

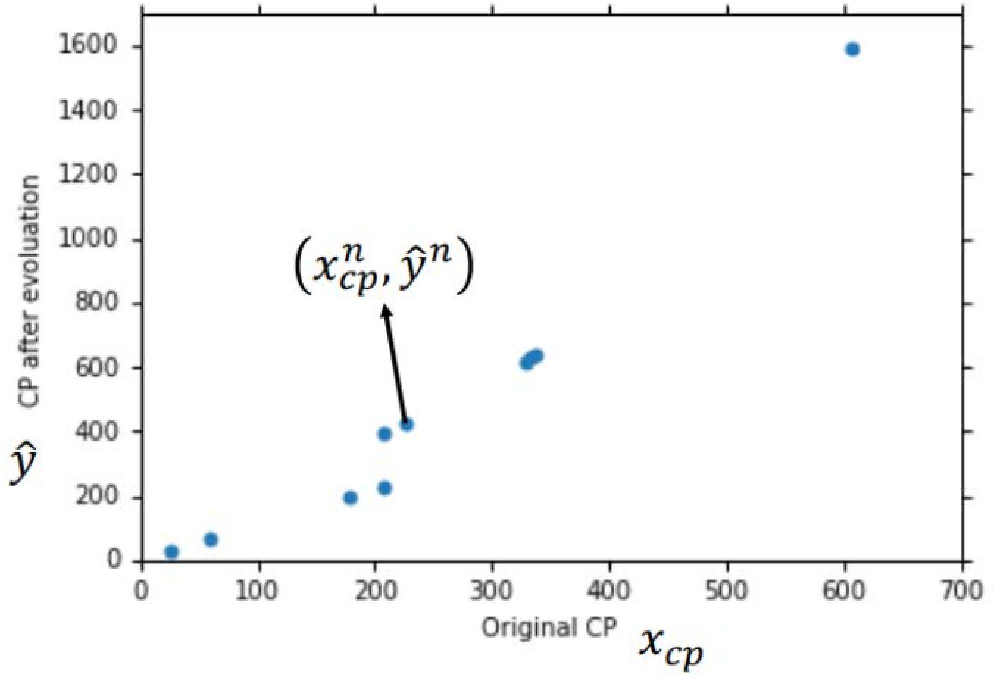
- Linear Regression
- Gradient Descent
- Error Analysis (for Engineer)
- Linear Classification
- High Level View
  - Different explain on LR
  - Development / History

机器学习基础思想(线性模型):

- 线性回归与梯度下降
- 梯度下降细节与技巧
- 偏差与方差——误差从何而来?
- 线性分类与逻辑回归
- 机器学习思想比较
- 机器学习模型发展
- 数学思维强化，感受抽象的力量:
  - 高屋建瓴之线性回归
  - 高屋建瓴之线性分类 [计划中]
  - 正态分布 [计划中]
  - 指数族分布 [计划中]

Linear Regression

- Step 1 – Model
  - Linear Model: $y = f(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x}$
- Step 2 – Evaluate
  - Loss Function: $L(f) = \sum^{n} (\hat{y} - f(x))^2$
- Step 3 - Optimize



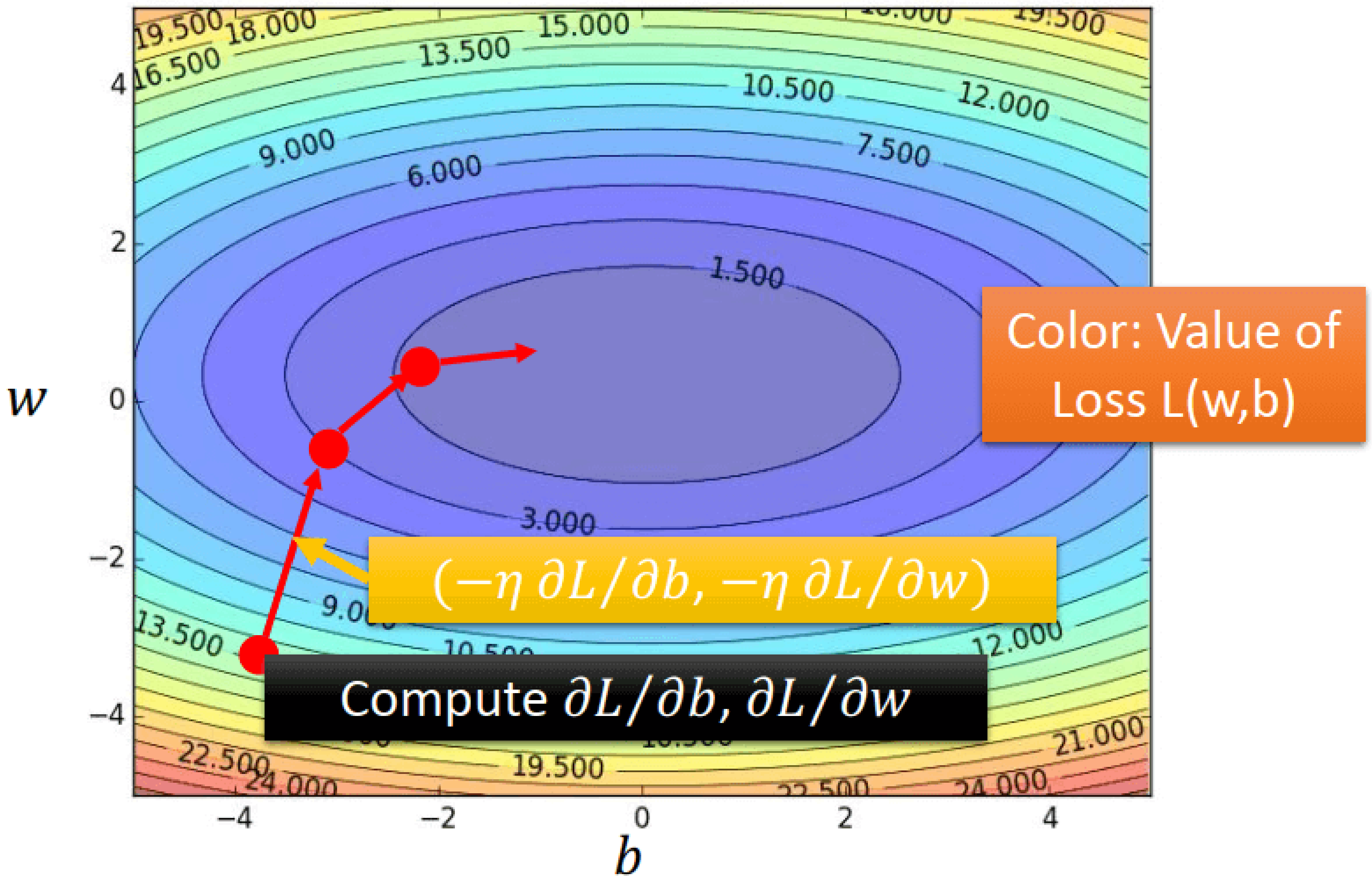
$$f^* = \arg\min_f L(f)$$

$$w^*, b^* = \arg\min_{w,b} L(w, b)$$

$$= \arg\min_{w,b} \sum_{n=1}^{10} \left(\hat{y}^n - (b + w \cdot x_{cp}^n)\right)^2$$

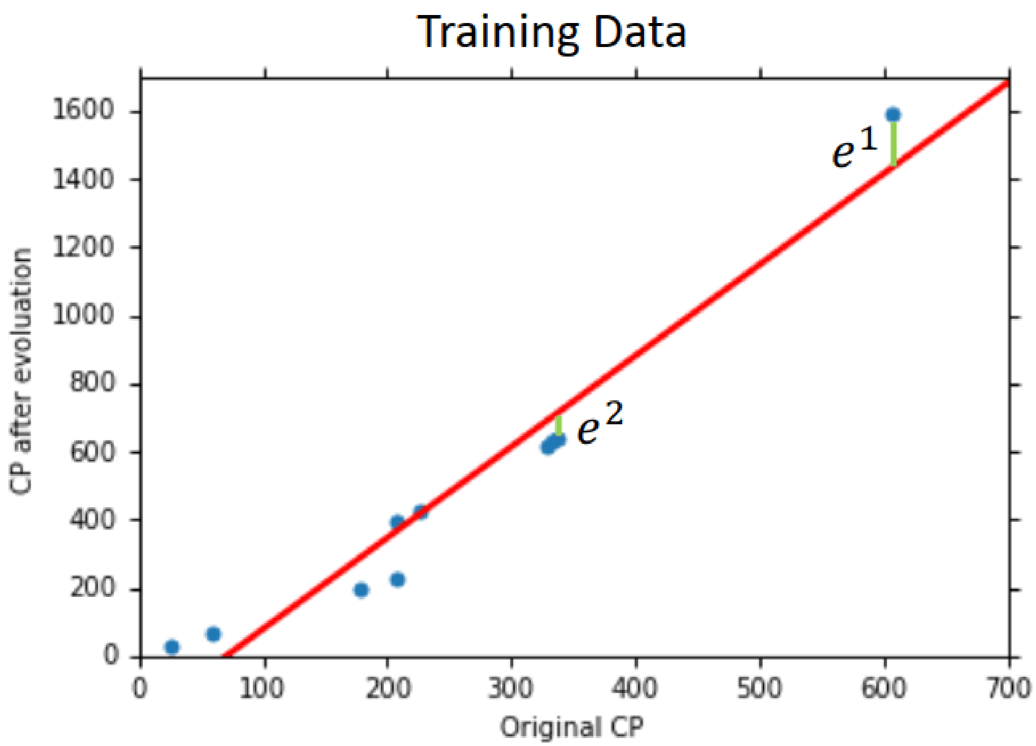
https://accepteddoge.com/cnblogs/regression-and-gradient-descent

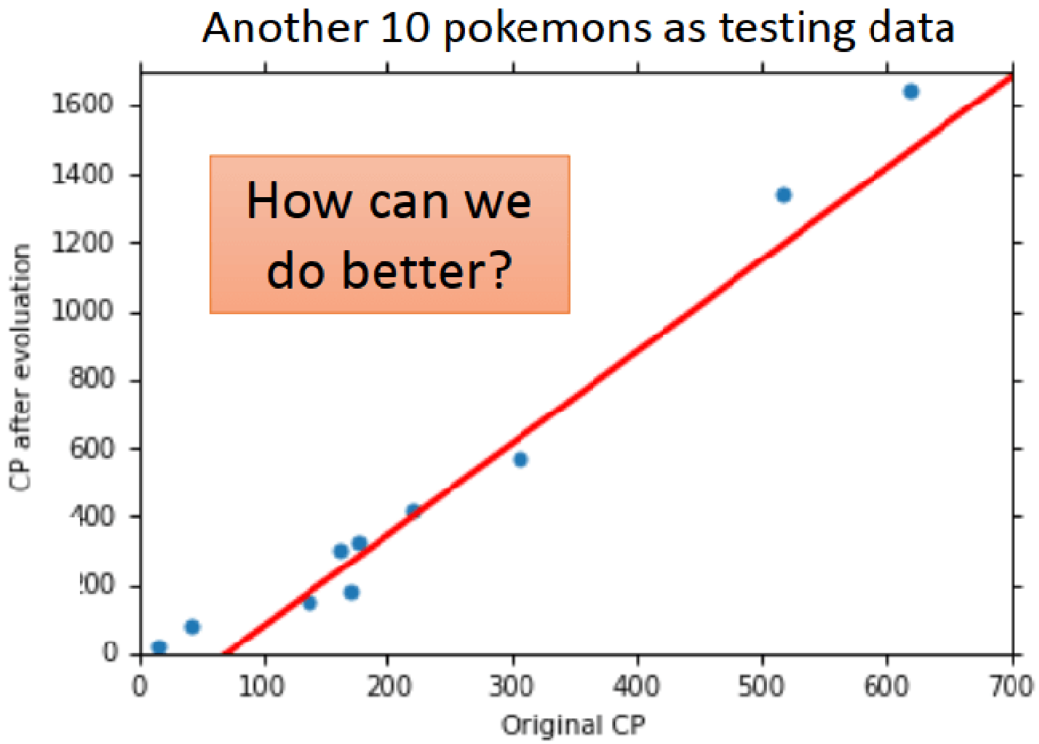
Gradient Descent



https://accepteddoge.com/cnblogs/gradient-descent

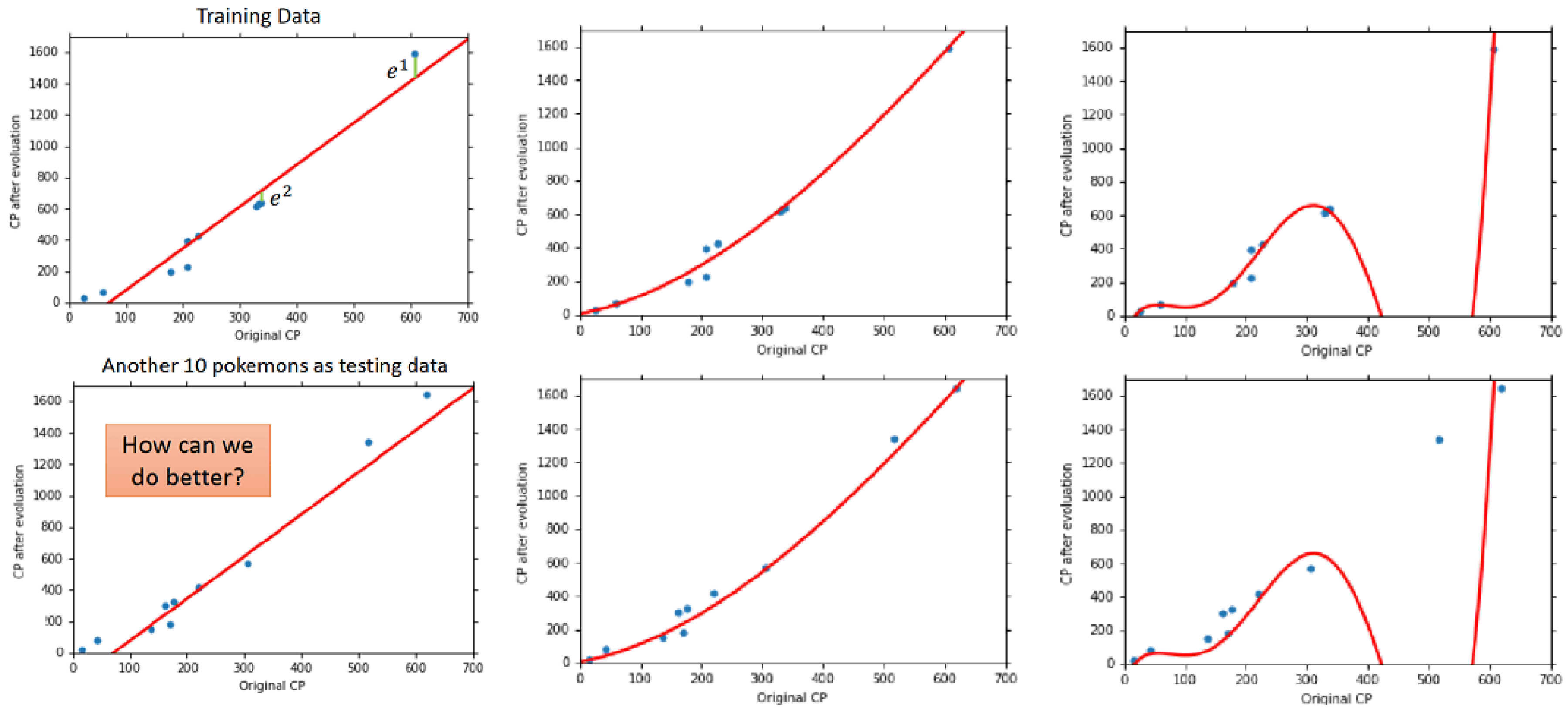
Generalization





https://accepteddoge.com/cnblogs/regression-and-gradient-descent

Overfitting



$$y = b + w \cdot x_{cp}$$

$$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3$$

$$y = b + w_1 \cdot x_{cp} + w_2 \cdot (x_{cp})^2 + w_3 \cdot (x_{cp})^3 + w_4 \cdot (x_{cp})^4 + w_5 \cdot (x_{cp})^5$$

Regularization

- In Linear Models:

$$L = \sum_n \left(\hat{y}^n - \left(b + \sum w_i x_i\right)\right)^2 + \lambda \sum (w_i)^2$$

- In Paper:

$$\arg\min_{\boldsymbol{w}} J(\boldsymbol{w}) = [L(\boldsymbol{w}) + \lambda P(\boldsymbol{w})]$$

https://accepteddoge.com/cnblogs/regression-and-gradient-descent

Closed-form Solution

- Loss Function:

$$E_{(w,b)} = \sum_{i=1}^{m} \left(y_i - f\left(x_i\right)\right)^2 = \sum_{i=1}^{m} \left(y_i - wx_i - b\right)^2$$

- Parameter Estimation:

$$b = \frac{1}{m} \sum_{i=1}^{m} \left(y_i - wx_i\right) = \bar{y} - w\bar{x}$$

$$w = \frac{\sum_{i=1}^{m} y_i \left(x_i - \bar{x}\right)}{\sum_{i=1}^{m} x_i^2 - \frac{1}{m} \left(\sum_{i=1}^{m} x_i\right)^2}$$

- Vectorization:

$$w = \frac{\boldsymbol{y}_d^T \boldsymbol{x}_d}{\boldsymbol{x}_d^T \boldsymbol{x}_d}$$

https://accepteddoge.com/cnblogs/regression-and-gradient-descent

Multiple linear regression

- Data Matrix and Label Vector

$$X = (\boldsymbol{x_1}, \boldsymbol{x_2} \ldots \boldsymbol{x_N})^T = \begin{bmatrix} - & \boldsymbol{x_1}^T & - \\ - & \boldsymbol{x_2}^T & - \\ & \vdots & \\ - & \boldsymbol{x_N}^T & - \end{bmatrix} = \begin{pmatrix} x_{11} & x_{12} & \ldots & x_{1P} \\ x_{21} & x_{22} & \ldots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \ldots & x_{NP} \end{pmatrix}_{N\times P} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N\times 1}$$

- Parameter Estimation (Closed-form Solution):

$$\boldsymbol{w} = (X^T X)^{-1} X^T Y$$

$$w = \frac{\boldsymbol{y}_d^T \boldsymbol{x}_d}{\boldsymbol{x}_d^T \boldsymbol{x}_d}$$

https://accepteddoge.com/cnblogs/high-level-linear-regression

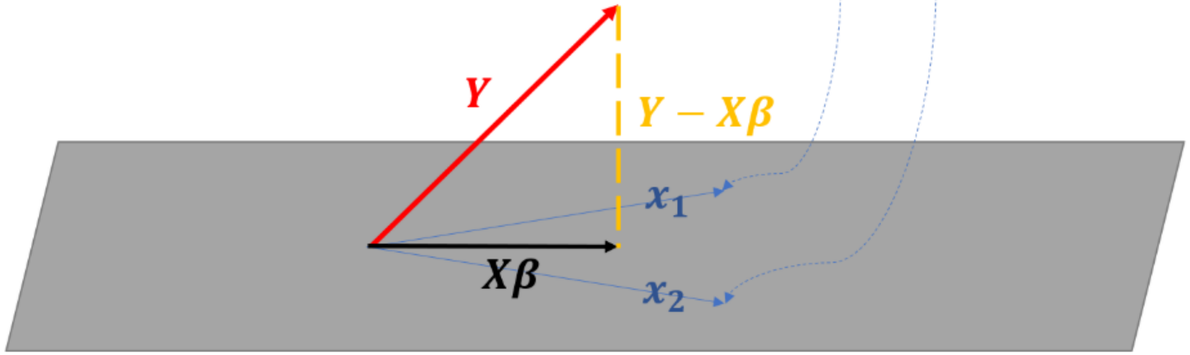
Another View

- We know that: $f(\boldsymbol{w}) = \boldsymbol{w}^T \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{\beta} = f(\boldsymbol{\beta})$

$$X = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1P} \\ x_{21} & x_{22} & \dots & x_{2P} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \dots & x_{NP} \end{pmatrix}_{N\times P} \qquad Y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}_{N\times 1}$$

$N = 3, P = 2$

$$X = \begin{bmatrix} x_{1,1} & x_{1,2} \\ x_{2,1} & x_{2,2} \\ x_{3,1} & x_{3,2} \end{bmatrix} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix}$$



$$X^T(Y - X\boldsymbol{\beta}) = \overrightarrow{0}$$
$$X^T Y = X^T X \boldsymbol{\beta}$$
$$\boldsymbol{\beta} = (X^T X)^{-1} X^T Y$$

$$\hat{\boldsymbol{w}} = \arg\min L(\boldsymbol{w}) = \left(X^T X\right)^{-1} X^T Y$$

https://accepteddoge.com/cnblogs/high-level-linear-regression

Why Regularization Useful

$$L(\boldsymbol{w}) = \sum_{i=1}^{N} \left\| \boldsymbol{w}^T \boldsymbol{x}_i - y_i \right\|^2 = \sum_{i=1}^{N} \left(\boldsymbol{w}^T \boldsymbol{x}_i - y_i \right)^2$$

$$\hat{\boldsymbol{w}} = \arg\min L(\boldsymbol{w}) = \left(X^T X \right)^{-1} X^T Y$$

$$\arg\min_{\boldsymbol{w}} \left[L(\boldsymbol{w}) + \lambda P(\boldsymbol{w}) \right]$$

$$\begin{aligned} J(\boldsymbol{w}) &= \sum_{i=1}^{N} \left\| \boldsymbol{w}^T \boldsymbol{x}_i - y_i \right\|^2 + \lambda \boldsymbol{w}^T \boldsymbol{w} \\ &= (\boldsymbol{w}^T X^T - Y^T)(X\boldsymbol{w} - Y) + \lambda \boldsymbol{w}^T \boldsymbol{w} \\ &= \boldsymbol{w}^T X^T X \boldsymbol{w} - 2\boldsymbol{w}^T X^T Y + Y^T Y + \lambda \boldsymbol{w}^T \boldsymbol{w} \\ &= \boldsymbol{w}^T (X^T X + 2\lambda I)\boldsymbol{w} - 2\boldsymbol{w}^T X^T Y + Y^T Y \end{aligned}$$

$$\frac{\partial J(\boldsymbol{w})}{\partial \boldsymbol{w}} = 2\left(X^T X + \lambda I \right) \boldsymbol{w} - 2X^T Y = 0$$

$$\hat{\boldsymbol{w}} = \arg\min \left[L(\boldsymbol{w}) + \lambda P(\boldsymbol{w}) \right] = \boxed{(X^T X + \lambda I)^{-1}} X^T Y$$

https://accepteddoge.com/cnblogs/high-level-linear-regression

Why Gradient Descent Useful

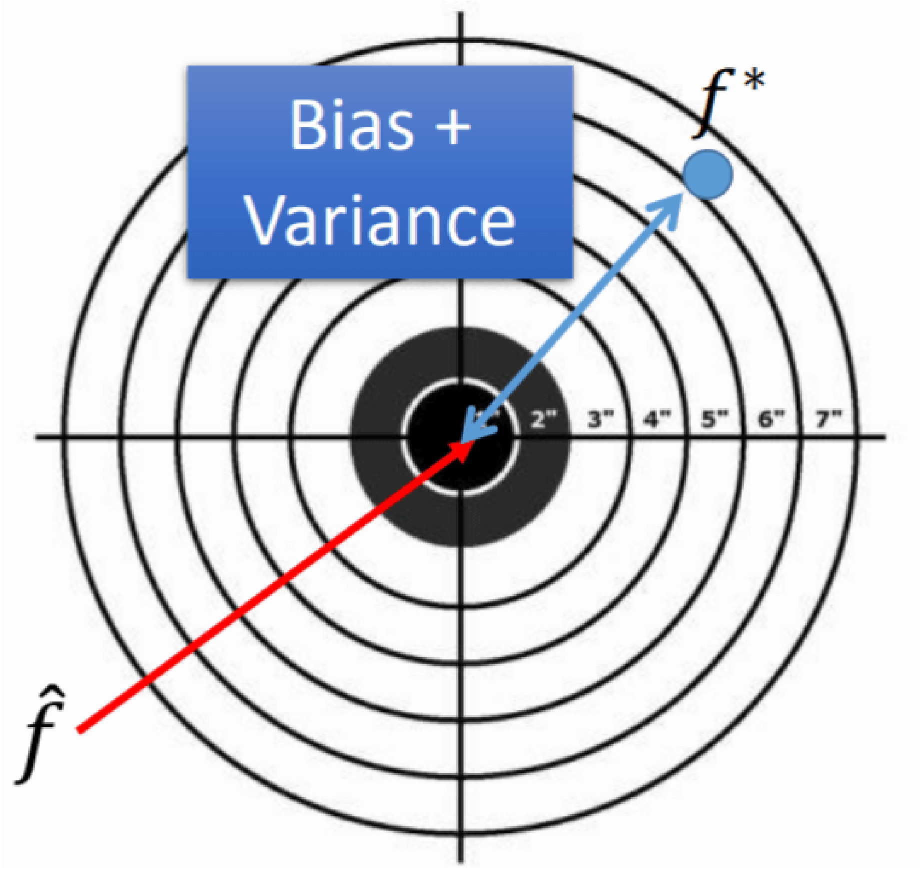
$$J(\boldsymbol{w}) = \sum_{i=1}^{N} \left\| \boldsymbol{w}^T \boldsymbol{x}_i - y_i \right\|^2 + \lambda \boldsymbol{w}^T \boldsymbol{w}$$

$$\nabla J(\boldsymbol{w}) = 2\left(X^T X + \lambda I\right)\boldsymbol{w} - 2X^T Y$$

$$H = \nabla^2 J(\boldsymbol{w}) = 2\left(X^T X + \lambda I\right) > 0$$

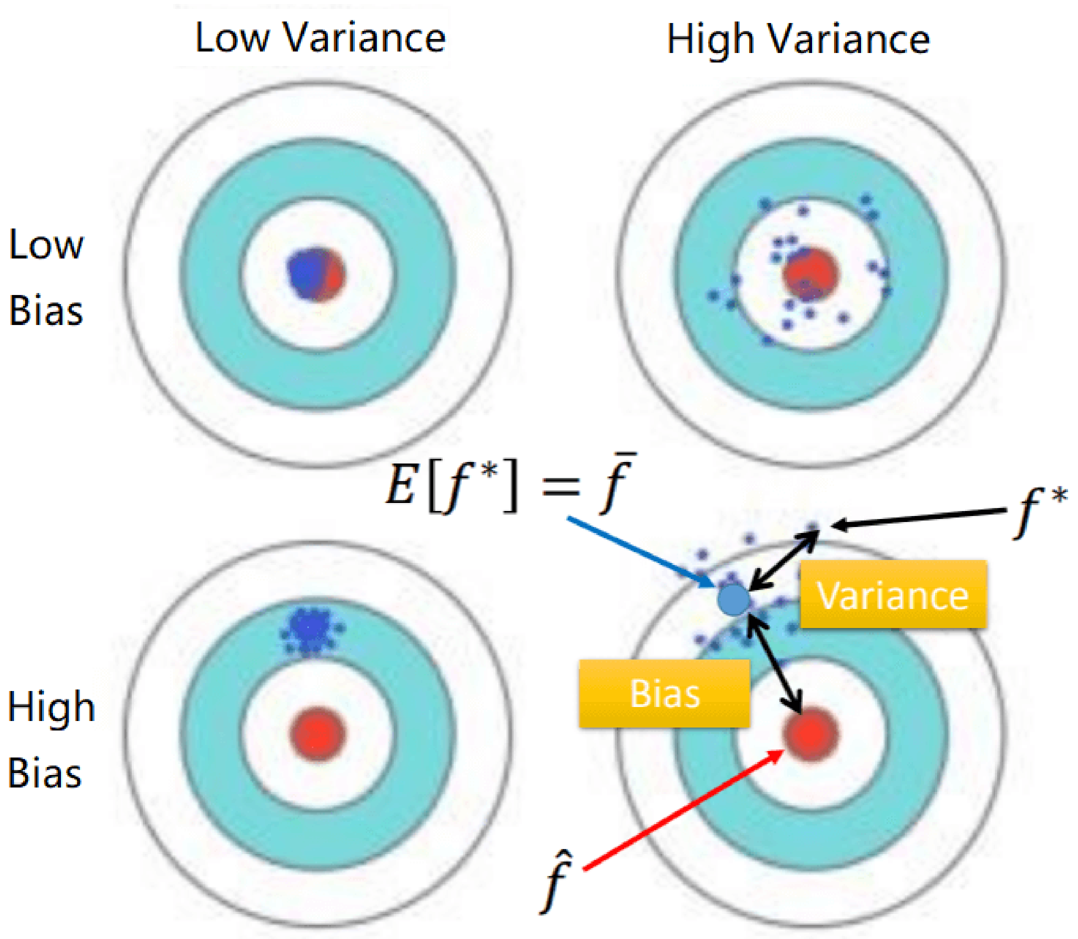
https://accepteddoge.com/cnblogs/high-level-linear-regression

Bias and Variance



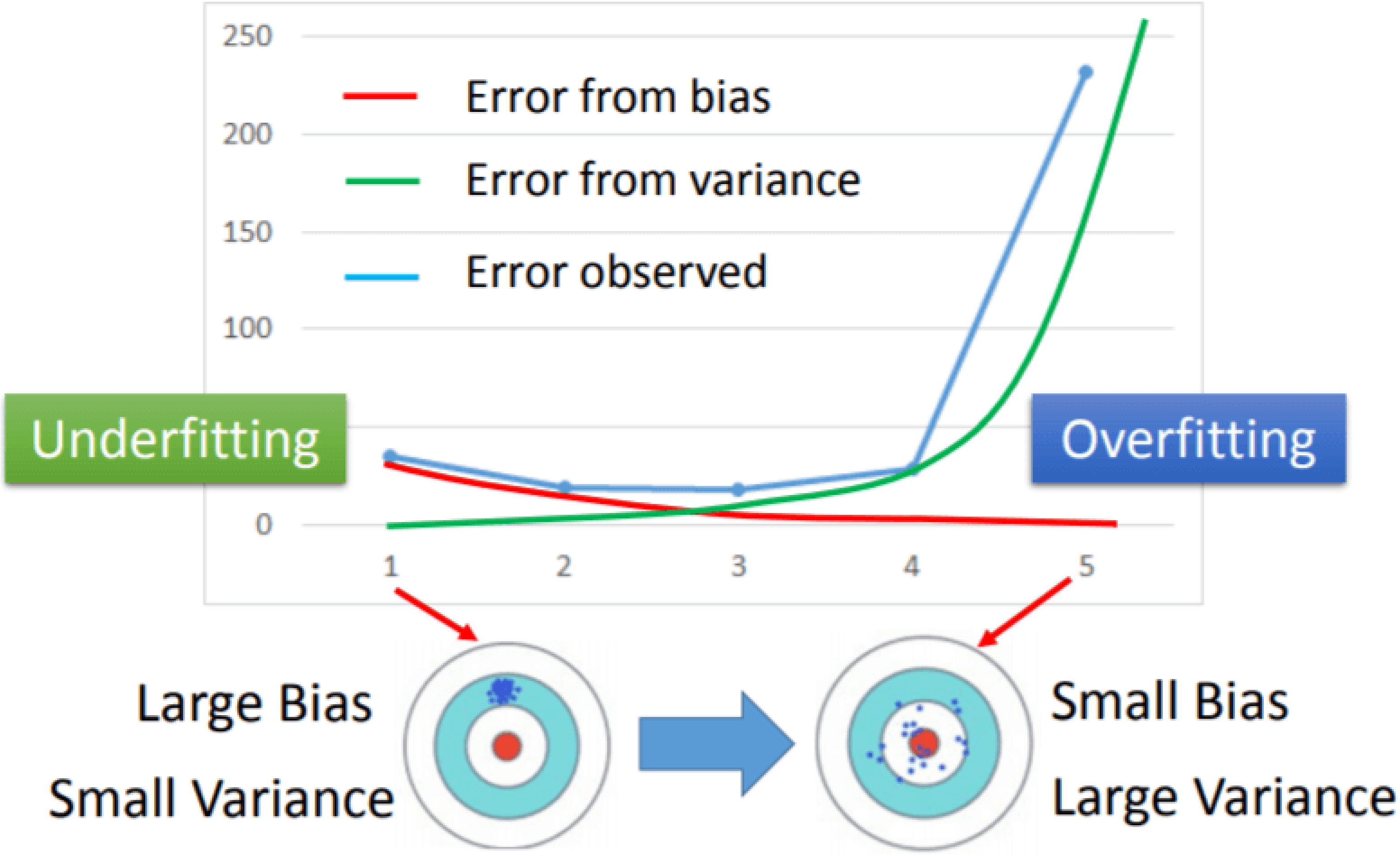
https://accepteddoge.com/cnblogs/bias-and-variance

Bias and Variance - Example



https://accepteddoge.com/cnblogs/bias-and-variance

Error Analysis

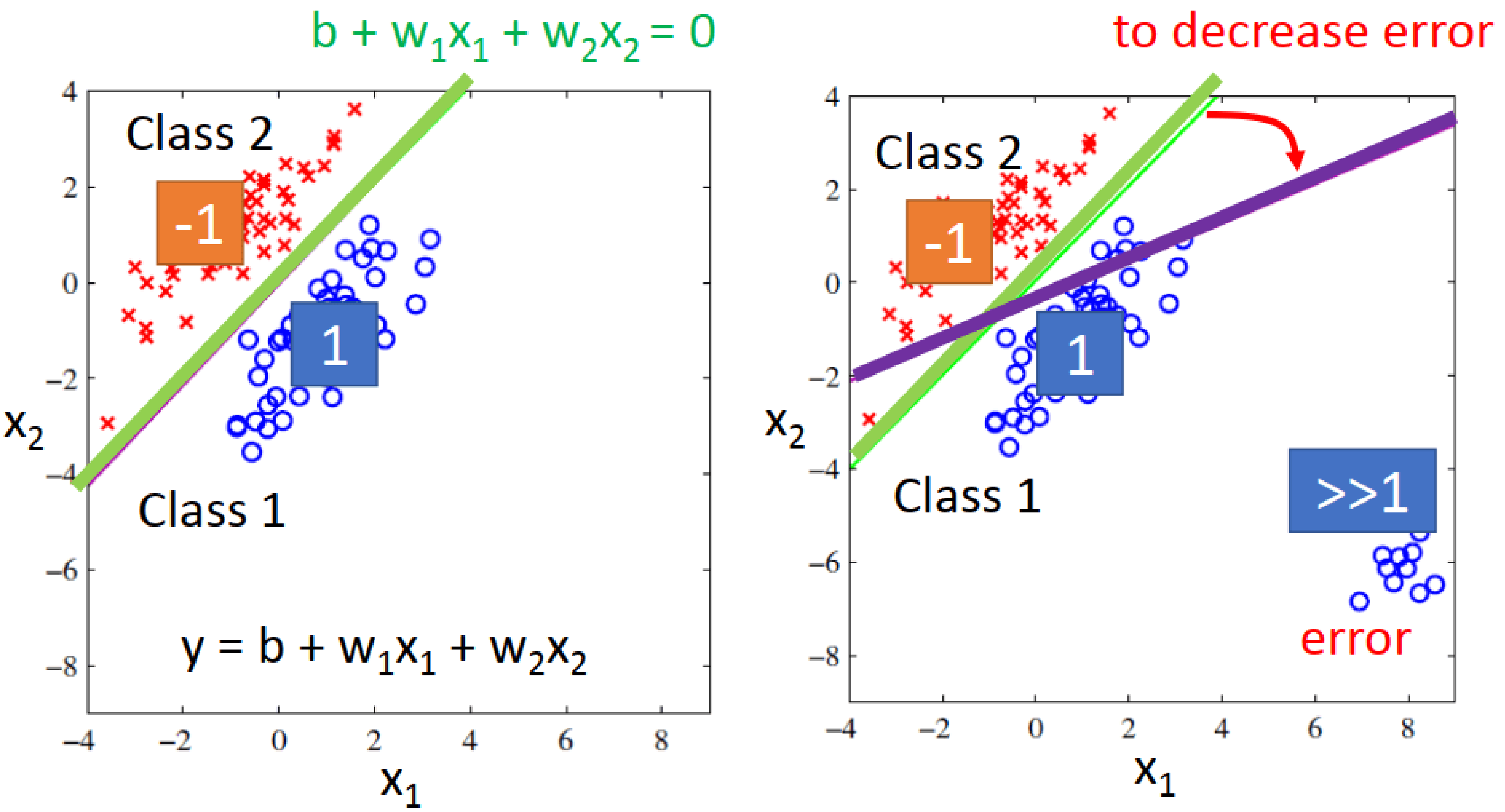


https://accepteddoge.com/cnblogs/bias-and-variance

Break

5 mins

From Regression to Classification

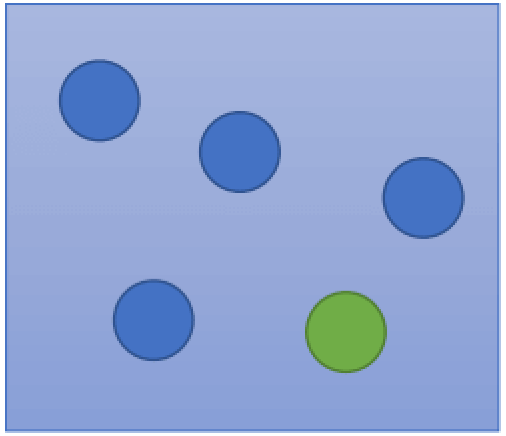


https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Classification Probability

Box 1

$P(B_1) = 2/3$

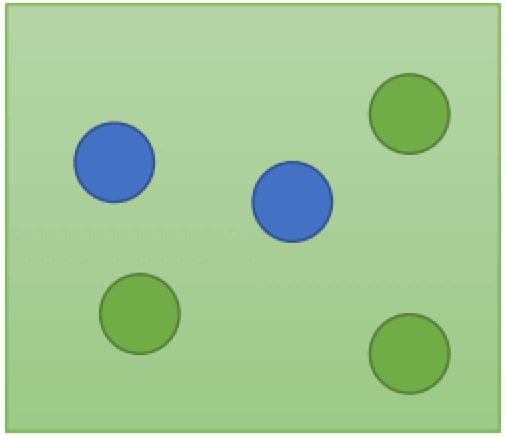


$P(Blue|B_1) = 4/5$

$P(Green|B_1) = 1/5$

Box 2

$P(B_2) = 1/3$



$P(Blue|B_1) = 2/5$

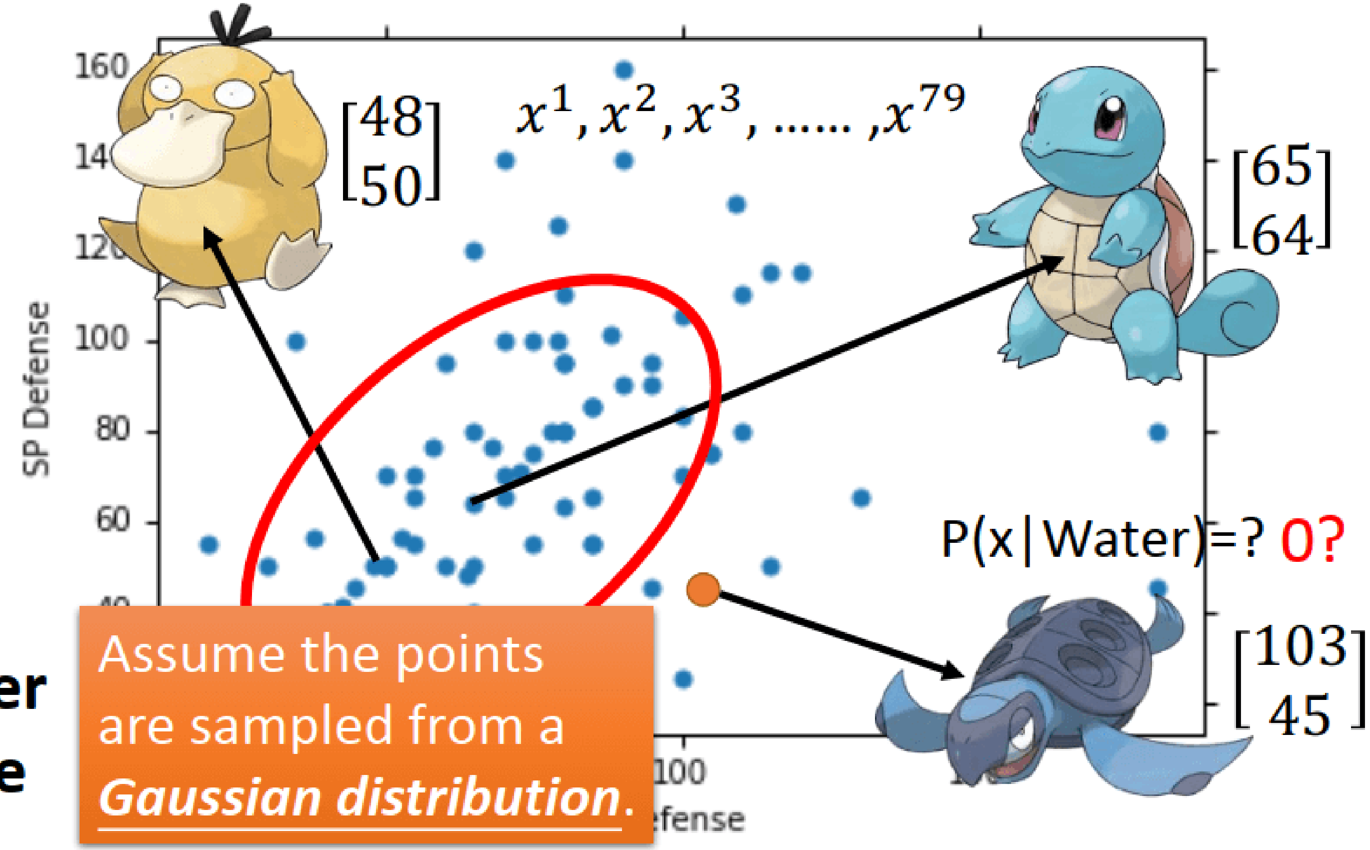
$P(Green|B_1) = 3/5$

$$\mathrm{P}\left(\mathrm{B}_1|\mathrm{Blue}\right) = \frac{P(\text{ Blue }|B_1)\, P(B_1)}{P(\text{ Blue }|B_1)\, P(B_1) + P(\text{ Blue }|B_2)\, P(B_2)}$$

https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Data with Noise

Water Type



Assume the points are sampled from a ***Gaussian distribution***.

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{ -\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu) \right\}$$

https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Maximum Likelihood

$$L(\mu, \Sigma) = f_{\mu,\Sigma}\left(x^1\right) f_{\mu,\Sigma}\left(x^2\right) f_{\mu,\Sigma}\left(x^3\right) \ldots\ldots f_{\mu,\Sigma}\left(x^{79}\right)$$

$$f_{\mu,\Sigma}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right\}$$

$$\mu^*, \Sigma^* = \arg\max_{\mu,\Sigma} L(\mu, \Sigma)$$

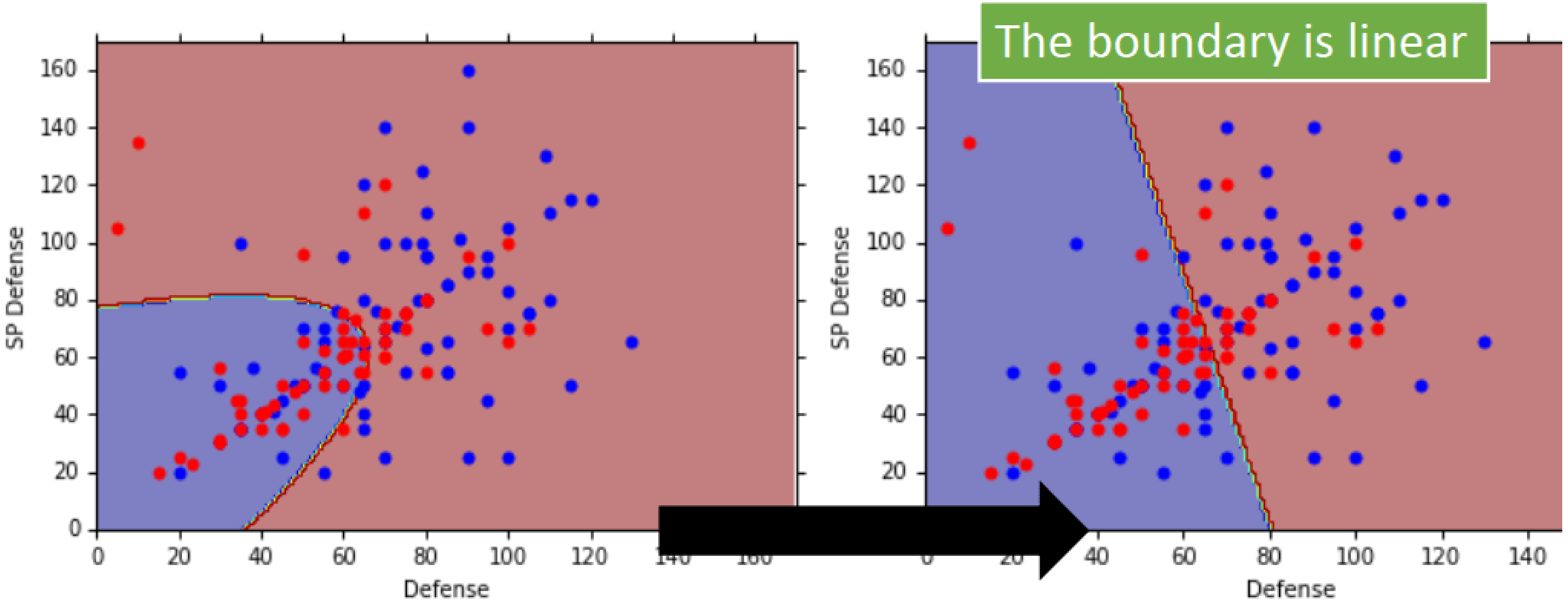
$$\mu^* = \frac{1}{79}\sum_{n=1}^{79} x^n$$

$$\Sigma^* = \frac{1}{79}\sum_{n=1}^{79} \left(x^n - \mu^*\right)\left(x^n - \mu^*\right)^T$$

$$P(C|x) = f_{\mu^*,\Sigma^*}(x) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\Sigma^*|^{1/2}} \exp\left\{-\frac{1}{2}(x-\mu^*)^T (\Sigma^*)^{-1}(x-\mu^*)\right\}$$

https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Covariance Matrix



The same covariance matrix

https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Probability Model

$$P(C_1|x) = \frac{P(x|C_1)P(C_1)}{P(x|C_1)P(C_1) + P(x|C_2)P(C_2)}$$

$$= \frac{1}{1 + \frac{P(x|C_2)P(C_2)}{P(x|C_1)P(C_1)}}$$

$$z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)}$$

$$P(C_1|x) = \frac{1}{1+\exp(-z)} = \boxed{\sigma(z)}$$

$$z = \ln \frac{P(x|C_1)P(C_1)}{P(x|C_2)P(C_2)} = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{P(C_1)}{P(C_2)}$$

...

$$z = \ln \frac{|\Sigma^2|^{1/2}}{|\Sigma^1|^{1/2}} - \frac{1}{2}x^T(\Sigma^1)^{-1}x + (\mu^1)^T(\Sigma^1)^{-1}x - \frac{1}{2}(\mu^1)^T(\Sigma^1)^{-1}\mu^1$$

$$+ \frac{1}{2}x^T(\Sigma^2)^{-1}x - (\mu^2)^T(\Sigma^2)^{-1}x + \frac{1}{2}(\mu^2)^T(\Sigma^2)^{-1}\mu^2 + \ln\frac{N_1}{N_2}$$

$$\Sigma_1 = \Sigma_2 = \Sigma$$

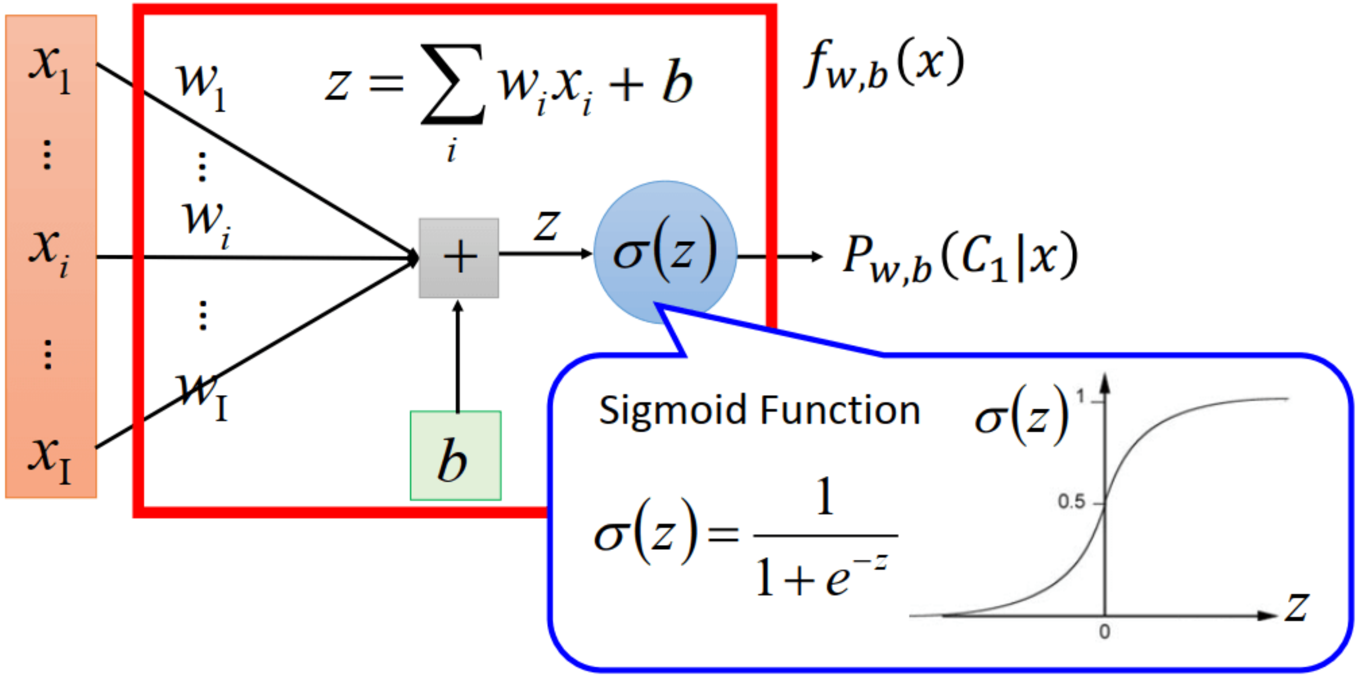
$$z = \underline{(\mu^1 - \mu^2)^T\Sigma^{-1}}x \underline{- \frac{1}{2}(\mu^1)^T\Sigma^{-1}\mu^1 + \frac{1}{2}(\mu^2)^T\Sigma^{-1}\mu^2 + \ln\frac{N_1}{N_2}}$$

$$z = \underline{\boldsymbol{W}^T}x + \underline{b}$$

$$P(C_1|x) = \sigma(w \cdot x + b)$$

https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Logistic Regression



$$P_{w,b}(C_1|x) = \sigma(z)$$

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

$$z = w \cdot x + b = \sum_i w_i x_i + b$$

$$z = \ln \frac{P(x|C_1)\,P(C_1)}{P(x|C_2)\,P(C_2)} = \ln \frac{P(x|C_1)}{P(x|C_2)} + \ln \frac{P(C_1)}{P(C_2)}$$

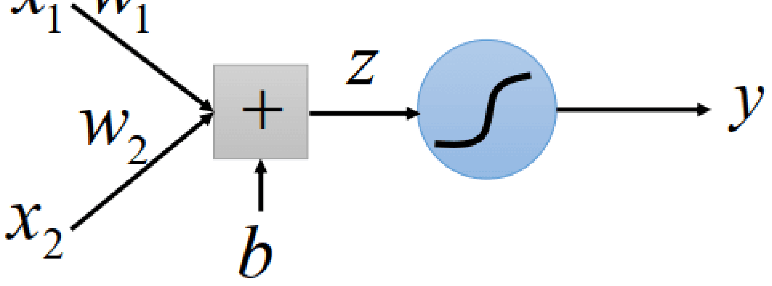
https://accepteddoge.com/cnblogs/classification-and-logistic-regression

More Logistic Regression

- Loss Function
  - MLE
- Optimize
  - Gradient Descent

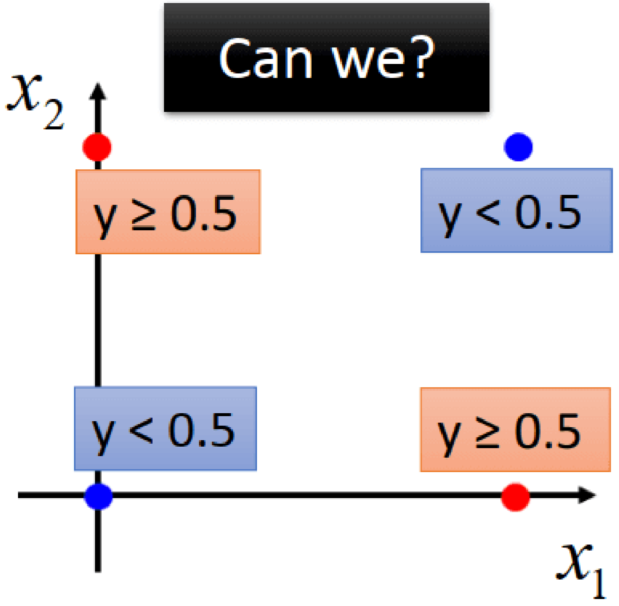
Logistic Regression

$$z = w_1 x_1 + w_2 x_2 + b$$



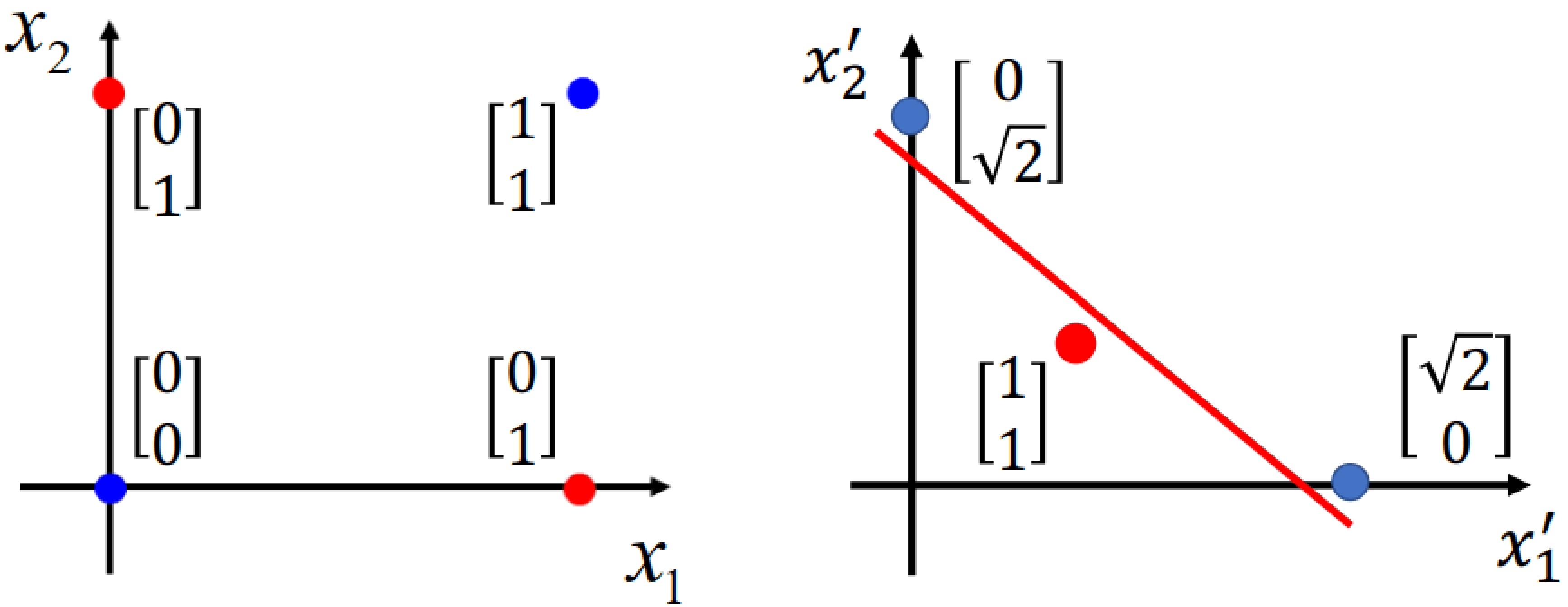
$$\begin{cases} Class1 & y \geq 0.5 \\ Class2 & y < 0.5 \end{cases}$$

Input Feature		Label
x_1	x_2	
0	0	Class 2
0	1	Class 1
1	0	Class 1
1	1	Class 2



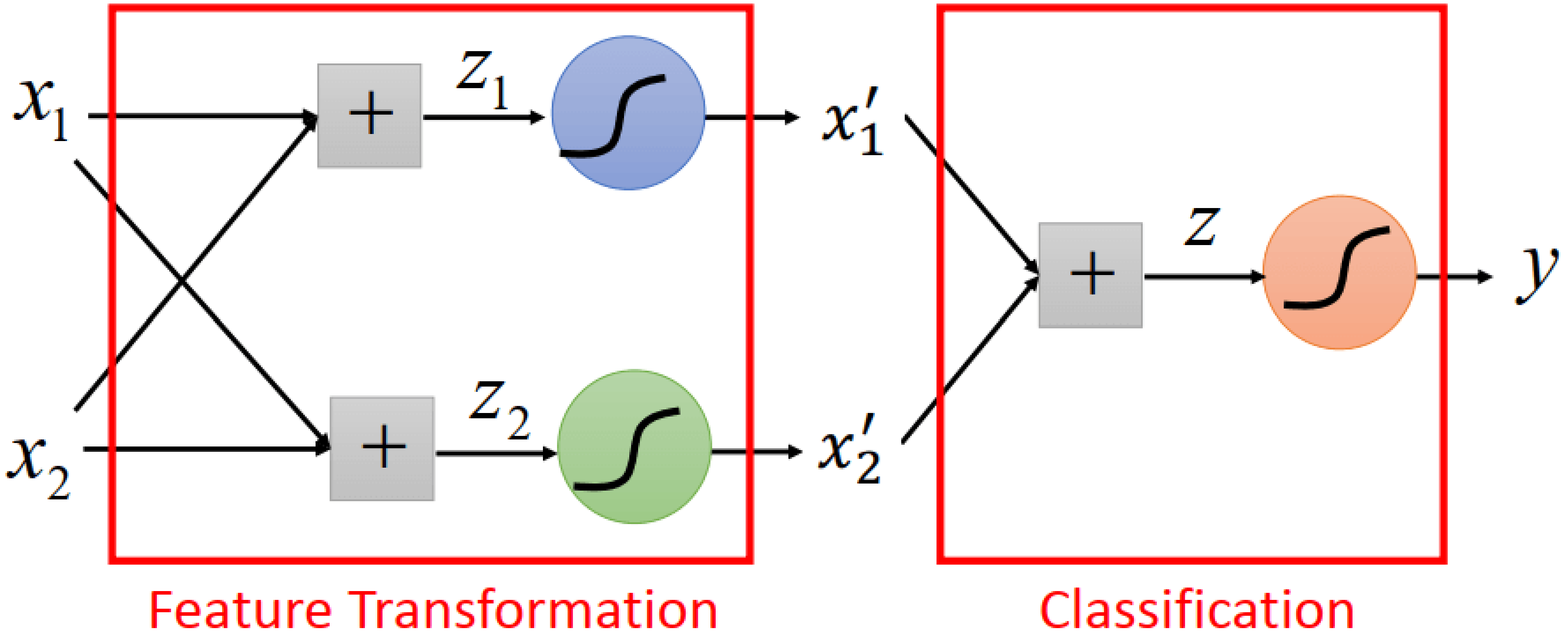
https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Logistic Regression



https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Logistic Regression vs. NN



https://accepteddoge.com/cnblogs/classification-and-logistic-regression

Machine Learning Algorithm

Classification

Identifying to which category an object belongs to.

Applications: Spam detection, Image recognition.

Algorithms: SVM, nearest neighbors, random forest, … — Examples

Regression

Predicting a continuous-valued attribute associated with an object.

Applications: Drug response, Stock prices.

Algorithms: SVR, ridge regression, Lasso, … — Examples

Clustering

Automatic grouping of similar objects into sets.

Applications: Customer segmentation, Grouping experiment outcomes

Algorithms: k-Means, spectral clustering, mean-shift, … — Examples

Dimensionality reduction

Reducing the number of random variables to consider.

Applications: Visualization, Increased efficiency

Algorithms: PCA, feature selection, non-negative matrix factorization. — Examples

Model selection

Comparing, validating and choosing parameters and models.

Goal: Improved accuracy via parameter tuning

Modules: grid search, cross validation, metrics. — Examples

Preprocessing

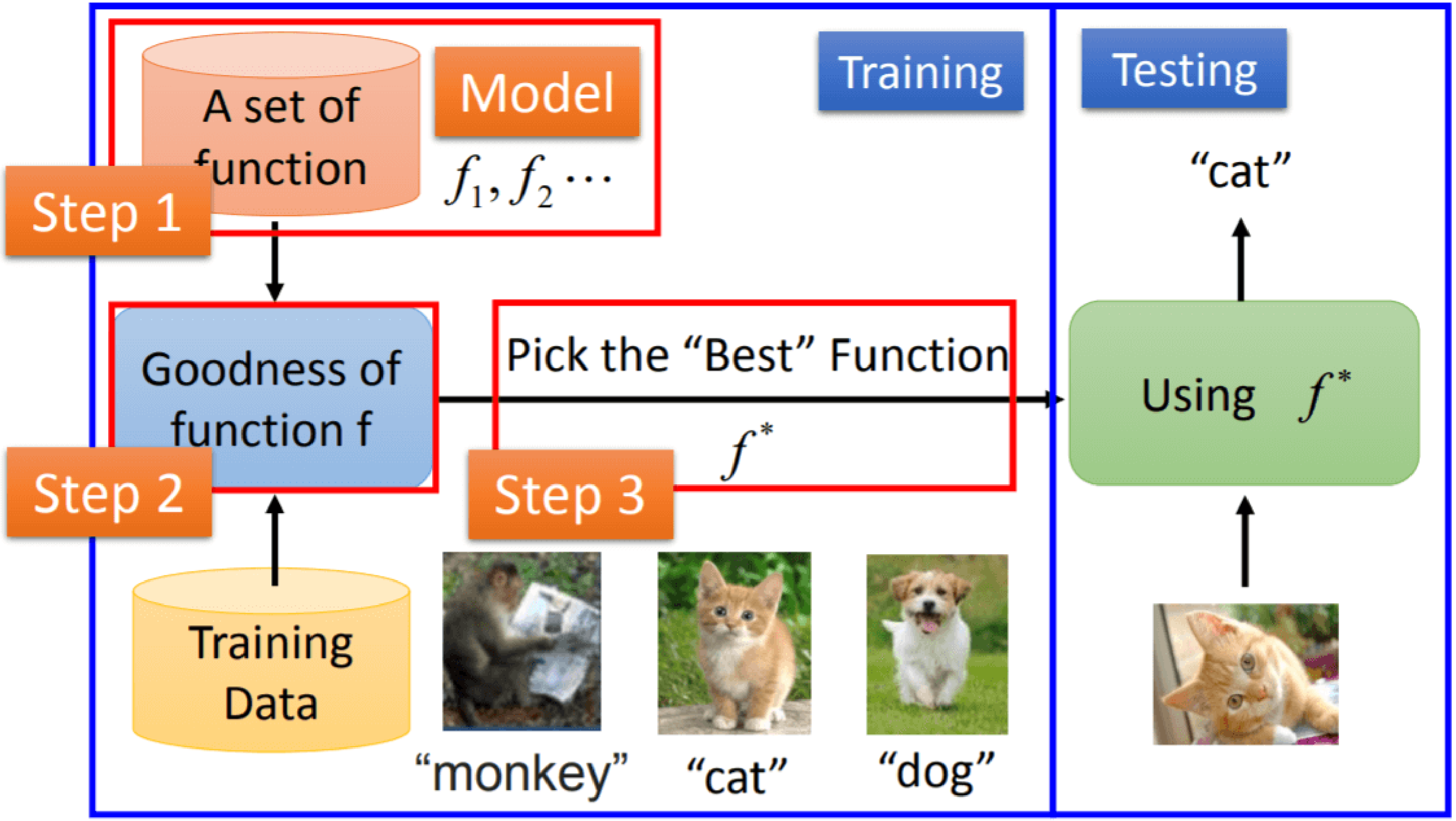
Feature extraction and normalization.

Application: Transforming input data such as text for use with machine learning algorithms.

Modules: preprocessing, feature extraction. — Examples

https://scikit-learn.org/stable/

Machine Learning Basic Concepts



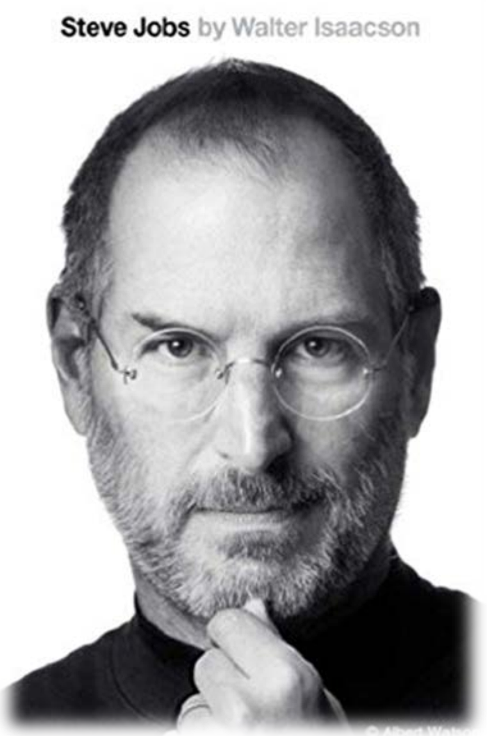
http://speech.ee.ntu.edu.tw/~tlkagk/courses_ML17_2.html

Contents

- Quick Review
  - Decision Tree
  - Linear Models
- Quiz Discussion
- Supplement
  - Machine Learning Again
  - More Linear Models
- Warm Up for Next Week

Steve Jobs' 2005 Stanford Commencement Address

- Just three stories:
  - Connecting the dots
  - Love and loss
  - Death
- Stay Hungry
- Stay Foolish



Steve Jobs
Co-founder, Chairman, and CEO of Apple Inc.

- 'You've got to find what you love,' Jobs says – Stanford (Text of the speech)
- Steve Jobs' 2005 Stanford Commencement Address - YouTube

Have a nice weekend~

"You can't connect the dots looking forward; you can only connect them looking backward. So you have to trust that the dots will somehow connect in your future." -- Steve Jobs