The Introduction To Artificial Intelligence

Yuni Zeng yunizeng@zstu.edu.cn

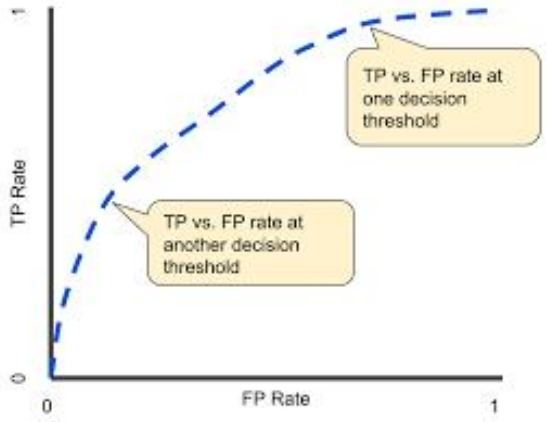
2024-2025-1

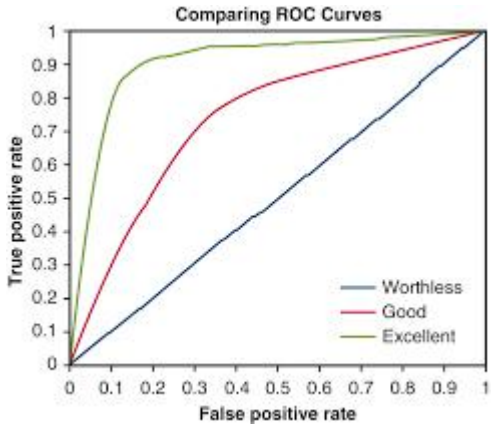
A brief review

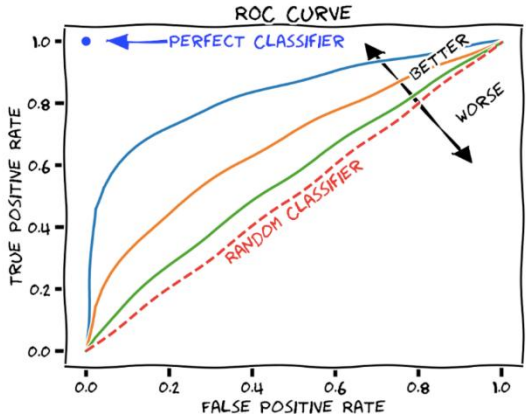
- How to make a model convincible?
  - Error, Training error, Generalization error
  - Overfitting and Underfitting
  - Evaluation Methods: Hold-out method, Cross Validation, Bootstrapping
- How to evaluate a model?
  - Measure metrics: ACC, Recall, F1,AUC…

1.3 Performance Measure

- ROC Curve (Receiver Operating Characteristic)







- An ROC curve (receiver operating characteristic curve) is a graph showing the performance of a classification model at all classification thresholds.
- TPR – FPR:
  - TPR: True positive rate
  - FPR: False positive rate

Test

- ROC Curve (Receiver Operating Characteristic)

样本编号 (No.)	真实标签 (True label)	模型输出概率 (output probability)	样本编号(No.)	真实标签 (True label)	模型输出概率 (output probability)
1	p	0.9	11	p	0.4
2	p	0.8	12	n	0.39
3	n	0.7	13	p	0.38
4	p	0.6	14	n	0.37
5	p	0.55	15	n	0.36
6	p	0.54	16	n	0.35
7	n	0.53	17	p	0.34
8	n	0.52	18	n	0.33
9	p	0.51	19	p	0.30
10	n	0.505	20	n	0.10

- p : positive sample, n: negative sample

Test

- ROC Curve (Receiver Operating Characteristic)

- P: number of positive samples; TP: number of true positive samples
- N: number of negative samples; FP: number of false positive samples

Thresholds	0.9	0.8	0.7	0.6	0.55	0.54	0.53	0.52	0.51	0.505
TPR										
FPR										

Thresholds	0.4	0.39	0.38	0.37	0.36	0.35	0.34	0.33	0.30	0.10
TPR										
FPR										

Test

- ROC Curve (Receiver Operating Characteristic)

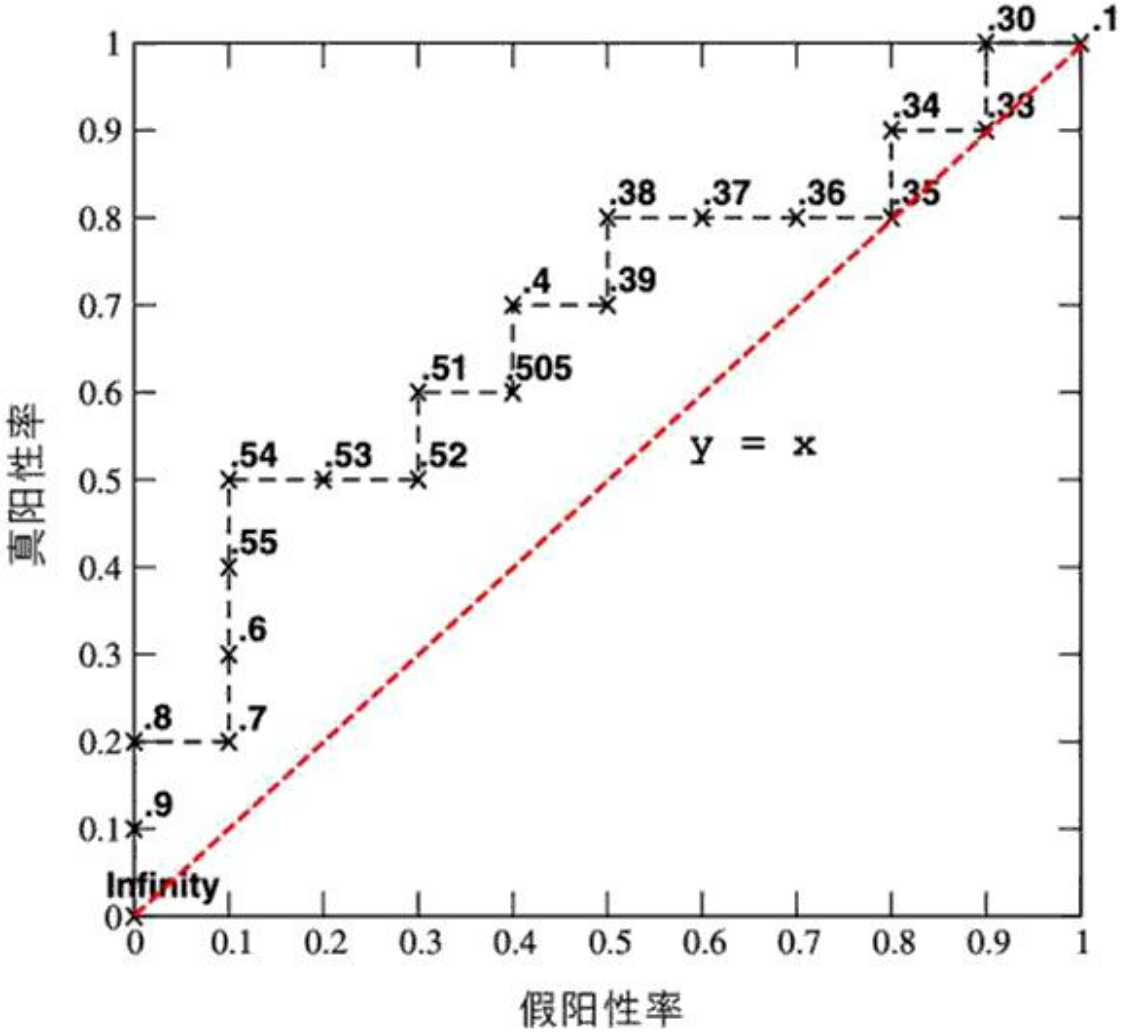
- P: number of positive samples; TP: number of true positive samples
- N: number of negative samples; FP: number of false positive samples

Thresholds	0.9	0.8	0.7	0.6	0.55	0.54	0.53	0.52	0.51	0.505
TPR = TP/P	0.1	0.2	0.2	0.3	0.4	0.5	0.5	0.5	0.6	0.6
FPR = FP/N	0	0	0.1	0.1	0.1	0.1	0.2	0.3	0.3	0.4

Thresholds	0.4	0.39	0.38	0.37	0.36	0.35	0.34	0.33	0.30	0.10
TPR = TP/P	0.7	0.7	0.8	0.8	0.8	0.8	0.9	0.9	1.0	1.0
FPR = FP/N	0.4	0.5	0.5	0.6	0.7	0.8	0.8	0.9	0.9	1.0

Test

- ROC Curve (Receiver Operating Characteristic)



The Introduction to Artificial Intelligence

- Part I Brief Introduction to AI & Different AI tribes
- Part II Knowledge Representation & Reasoning
- Part III AI GAMES and Searching
- Part IV Model Evaluation and Selection
- Part V Machine Learning

Machine Learning

Supervised learning

Unsupervised learning

Reinforcement learning

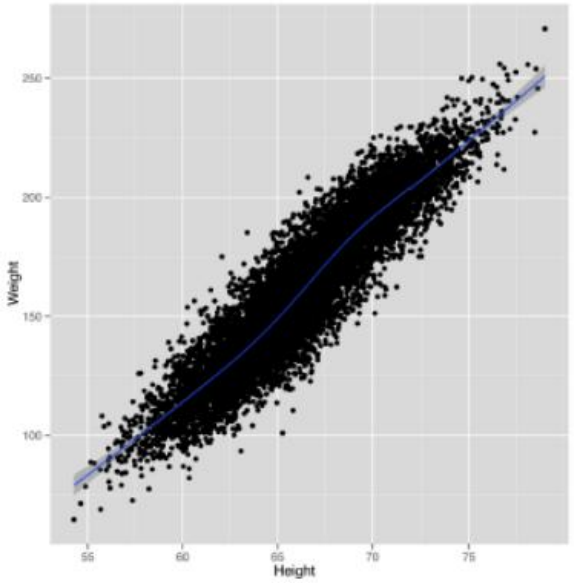
Supervised learning

- *Linear Regression*
- Logistic Regression
- Classification
  - Distance-based algorithms
  - Linear classifiers
  - Other classifiers
- ……

Linear Regression

- What is regression?

Regression is to relate input variables to the output variable, to either predict outputs for new inputs and/or to interpret the effect of the input on the output.



Height is correlated with weight.

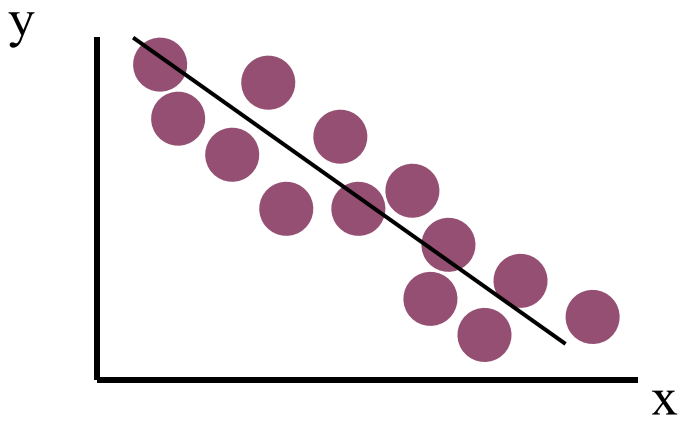
Linear Regression

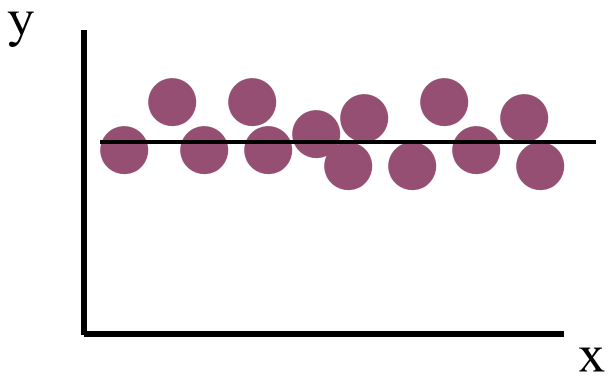
- Linear Regression Model
  - Only **one** independent variable, x
  - Relationship between x and y is described by a linear function
  - Changes in y are assumed to be related to changes in x

Linear Regression

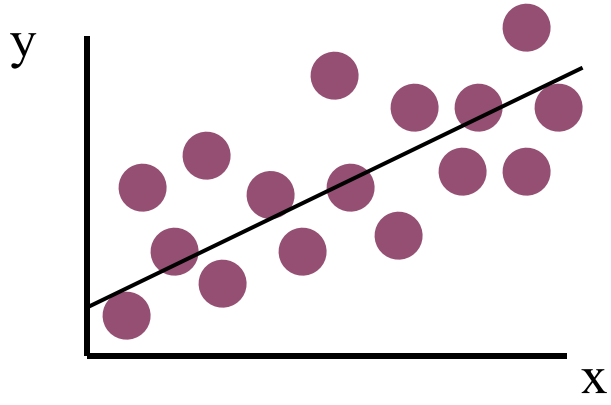
- Linear Regression Model

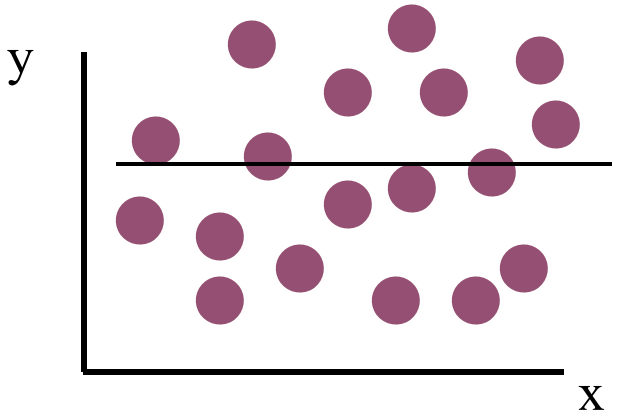
Linear relationships





Question: How to describe the linear relationships?

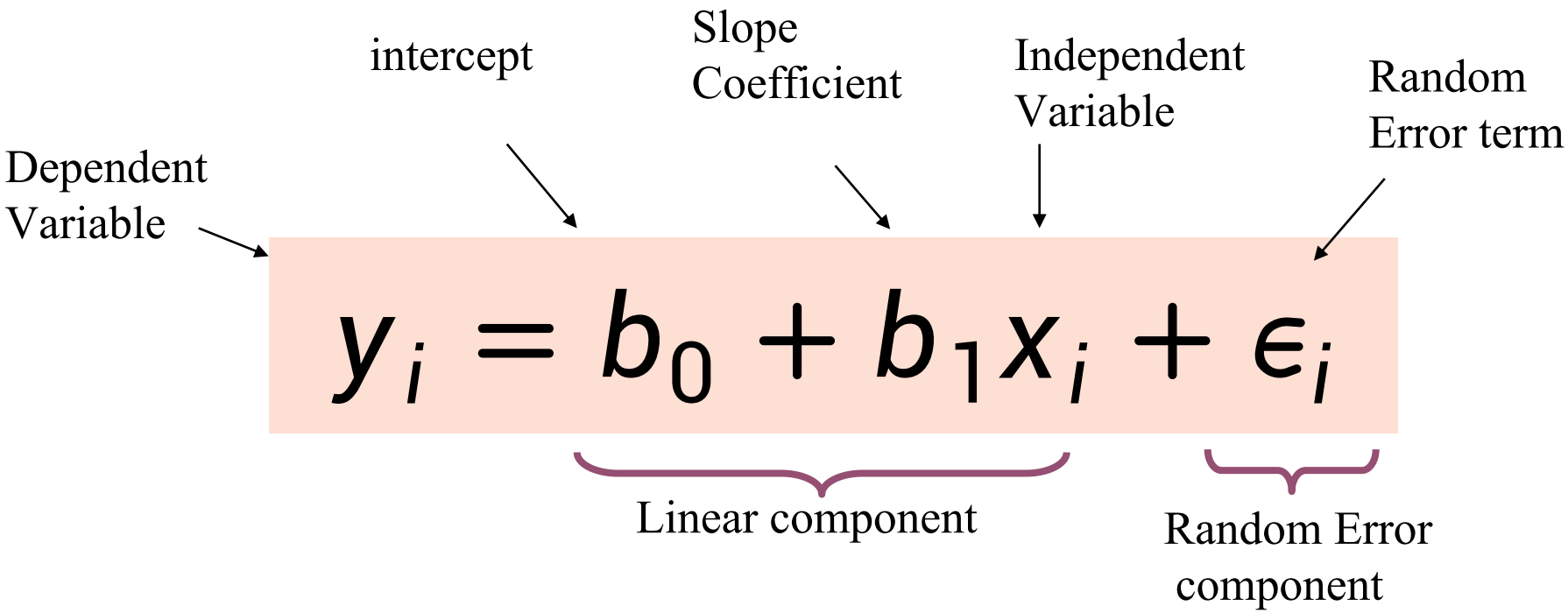




No relationship

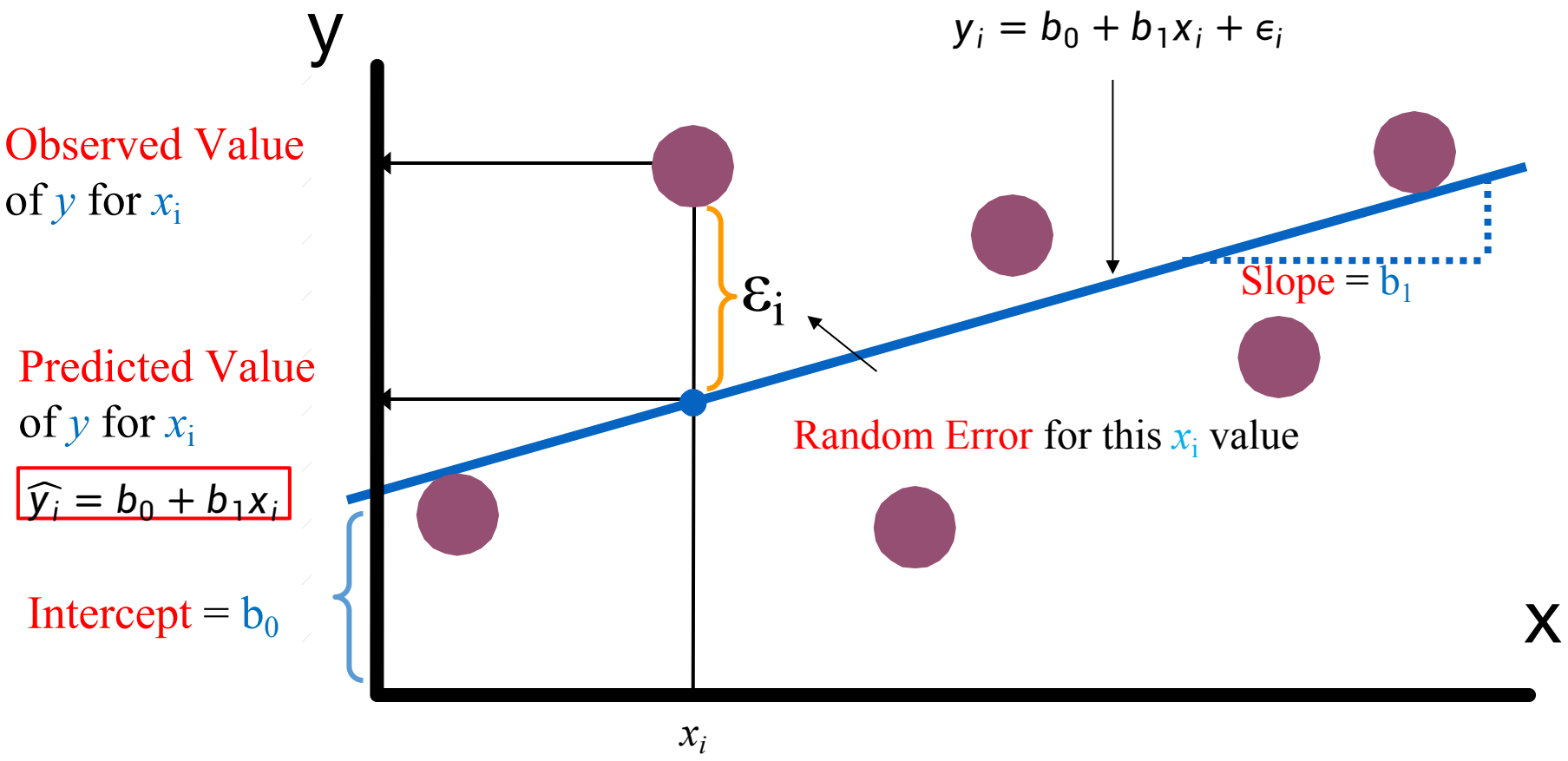
Linear Regression

- Linear Regression Model



Linear Regression

- Linear Regression Model



Question: How to obtain the best line?

Linear Regression

- The Least Squares Method

b_0 and b_1 are obtained by finding the values of that minimize the sum of the squared differences between y_i and $\hat{y}_i$ for all i :

$$\min \sum (y_i - \hat{y}_i)^2$$

$$\hat{y}_i = b_0 + b_1 x_i$$

$$\min \sum (y_i - (b_0 + b_1 x_i))^2 \rightarrow \text{Objective function}$$

Question: How to calculate b_0 and b_1?

$$\text{derivative}[\sum (y_i - (b_0 + b_1 x_i))^2] = 0 \quad \rightarrow \quad \text{solve for } b_0, b_1$$

Linear Regression

The Least Squares Method

- Considering the objective function:

$$J = \sum (y_i - (b_0 + b_1 x_i))^2$$

- Rewrite it in matrix form as:

$$J = \|Y - \theta^T X\|_2^2$$

$$\text{where } Y = [y_1, \cdots, y_n],\ X = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix},\ \text{and } \theta = \begin{bmatrix} b_0 \\ b_1 \end{bmatrix}$$

$$\frac{\partial J}{\partial \theta} = -2(Y - \theta^T X)X^T = 0$$

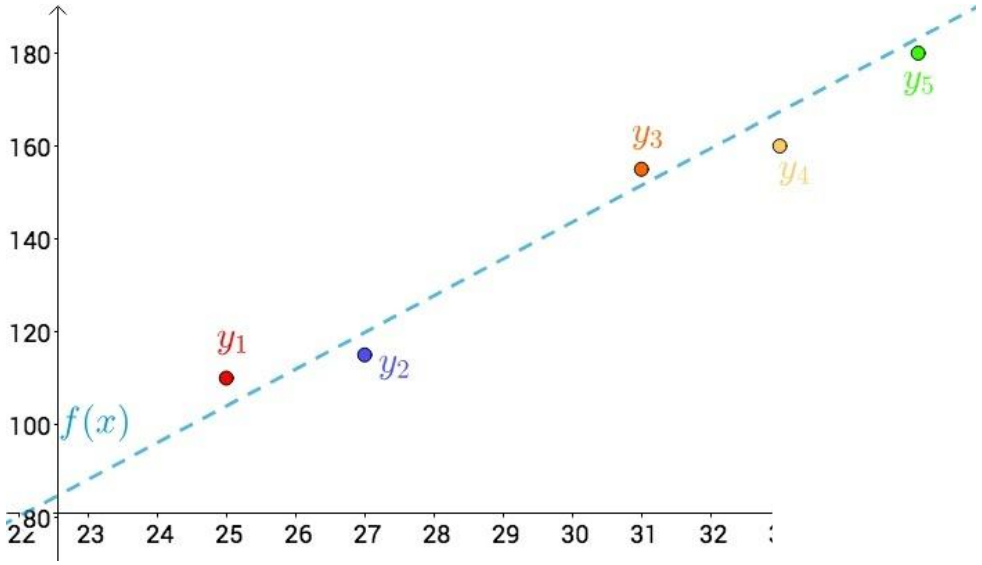
$$\theta^* = (XX^T)^{-1} XY^T$$

Linear Regression

An Example

- between temperature and ice cream sales:

Temperature	Sales
25°	110
27°	115
31°	155
33°	160
35°	180



Seems like a linear relationship

Linear Regression

An Example

- between temperature and ice cream sales:
- Set: $y = ax + b$

Temperature	Sales
25°	110
27°	115
31°	155
33°	160
35°	180

i	x	y
1	25	110
2	27	115
3	31	155
4	33	160
5	35	180

Linear Regression

An Example

- between temperature and ice cream sales:
- Set: $y = ax + b$
- $J = \sum (f(x_i) - y_i)^2 = \sum (ax_i + b - y_i)^2$
- $\begin{cases} \frac{\partial}{\partial a} J = 2 \sum (ax_i + b - y) x_i = 0 \\ \frac{\partial}{\partial b} J = 2 \sum (ax_i + b - y) = 0 \end{cases}$
- $\begin{cases} a \approx 7.2 \\ b \approx -73 \end{cases}$

i	x	y
1	25	110
2	27	115
3	31	155
4	33	160
5	35	180

Linear Regression

- Another Example
  - A real estate agent wishes to examine the relationship between the selling price of a houses and its size (measured in square feet)
  - A random sample of 10 houses is selected
    - Dependent variable (y) = house price in \$1000s
    - Independent variable (x) = square feet

Linear Regression

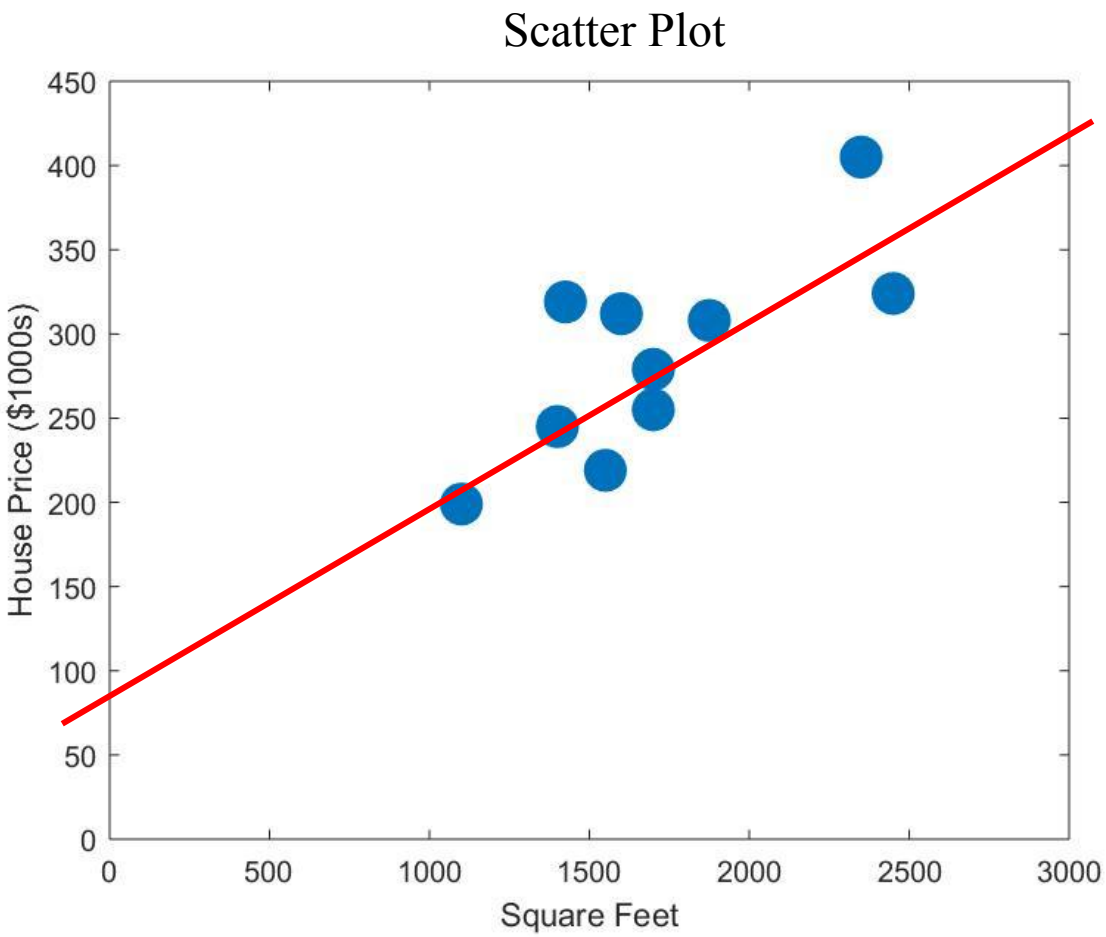
- An Example

House Price (y) in \$1000s	Square Feet (x)
245	1400
312	1600
279	1700
308	1875
199	1100
219	1550
405	2350
324	2450
319	1425
255	1700

Linear Regression

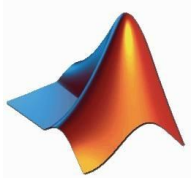
- An Example

$$\theta^* = (XX^T)^{-1}XY^T$$




```
>> theta = inv(X*X')*X*Y'

theta =

   98.2483
    0.1098
```

```
>> [epsilon, b1, b0] = regression(X, Y)

epsilon =

    0.7621


b1 =

    0.1098


b0 =

   98.2483
```



Linear Regression

- Conclusion: Linear Regression
- Uses least squares estimation to estimate parameters
  - Finds the line that minimizes total squared error around the line:
  - Sum of Squared Error (SSE)= $\Sigma(y_i-(b_0 + b_1x))^2$
  - Minimize the squared error function:

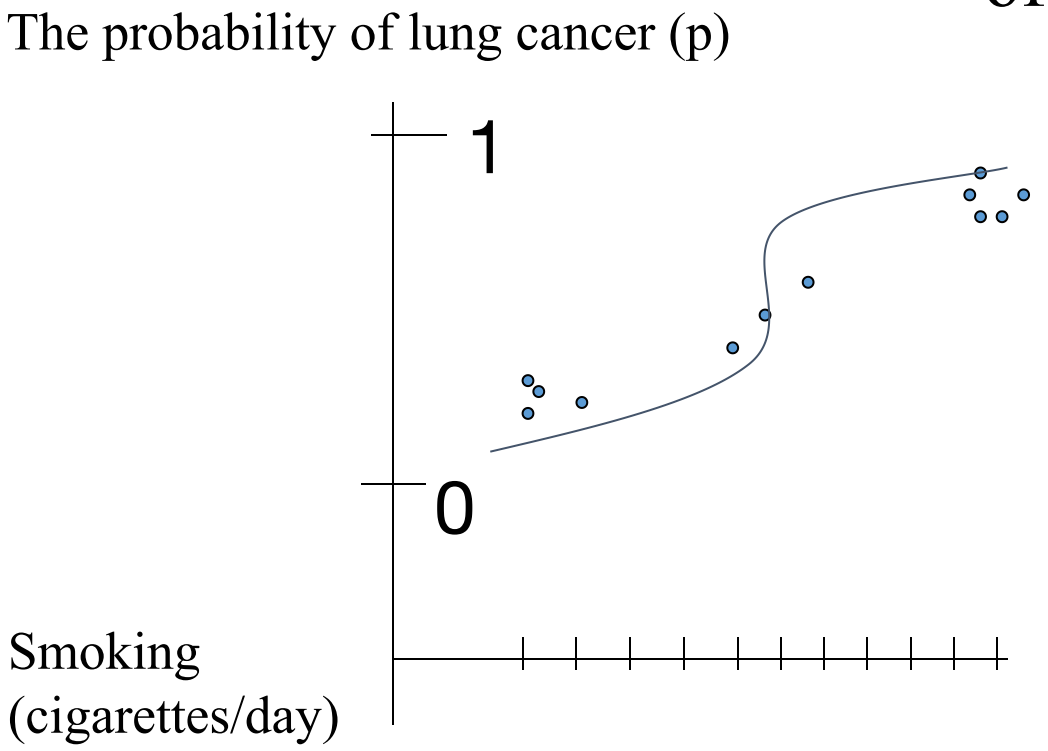
    derivative$[\Sigma(y_i-(b_0 + b_1x))^2]=0$→ solve for b_0, b_1

Linear Regression

- Thinking…

Could model probability of lung cancer…

$$P \leftarrow b_0 + b_1 x_i$$



But why might this not be best modeled as linear?

Logistic Regression

Logistic Regression Model

- In medical research, it is often necessary to analyze which factors are related to the outcome of a certain outcome.
- How do we find out which factors have a significant impact on the outcome?
- Logistic regression analysis can solve these problems better.

Logistic Regression

- Linear regression is written as:

$$y = b_0 + b_1 X \qquad -\infty \leq y \leq +\infty$$

- If we define y as disease or normal, it can not be modeled by the above equation.

- How about apply the probability to represent it?

$$p \leftarrow b_0, b_1, X$$

Logistic Regression

- Logistic Regression Model

Think about the probability…

probability of disease : p $0 \le p \le 1$

probability of no-disease : $1-p$ $0 \le p \le 1$

odds: $\frac{p}{1-p}$ $0 \le \frac{p}{1-p} < +\infty$

$\ln(\frac{p}{1-p})$ $-\infty < \ln(\frac{p}{1-p}) < +\infty$

Logistic Regression

- Logistic Regression Model

Define logistic model as

$$\ln \frac{p}{1-p} = b_0 + b_1 X$$

We obtained that,

$$p = \frac{1}{1 + e^{-(b_0 + b_1 X)}}$$

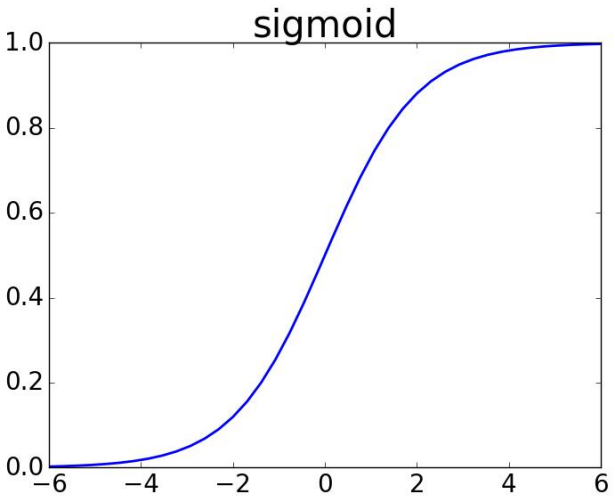
$$h_\theta(X) = \frac{1}{1 + e^{-\theta^T X}}$$

Therefore,

$$P(class = 1 | x; \theta) = h_\theta(X)$$

$$P(class = 0 | x; \theta) = 1 - h_\theta(X)$$

$$g(z) = \frac{1}{1 + e^{-z}}$$



The output of sigmoid function could be used to indicate the probability.

Logistic Regression

- Logistic Regression Model

$$P(class = 1|x;\theta) = h_\theta(X)$$

$$P(class = 0|x;\theta) = 1 - h_\theta(X)$$

$$P(class = y|x;\theta) = h_\theta(X)^y \left(1 - h_\theta(x)\right)^{1-y}$$

Considering all the given data (training set):

$$X = [x_1, \cdots, x_n], \qquad Y = [y_1, \cdots, y_n],$$

$$L(\theta) = \prod_i^n h_\theta(x_i)^{y_i} \left(1 - h_\theta(x_i)\right)^{1-y_i}$$

The cost function : $J = -\frac{1}{n} \log\left(L(\theta)\right)$

Logistic Regression

- Conclusion
  - Logistic regression
  - Uses sigmoid and log function and to estimate the parameters
  - According to the Maximum Likelihood Estimate, construct the loss function:

$$J = -\frac{1}{m}\log\left(L(\theta)\right)$$

where,

$$L(\theta) = \prod_{i}^{n} h_\theta(x_i)^{y_i}\left(1 - h_\theta(x_i)\right)^{1-y_i}$$

  - Minimize the cost:

$$\frac{\partial J}{\partial \theta} = 0 \quad \rightarrow \quad \text{solve for } \theta$$

HOW?

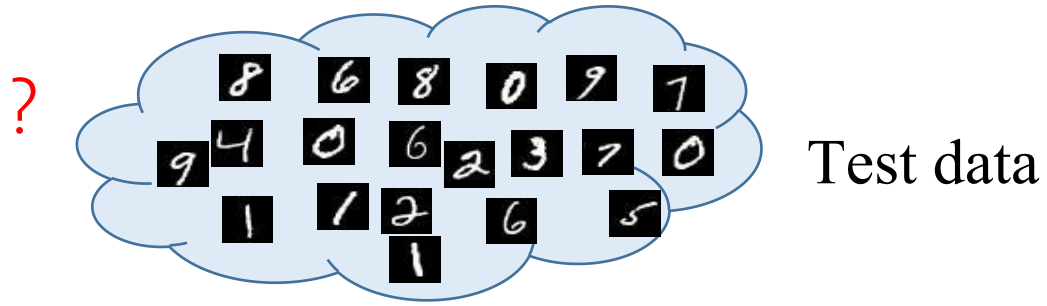
Try to solve it by yourself.

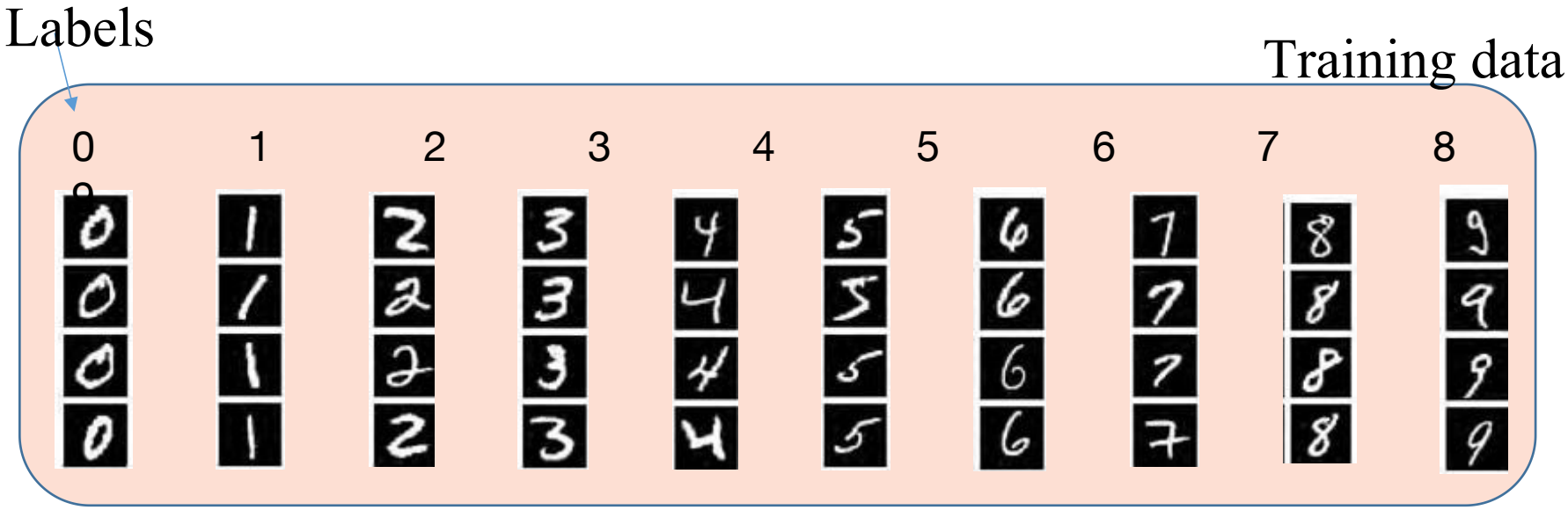
Supervised learning

- Linear Regression
- Logistic Regression
- *Classification*
  - *Distance-based algorithms*
  - Linear classifiers
  - Other classifiers
- ……

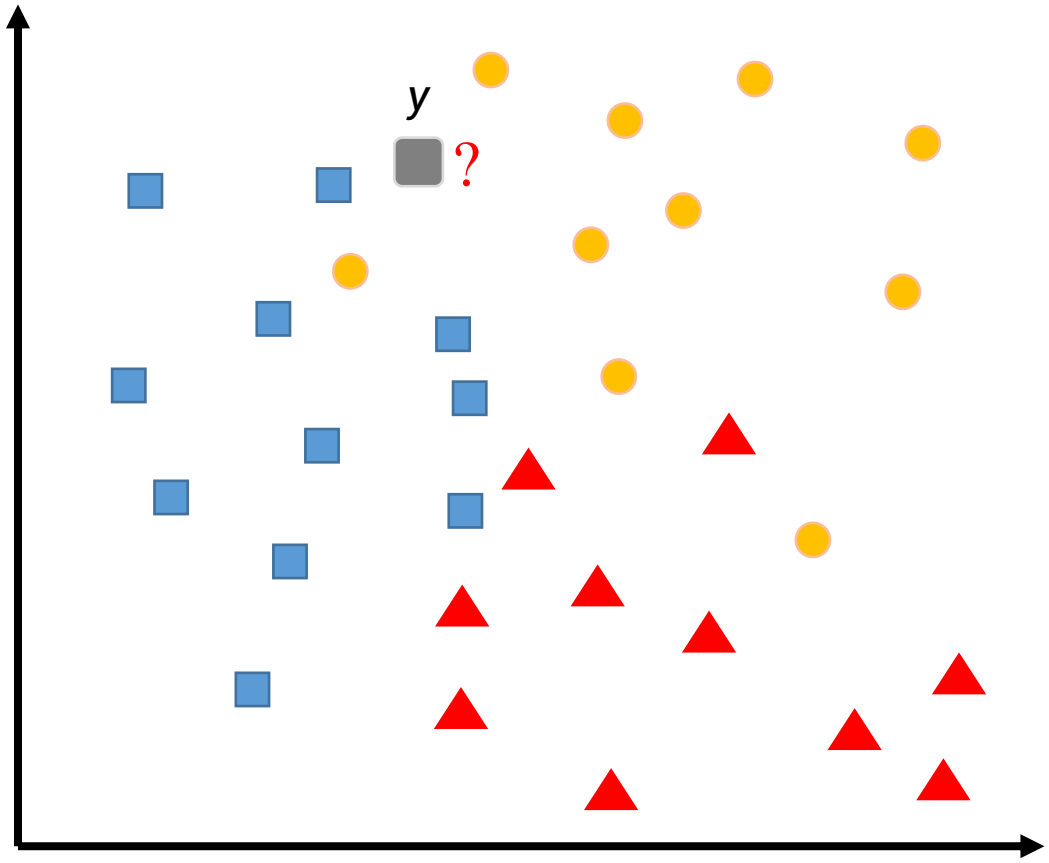
Classification

Multi-class classification assigns test samples to a certain class.





Classification



Training data:

$$X = \{x^{(1)}, x^{(2)}, \ldots, x^{(N)}\}$$

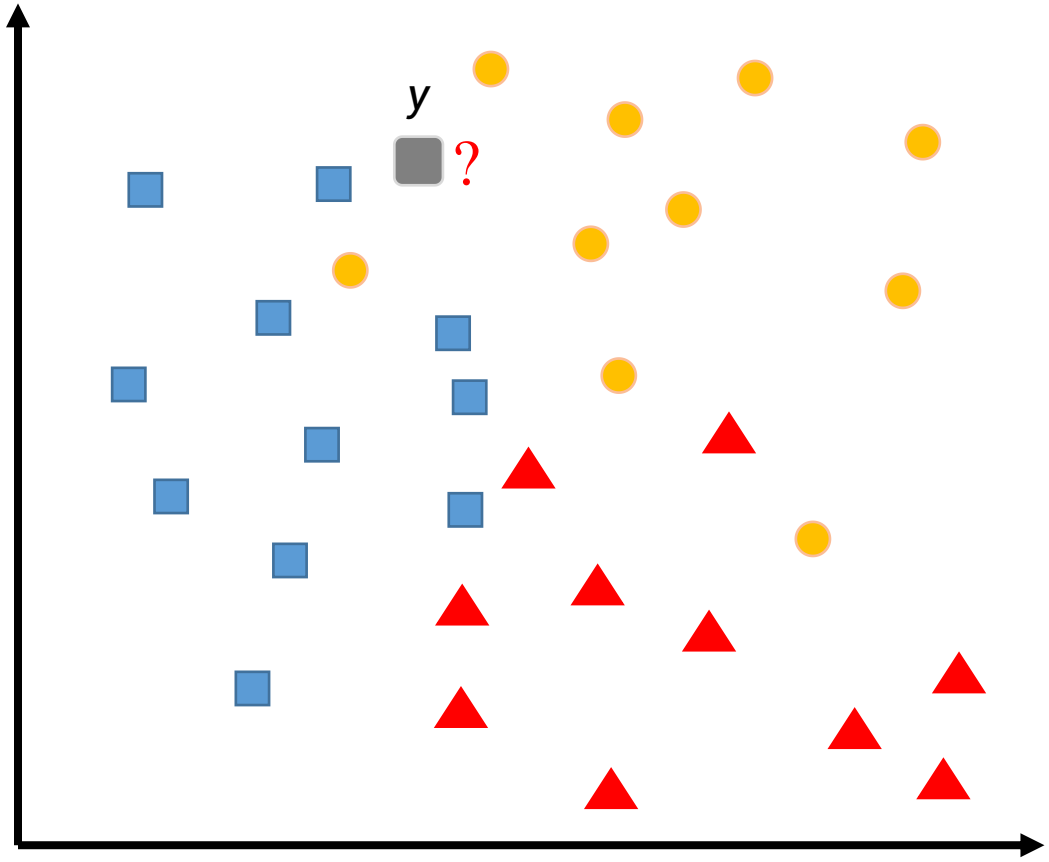
and
training labels:

$$L = \{l^{(1)}, l^{(2)}, \ldots, l^{(N)}\}$$

N: the *number of training data*

Classification

- Nearest neighbor



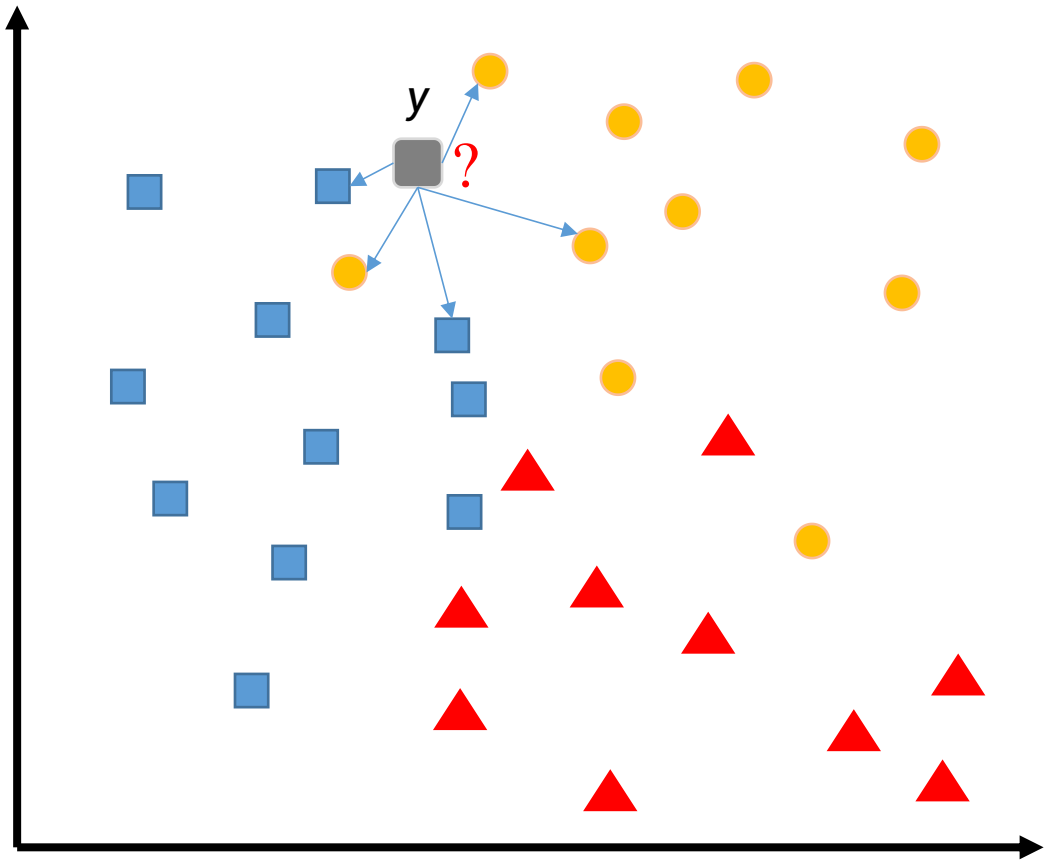
How to decide which is the nearest one?

The distance $d(\boldsymbol{x}, \boldsymbol{y})$ between two points $\boldsymbol{x} \in R^n$ and $\boldsymbol{y} \in R^n$ can for example be measured by the Euclidean distance.

$$d(x^{(1)}, x^{(2)}) = \sqrt{\sum_{i=1}^{n} \left(x_i^{(1)} - x_i^{(2)}\right)^2}$$

Classification

- Nearest neighbor



How to decide which is the nearest

$$d^j(x^{(y)}, y) = \sqrt{\sum_{i=1}^{n} \left(x_i^{(j)} - y\right)^2}$$

Calculate all the distances from the training data to the test data y, and we obtain:

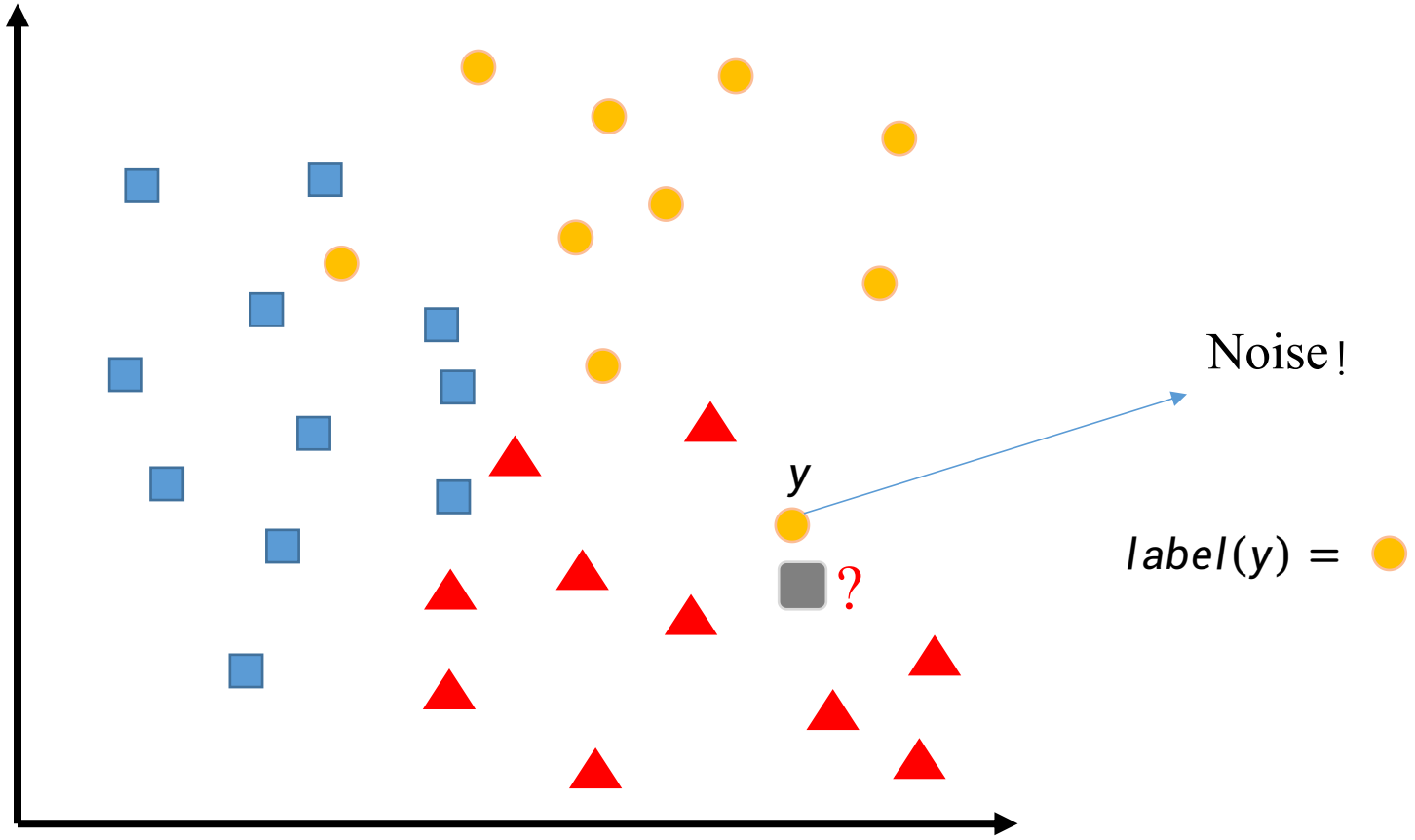
$$D = [d^{(1)}, d^{(2)}, \ldots, d^{(N)}]$$

$$s = \operatorname{argmin}_i d^{(i)}$$

$$label(y) = label(x^{(s)}) =$$ 

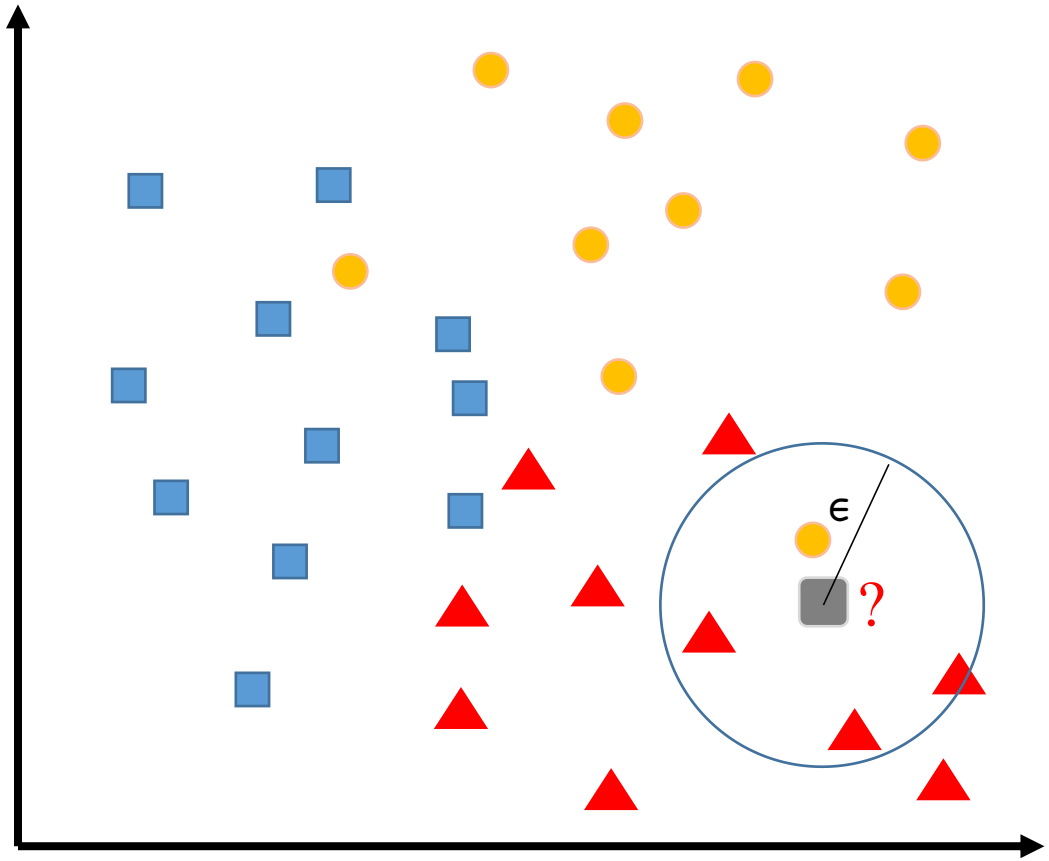
Classification

- Nearest neighbor



Classification

- ϵ-ball Nearest neighbor



Select a value ϵ, then draw a ball in R^n with y as the center and ϵ as the radius.

The label of y is decided by majority labels of points in this ball.

In this ball:

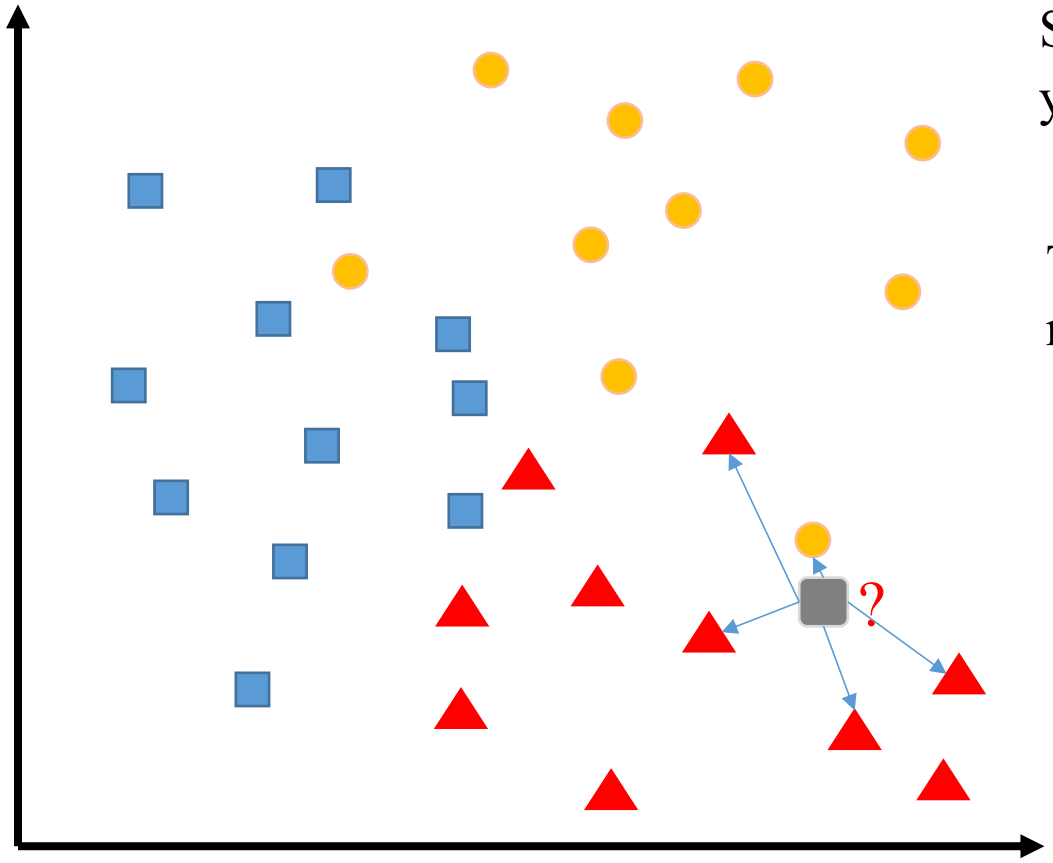
▲ : 3

● : 1

■ belongs to ▲

Classification

- K Nearest neighbor



Select a value *k*, then find y's k nearest neighbor.

The label of y is decided by majority labels of y's k neighbors.

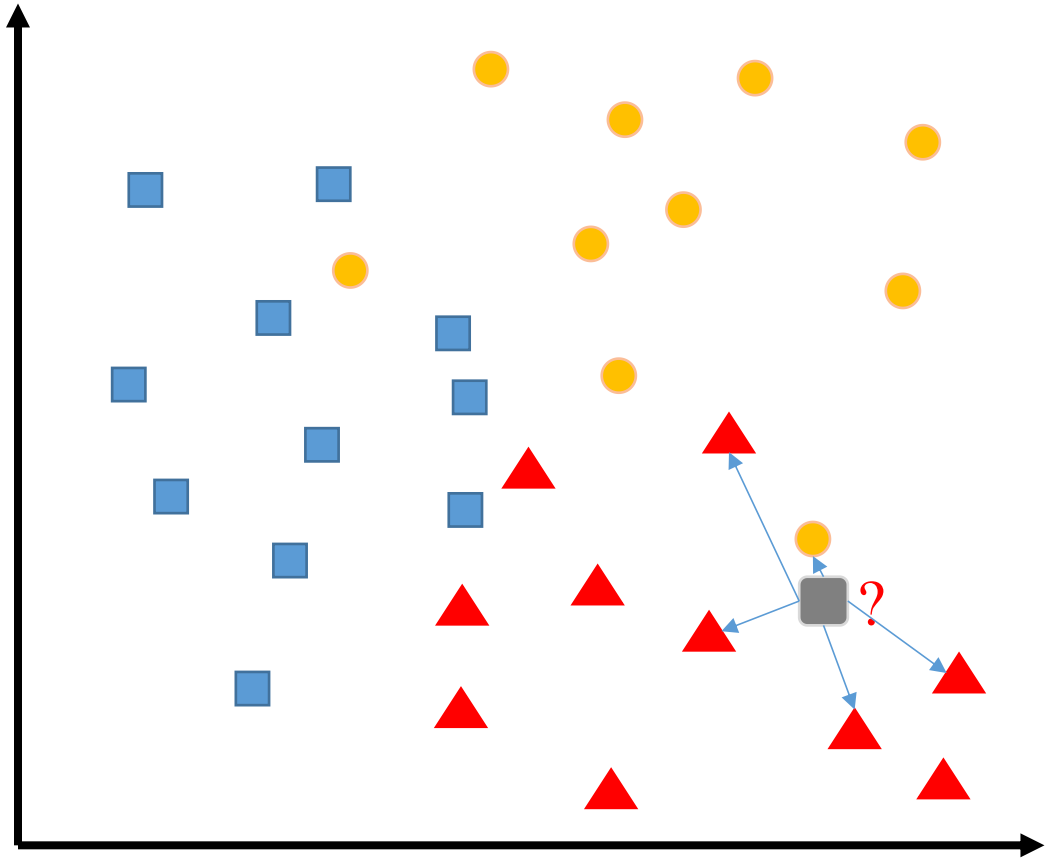
Let k be 5,

▲ : 5 ● : 1

■ belongs to ▲

Classification

- K Nearest neighbor



Question:
How to decide k?
Which algorithm achieve better performance?

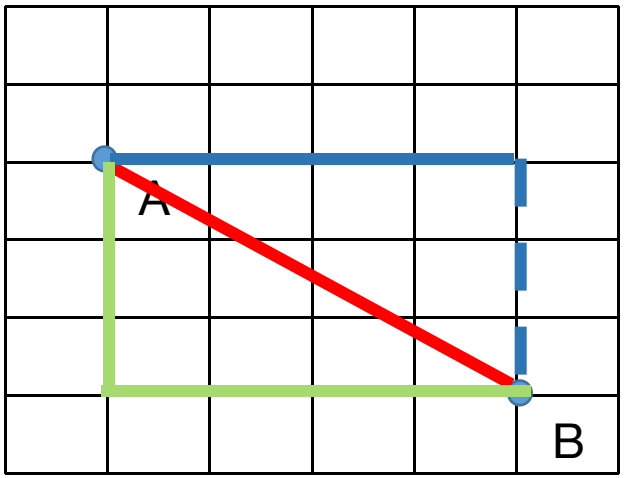
: 5

: 1

belongs to

Classification

- Distance Metrics
  - Euclidean distance
  - $d_e(x, y) = \sqrt{\sum_{i=1}^{n} (x_i - y_i)^2}$
  - Sum of squared distance
  - $d_q(x, y) = \sum_{i=1}^{n} (x_i - y_i)^2$
  - Manhattan distance
  - $d_m(x, y) = \sum_{i=1}^{n} |x_i - y_i|$
  - Chebyshev distance
  - $d_c(x, y) \max_{i=1,\cdots,n} |x_i - y_i|$



Classification

- Nearest neighbor classifier

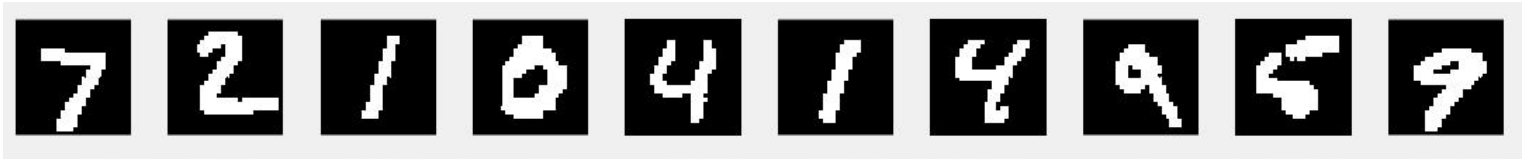
  Problem:

  - Need to determine value of parameter K
  - Distance based learning is not clear which type of distance to use and which attribute to use to produce the best results.
  - Computation cost is quite high because we need to compute distance of each query instance to all training samples.

Classification

- Example
  - Each image is represented by a vector of dimension 784.

The matrix indicates the pairwise distances.



7	0	2.8735	2.1766	2.6559	2.2201	2.2500	2.0893	2.4795	2.8443	2.1202
2	2.8735	0	2.5055	2.8681	2.9475	2.6062	2.8493	2.8330	2.9434	3.1619
1	2.1766	2.5055	0	2.9024	2.3556	0.7858	2.3561	2.2060	2.5274	2.4331
0	2.6559	2.8681	2.9024	0	2.7428	2.9531	3.0539	2.8362	2.8488	2.6425
4	2.2201	2.9475	2.3556	2.7428	0	2.5284	2.1733	2.4262	2.3432	2.5895
1	2.2500	2.6062	0.7858	2.9531	2.5284	0	2.4679	2.2906	2.5549	2.3900
4	2.0893	2.8493	2.3561	3.0539	2.1733	2.4679	0	2.5580	2.7456	2.3759
9	2.4795	2.8330	2.2060	2.8362	2.4262	2.2906	2.5580	0	2.8885	2.5823
5	2.8443	2.9434	2.5274	2.8488	2.3432	2.5549	2.7456	2.8885	0	2.9773
9	2.1202	3.1619	2.4331	2.6425	2.5895	2.3900	2.3759	2.5823	2.9773	0

The distance between the data is inconsistent with similarity of the content of the image .

Supervised learning

- Linear Regression
- Logistic Regression
- Classification
  - Distance-based algorithms
  - *Linear classifiers*
  - Other classifiers
- ……

Classification

- Linearly separable

Apple = [diameter, color, shape, spots, place of production]

$A_1 = \begin{bmatrix} 7.8 \\ 0.2 \end{bmatrix}$

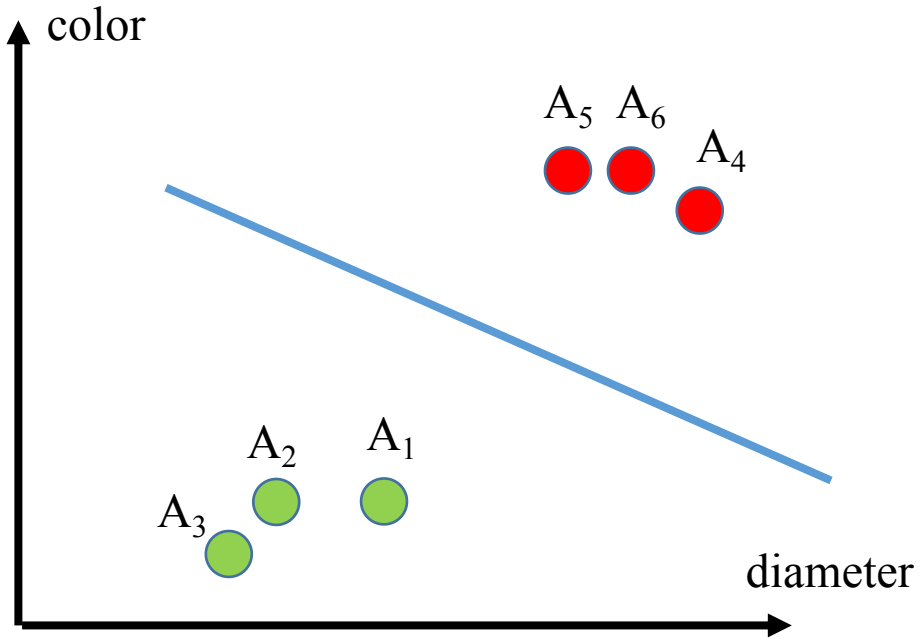
$A_2 = \begin{bmatrix} 7.4 \\ 0.2 \end{bmatrix}$

$A_3 = \begin{bmatrix} 7.1 \\ 0.1 \end{bmatrix}$

$A_4 = \begin{bmatrix} 8.5 \\ 0.7 \end{bmatrix}$

$A_5 = \begin{bmatrix} 8.1 \\ 0.8 \end{bmatrix}$

$A_6 = \begin{bmatrix} 8.3 \\ 0.8 \end{bmatrix}$



These training data are *linearly separable*

Classification

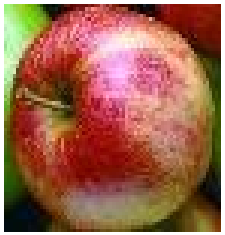
- Linearly separable

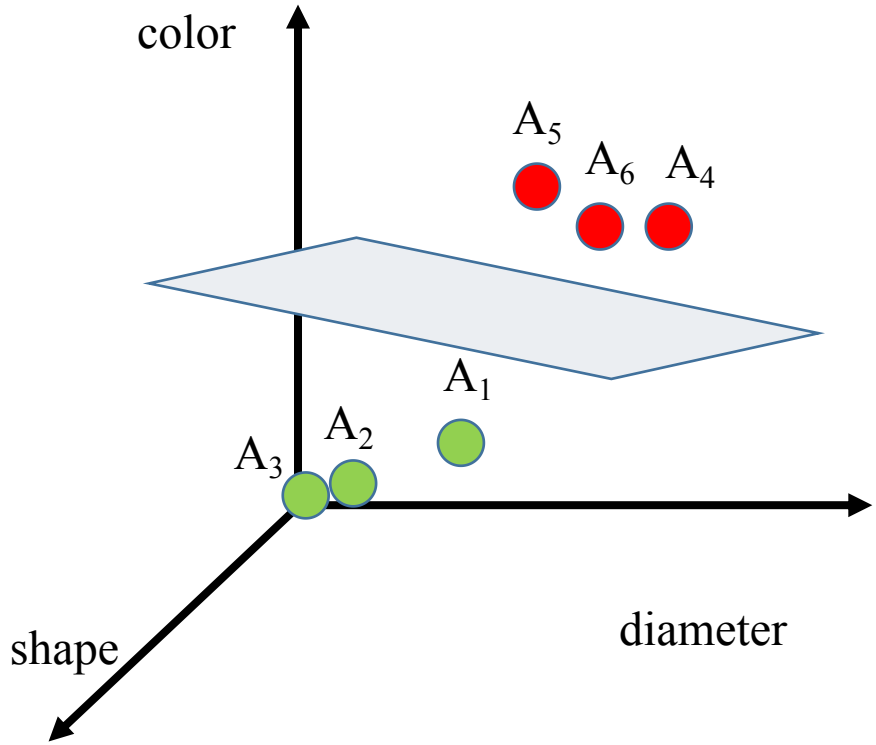
Apple = [diameter, color, shape, spots, place of production]

$A_1 = \begin{bmatrix} 7.8 \\ 0.2 \\ 0.6 \end{bmatrix}$

$A_2 = \begin{bmatrix} 7.4 \\ 0.2 \\ 0.7 \end{bmatrix}$

$A_3 = \begin{bmatrix} 7.1 \\ 0.1 \\ 0.6 \end{bmatrix}$

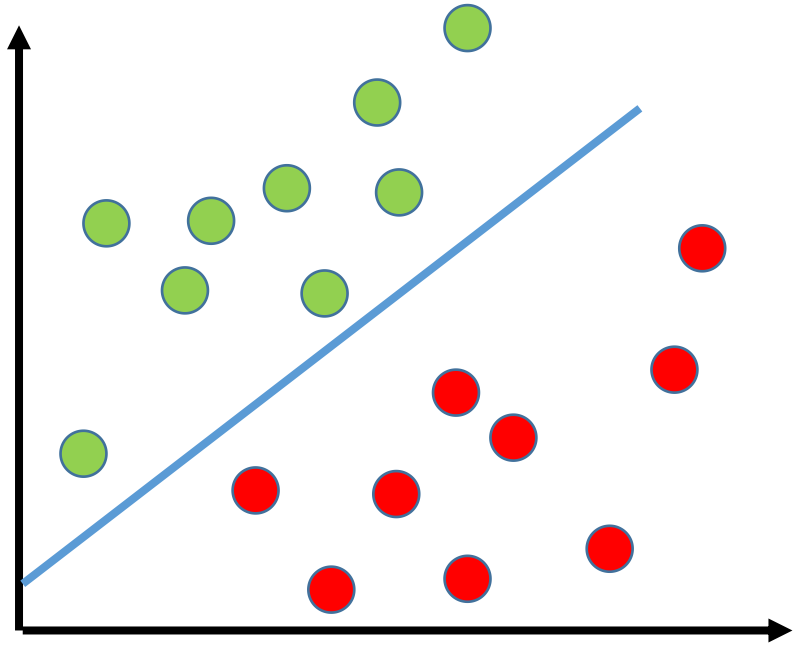




In *n* dimensions a hyperplane is needed for the separation.

Classification

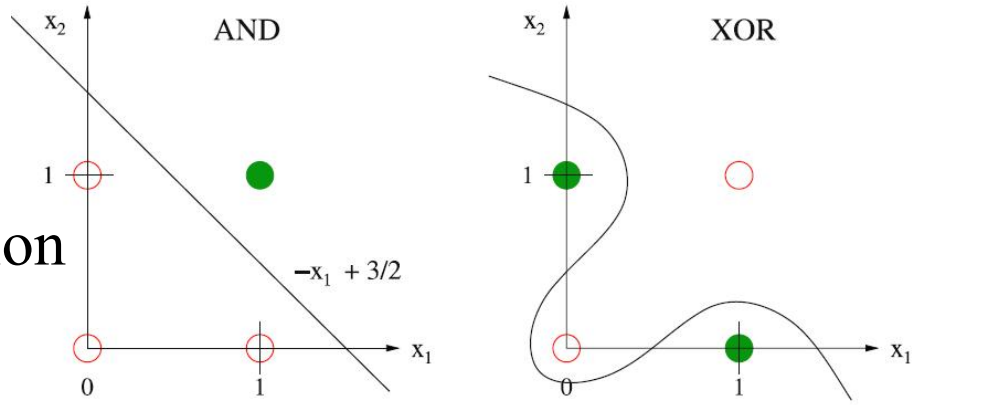
- Linearly separable



A linearly separable two dimensional data set. The equation for the dividing straight line is $w_1x_1+w_2x_2=1$

Every $(n-1)$-dimensional hyperplane in R^n can be described by an equation

$$\sum_{i=1}^{n} w_i x_i + b = 0$$



The boolean function AND is linearly separable, but XOR is not (● ≙ true, ○ ≙ false)

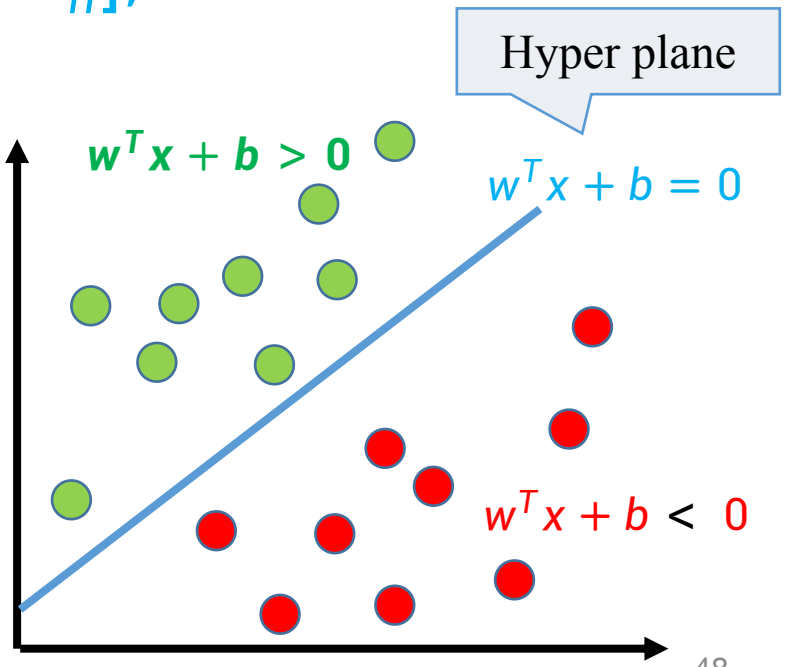
Classification

- Linearly separable

  - **Definition** Two sets $M_1 \subset R^n$ and $M_2 \subset R^n$ are called *linearly separable.*
  - if real vector $w=[w_1, w_2, \cdots, w_n], b$ exist with

$$w^T x + b > 0 \; for \; all \; x \in M_1$$
$$and$$
$$w^T x + b \leq 0 \; for \; all \; x \in M_2$$

$$classify(x; w, b) = \text{sign}(w\,x + b)$$



Classification

Linearly separable

- Given a training set which is linearly separable:

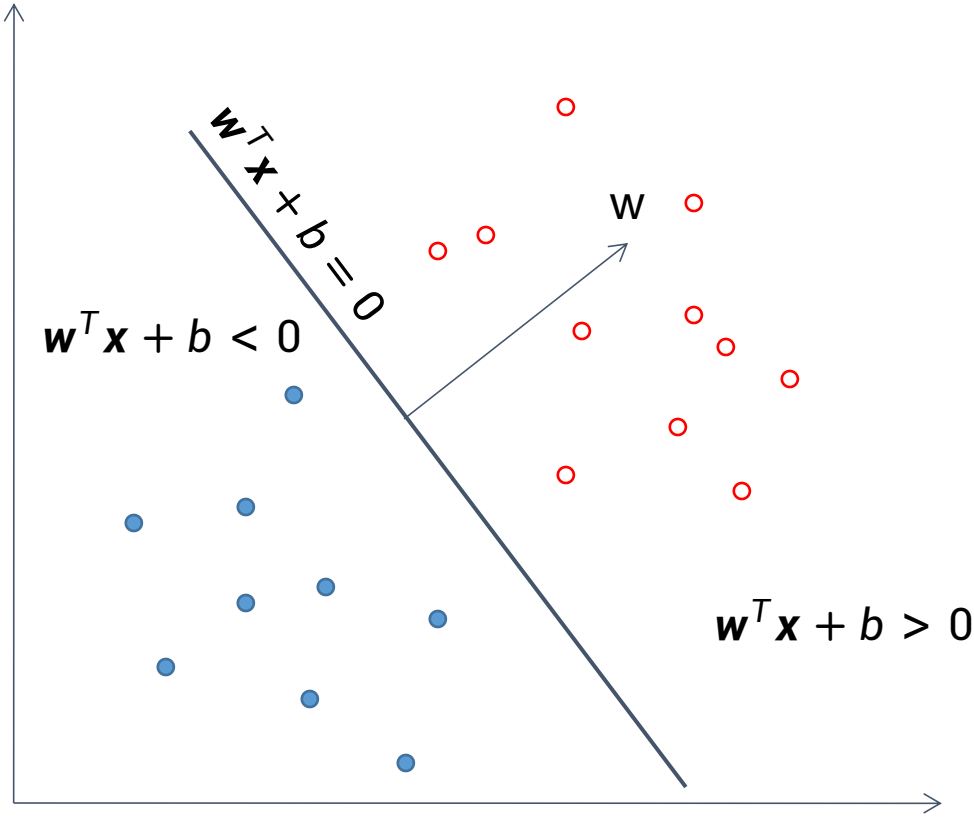
$$D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\},$$

$$\boldsymbol{x}_i \in X = R^d,$$

$$y_i \in Y = \{-1, +1\}, \quad i = 1,2,\ldots,m$$

- Goal: solve a separating hyperplane

$$g(\boldsymbol{x}) = w^T x + b = 0$$

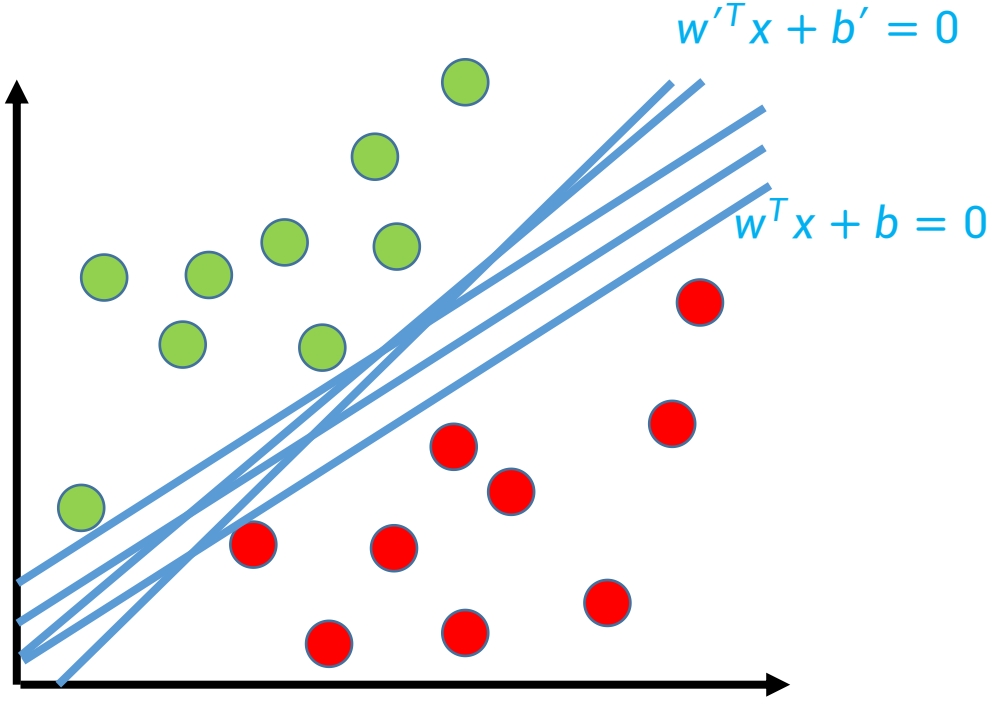


Classification

- Linearly separable

Any of these would be fine.

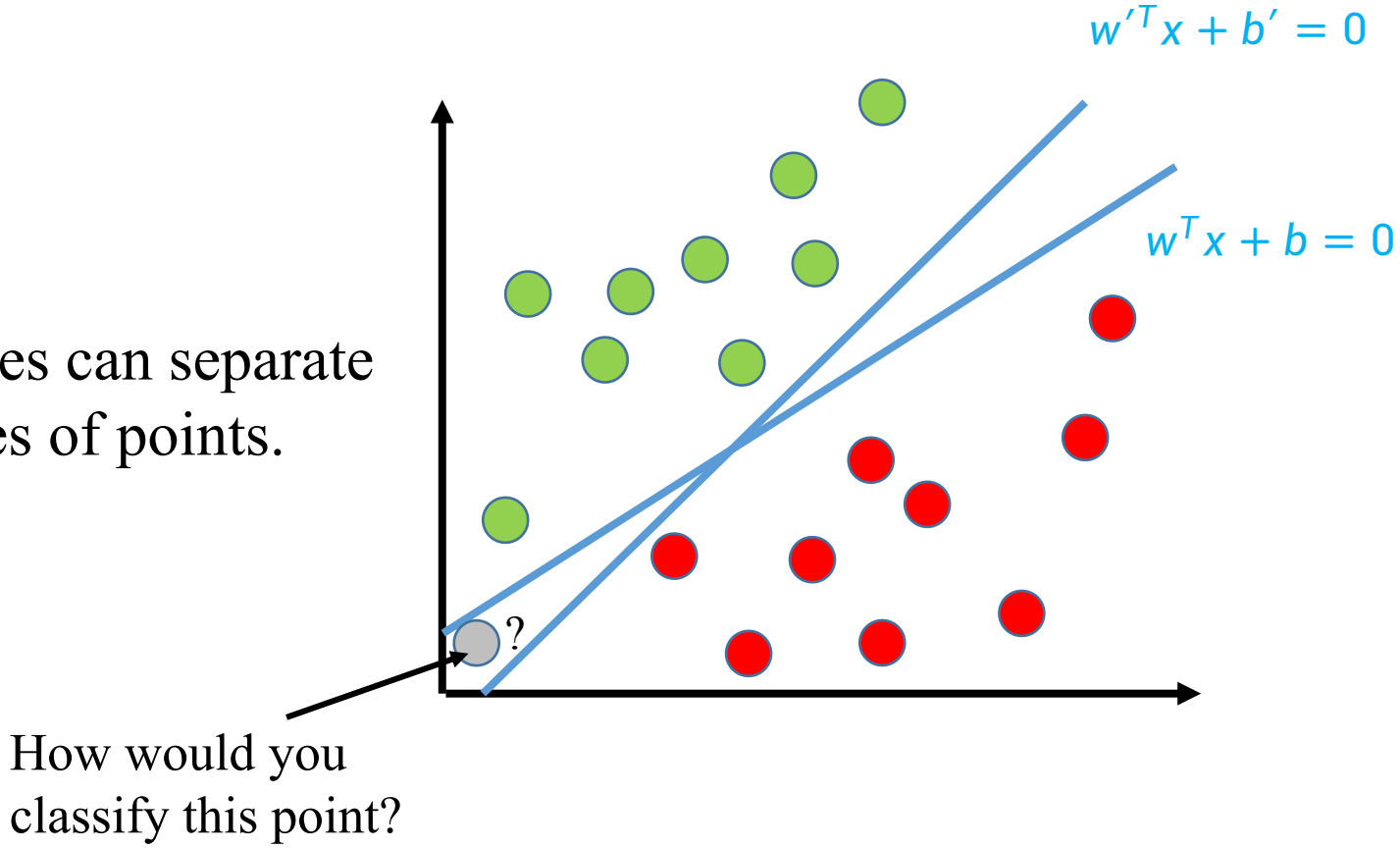
But which is best?



Classification

- Linearly separable

These two lines can separate the two classes of points.

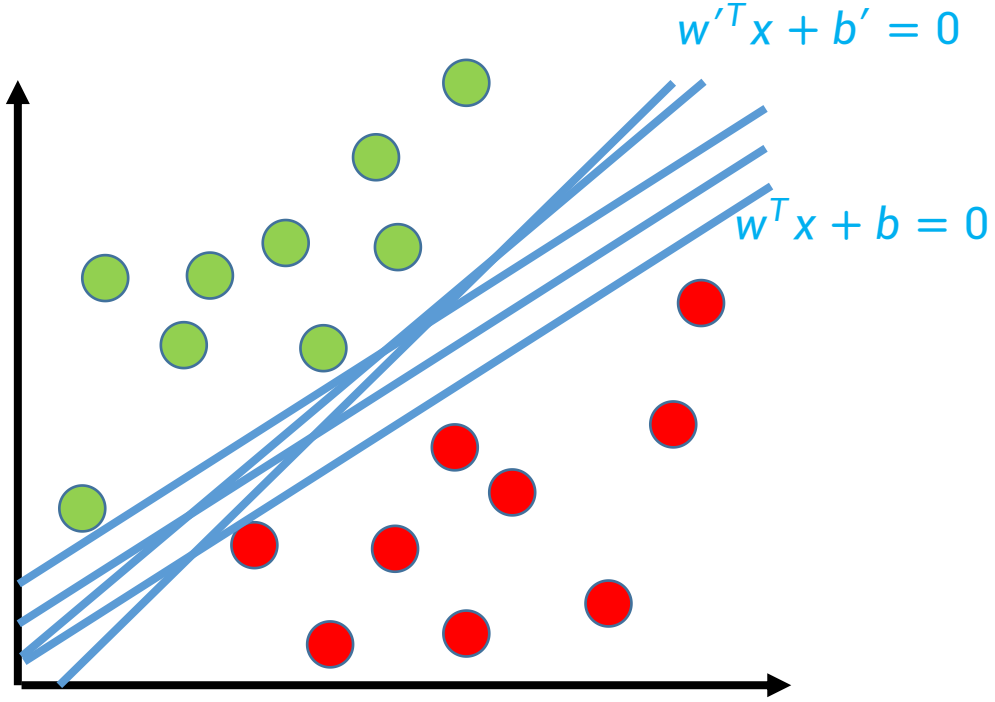


Classification

- Linearly separable

The distance between a sample and hyperplane indicates the classification confidence:

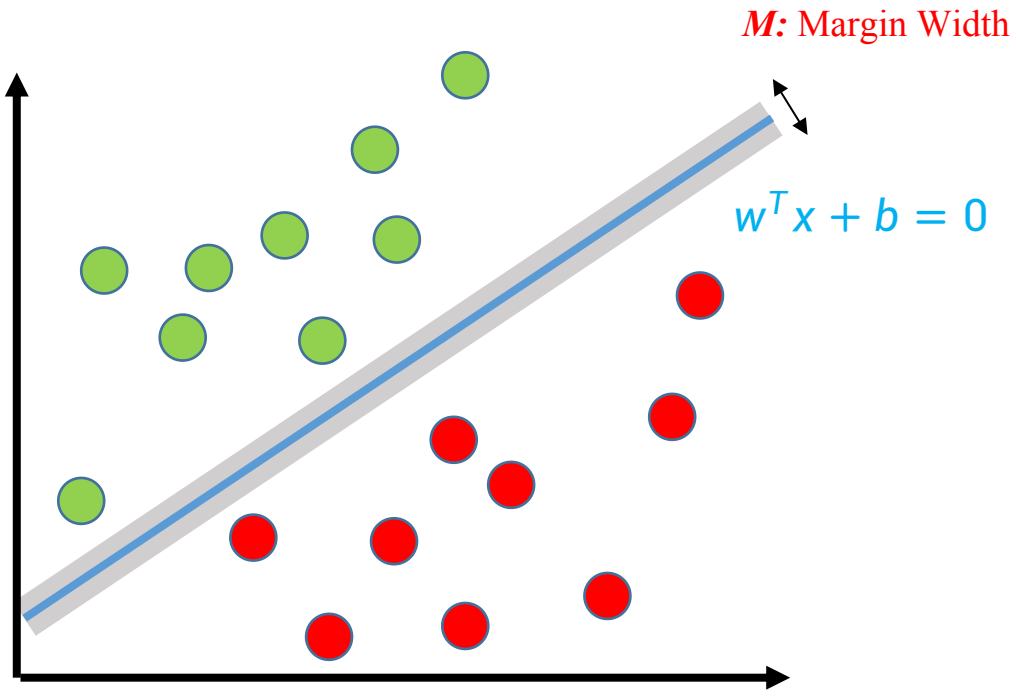
- The farther the sample is from the hyperplane, the higher the confidence that it will be correctly classified.
- The closer the sample is to the hyperplane, the lower the confidence that it will be correctly classified.



Classification

- Linearly separable

Define the **margin** of a linear classifier as the width that the boundary could be increased by before hitting a data point.

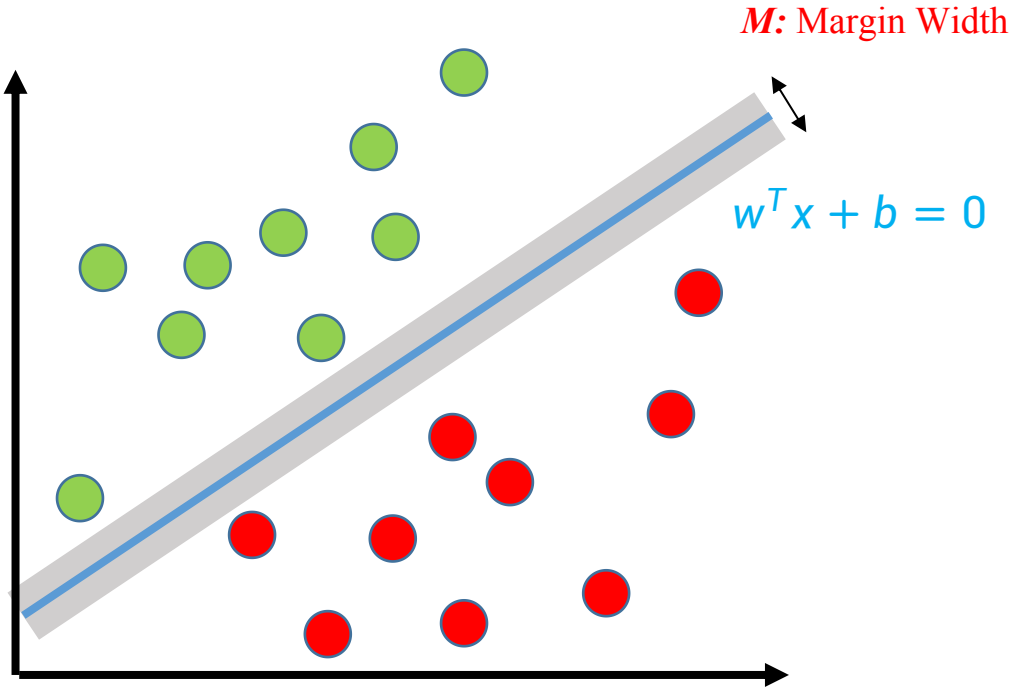


Classification

Linearly separable

Hyperplane with maximum margin: the hyperplane separating samples with maximum margin, which is more robust for classification.

- Two-class samples are separating on corresponding side of the hyperplane;
- The distance from the closest sample point to the hyperplane on both sides to the hyperplane is maximized.

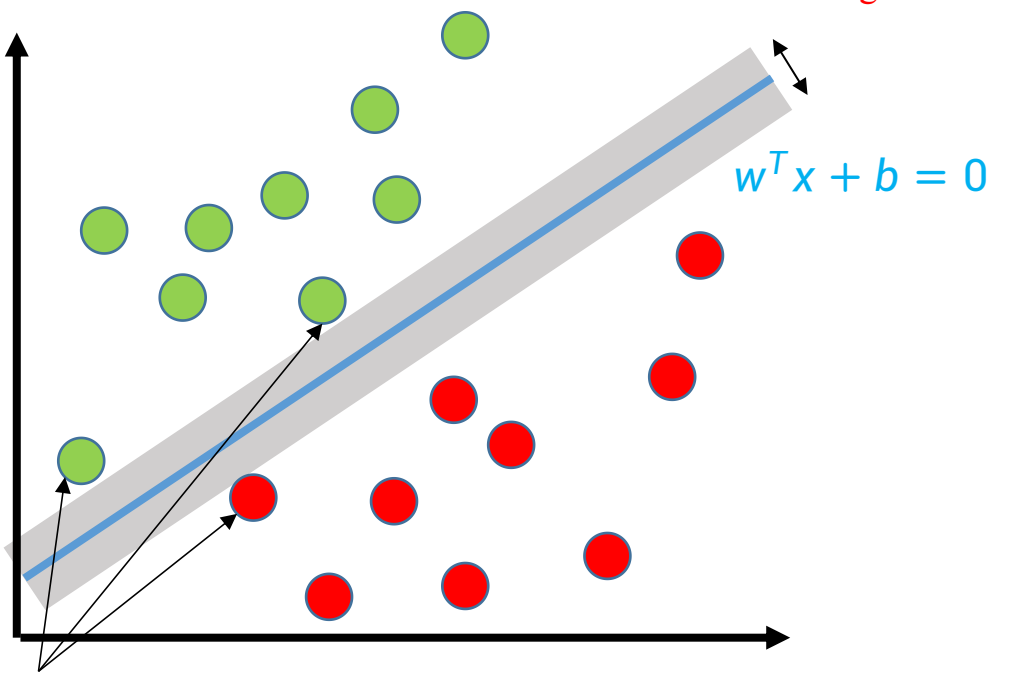


Support Vector Machine (SVM): the optimal separating hyperplane when samples are linearly separable.

Classification

- Support vector machine (SVM)

1. Maximizing the margin is good.
2. Implies that only support vectors are important; other training examples are ignorable.
3. Empirically it works very well.



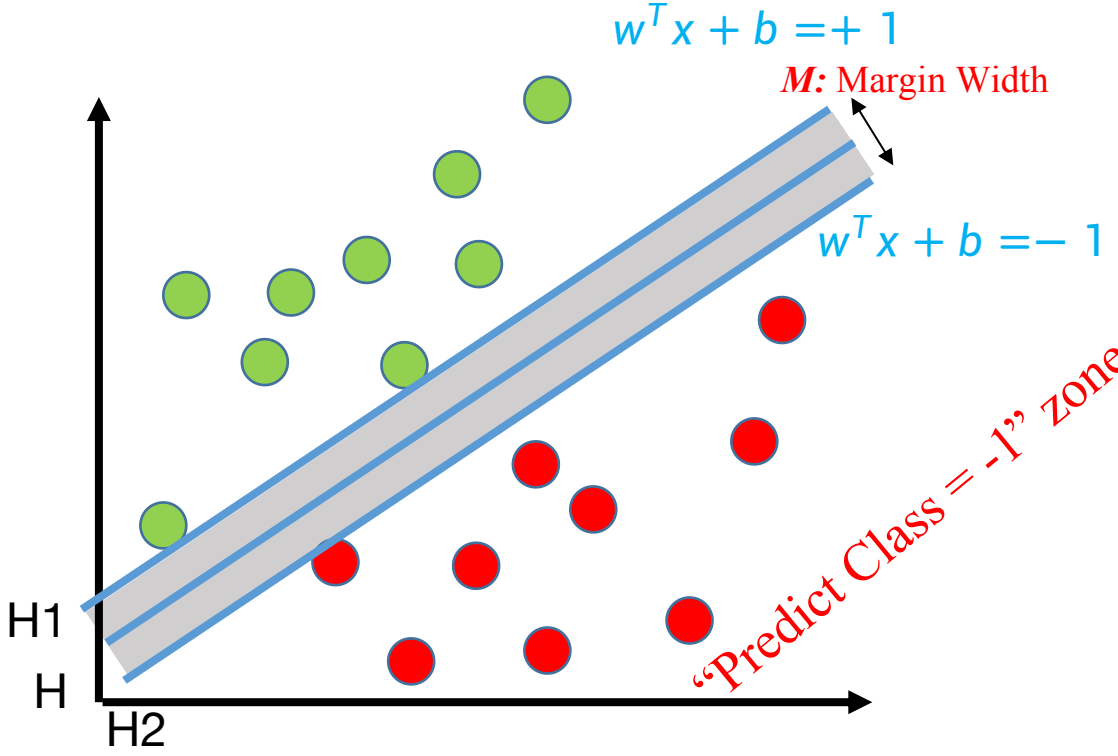
Support Vectors: are those data points that the margin pushes up against.

This is the simplest kind of SVM (Called an LSVM)

Classification

Linearly separable

- **Margin**: H1 and H2 are boundary hyperplanes that pass through the samples closest to the H and parallel to H. The distance between H1 and H2 is called separating margin.
- **Optimal separating hyperplane:** A hyperplane separating samples correctly, and samples (two-class) closest to the hyperplane also has the maximum distance from the hyperplane.



Classification

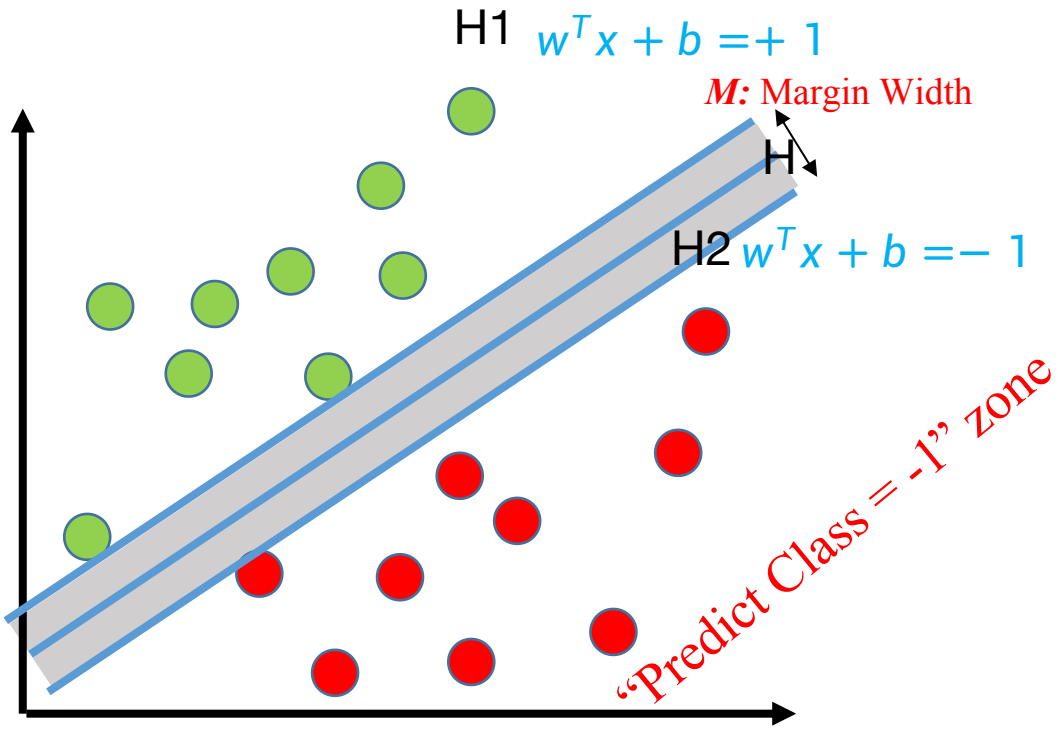
- Linearly separable

For any line for classification, we can calculate the margin.

Then, which line is the best?

Goal:
1) Correctly classify all training data
2) Maximize the Margin

How to achieve this goal?



Classification

SVM

- Given a training set which is linearly separable: $D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_m, y_m)\}, \quad \boldsymbol{x}_i \in X = R^d, \quad y_i \in Y = \{-1, +1\}$
- Hyperplane H: $\boldsymbol{w}^T \boldsymbol{x} + b = 0$
- The distance between any sample $\boldsymbol{x}$ in feature space to H:

$$r = \frac{|\boldsymbol{w}^T \boldsymbol{x} + b|}{\|\boldsymbol{w}\|}$$

Classification

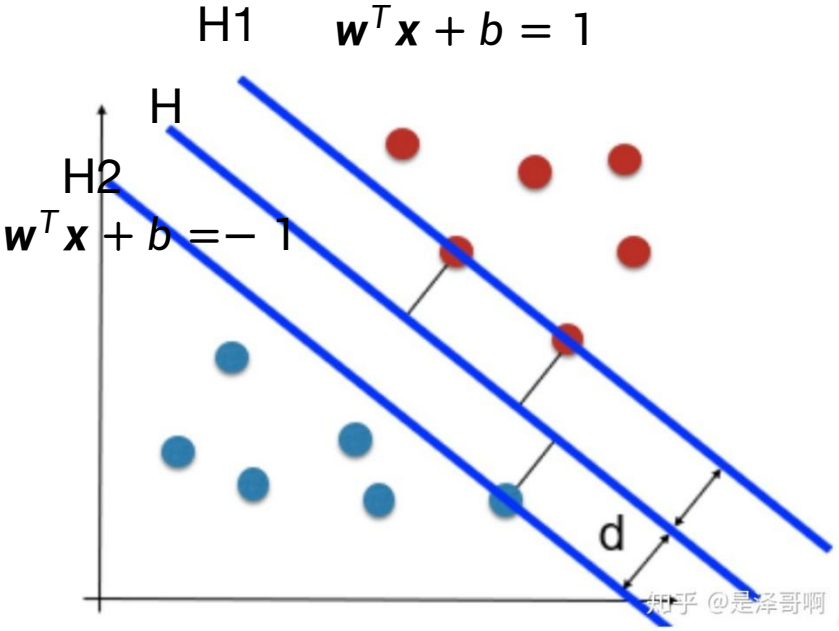
SVM-- Goal 1

- Linearly separable

$$\begin{cases} \boldsymbol{w}^T\boldsymbol{x}_i + b > 0, y_i = +1 \\ \boldsymbol{w}^T\boldsymbol{x}_i + b < 0, y_i = -1 \end{cases}$$

- Linearly separable sample with high confidence and accuracy

$$\begin{cases} \frac{\boldsymbol{w}^T\boldsymbol{x}_i + b}{\|\boldsymbol{w}\|} \geq d, y_i = +1 \\ \frac{\boldsymbol{w}^T\boldsymbol{x}_i + b}{\|\boldsymbol{w}\|} \leq -d, y_i = -1 \end{cases}$$



$$\Rightarrow \begin{cases} \frac{\boldsymbol{w}^T\boldsymbol{x}_i + b}{\|\boldsymbol{w}\|d} \geq 1, y_i = +1 \\ \frac{\boldsymbol{w}^T\boldsymbol{x}_i + b}{\|\boldsymbol{w}\|d} \leq -1, y_i = -1 \end{cases}$$

$\|\boldsymbol{w}\|d=1$ Normalization

$$\begin{cases} \boldsymbol{w}^T\boldsymbol{x}_i + w_0 \geq 1, y_i = +1 \\ \boldsymbol{w}^T\boldsymbol{x}_i + w_0 \leq -1, y_i = -1 \end{cases}$$

$$y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1,$$

Then $|\boldsymbol{w}^T\boldsymbol{x}_i + w_0| = y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0)$

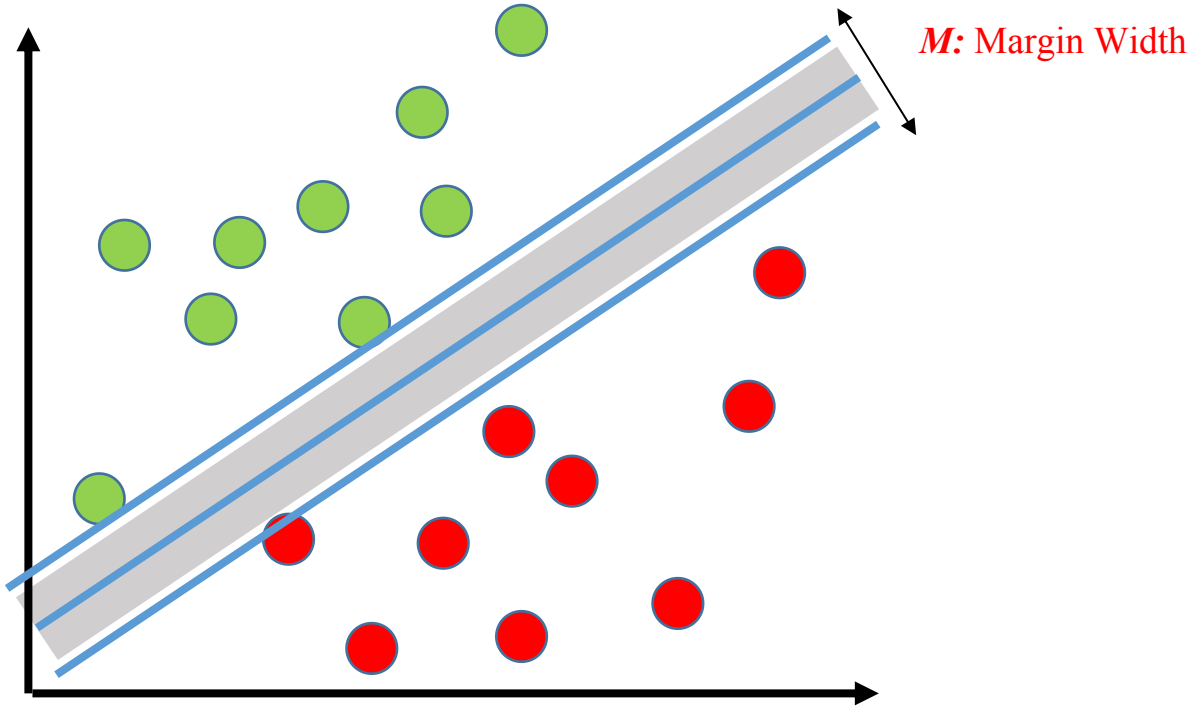
Classification

- SVM

Goal:

1) Correctly classify all training data:

$$y\,(w^T x + b) \geq 1 \; for \; all \; y$$

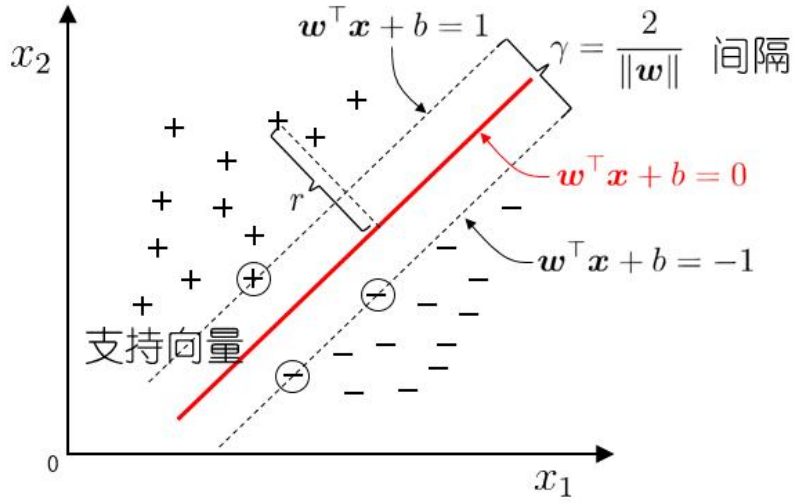


Classification

SVM

- Margin:

$$r = 2d = \frac{2}{\|\boldsymbol{w}\|}$$



$$w^T(x_i^+ - x_i^-) = 2$$

$$w^T(x_i^+ - x_i^-) = \|w\| \cdot \|x_i^+ - x_i^-\| \cdot cos\theta$$

$$= w^T(x_i^+ - x_i^-) = \|w\| \cdot \|x_i^+ - x_i^-\| \cdot \frac{r}{\|x_i^+ - x_i^-\|} = \|w\| \cdot r$$

- Maximum margin: solve $\boldsymbol{w}$ and b to get maximum margin γ

$$\max_{\boldsymbol{w},b} \frac{2}{\|\boldsymbol{w}\|} \; s.t.\; y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1,\; i = 1,2,\ldots,m$$

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$s.t.\; y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1,\; i = 1,2,\ldots,m$$

Classification

- SVM

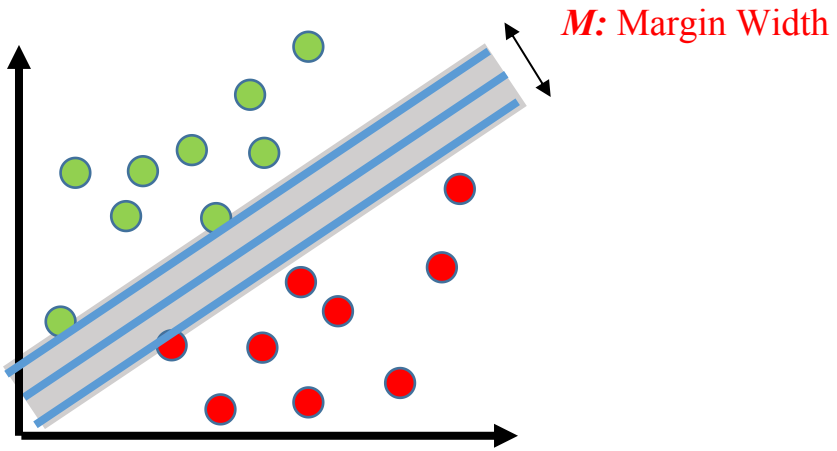
Goal:

1) Correctly classify all training data:

$$y_i\,(w^T x_i + b) \geq 1 \; for \; all \; i$$

2) Maximize the margin

$$\min_{w,b} \frac{1}{2}\|w\|^2$$



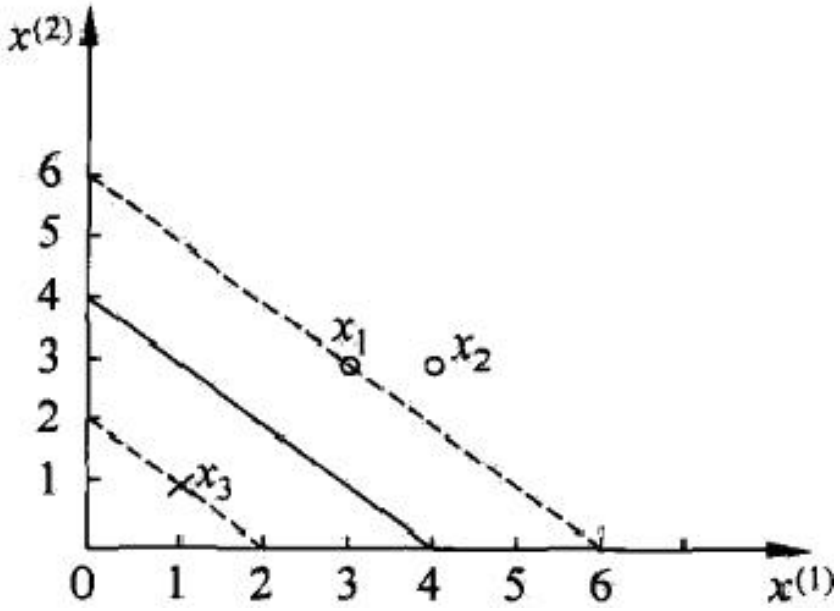
We can formulate a Quadratic Optimization Problem and solve for w and b

$$\text{Minimize } \Phi(w) = \frac{1}{2} w^T w$$
$$\text{subject to } y_i(wx_i + b) \geq 1 \; for \; all \; i$$

The primary problem of SVM: Solve $d+1$ variables (w, b) with m inequality constraints, which is suitable for low dimensions.

Example 1——Solve the primary problem of SVM

- Problem: Given a training set size of 3, in which (x_1, y_1) and (x_2, y_2) are positive samples, and (x_3, y_3) is negative sample.$x_1 = (3; 3), y_1 = +1, x_2 = (4; 3), y_2 = +1; x_3 = (1; 1), \quad y_3 = -1$. Solve the optimal separating hyperplane H with maximum margin.



Solve $\boldsymbol{w} = (w_1; w_2)$ and $w_0 (w_0 \, is \, b)$

Example 1——Solve the primary problem of SVM

Answer:

- Step 1: Build the primary problem of SVM upon the sample set

$$\min_{\boldsymbol{w},b} \frac{1}{2}(w_1^2 + w_2^2)$$
$$s.t. \begin{cases} 3w_1 + 3w_2 + w_0 \geq 1 \\ 4w_1 + 3w_2 + w_0 \geq 1 \\ -w_1 - w_2 + w_0 \geq 1 \end{cases}$$

$$\min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|^2$$
$$s.t.\ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1,\ i = 1,2,\ldots,m$$

Example 1——Solve the primary problem of SVM

- Step 2: Build Lagrange function by setting Lagrange multiplier $a_i \geq 0$ for each inequality constraint

$$L(w_1, w_2, w_0, a_1, a_2, a_3) = \frac{1}{2}(w_1^2 + w_2^2) - a_1(3w_1 + 3w_2 + w_0 - 1) - a_2(4w_1 + 3w_2 + w_0 - 1) - a_3(-w_1 - w_2 + w_0 - 1)$$

Set the partial as 0

$$\begin{cases} \frac{\partial L}{\partial w_1} = w_1 - 3a_1 - 4a_2 + a_3 = 0 \\ \frac{\partial L}{\partial w_2} = w_2 - 3a_1 - 3a_2 + a_3 = 0 \\ \frac{\partial L}{\partial w_0} = -a_1 - a_2 + a_3 = 0 \\ a_1(3w_1 + 3w_2 + w_0 - 1) = 0 \\ a_2(4w_1 + 3w_2 + w_0 - 1) = 0 \\ a_3(-w_1 - w_2 + w_0 - 1) = 0 \\ 3w_1 + 3w_2 + w_0 - 1 \geq 0 \\ 4w_1 + 3w_2 + w_0 - 1 \geq 0 \\ -w_1 - w_2 + w_0 - 1 \geq 0 \\ a_1 \geq 0, a_2 \geq 0, a_3 \geq 0 \end{cases} \Rightarrow \begin{cases} w_1 = 3a_1 + 4a_2 - a_3 \\ w_2 = 3a_1 + 3a_2 - a_3 \\ a_3 = a_1 + a_2 \\ a_1(3w_1 + 3w_2 + w_0 - 1) = 0 \\ a_2(4w_1 + 3w_2 + w_0 - 1) = 0 \\ a_3(-w_1 - w_2 + w_0 - 1) = 0 \\ 3w_1 + 3w_2 + w_0 - 1 \geq 0 \\ 4w_1 + 3w_2 + w_0 - 1 \geq 0 \\ -w_1 - w_2 + w_0 - 1 \geq 0 \\ a_1 \geq 0, a_2 \geq 0, a_3 \geq 0 \end{cases} \Rightarrow$$

Example 1——Solve the primary problem of SVM

- Step 3: only keep a_1, a_2, w_0

$$\begin{cases} w_1 = 3a_1 + 4a_2 - a_3 = 2a_1 + 3a_2 \\ w_2 = 3a_1 + 3a_2 - a_3 = 2a_1 + 2a_2 \\ a_3 = a_1 + a_2 \\ a_1(12a_1 + 15a_2 + w_0 - 1) = 0 \\ a_2(14a_1 + 18a_2 + w_0 - 1) = 0 \\ (a_1 + a_2)(-4a_1 - 5a_2 - w_0 - 1) = 0 \\ 12a_1 + 15a_2 + w_0 - 1 \geq 0 \\ 14a_1 + 18a_2 + w_0 - 1 \geq 0 \\ -4a_1 - 5a_2 - w_0 - 1 \geq 0 \\ a_1 \geq 0, a_2 \geq 0, a_3 \geq 0 \end{cases}$$

(1) If $a_1 = 0, a_2 = 0$, then

$w_1 = w_2 = 0$

However, $w_0 \geq 1$ and $w_0 \leq -1$ are contradictory!!!

(2) If $a_1 > 0, a_2 = 0$, then

$$\begin{cases} 12a_1 + w_0 - 1 = 0 \\ -4a_1 - w_0 - 1 = 0 \end{cases}$$

We will have

$$\begin{cases} a_1 = a_3 = \frac{1}{4} \\ a_2 = 0 \\ w_0 = -2 \\ w_1 = w_2 = \frac{1}{2} \end{cases}$$

Which satisfy all inequality constraints!

Example 1——Solve the primary problem of SVM

- Step 3: only keep a_1, a_2, w_0

$$\begin{cases} w_1 = 3a_1 + 4a_2 - a_3 = 2a_1 + 3a_2 \\ w_2 = 3a_1 + 3a_2 - a_3 = 2a_1 + 2a_2 \\ a_3 = a_1 + a_2 \\ a_1(12a_1 + 15a_2 + w_0 - 1) = 0 \\ a_2(14a_1 + 18a_2 + w_0 - 1) = 0 \\ (a_1 + a_2)(-4a_1 - 5a_2 - w_0 - 1) = 0 \\ 12a_1 + 15a_2 + w_0 - 1 \geq 0 \\ 14a_1 + 18a_2 + w_0 - 1 \geq 0 \\ -4a_1 - 5a_2 - w_0 - 1 \geq 0 \\ a_1 \geq 0, a_2 \geq 0, a_3 \geq 0 \end{cases}$$

(3) If $a_1 = 0, a_2 > 0$, then

$$\begin{cases} 18a_2 + w_0 - 1 = 0 \\ -5a_2 - w_0 - 1 = 0 \end{cases}$$

We will have

$$\begin{cases} a_1 = 0 \\ a_2 = a_3 = \frac{2}{13} \\ w_0 = -\frac{23}{13} \\ w_1 = \frac{6}{13} \\ w_2 = \frac{4}{13} \end{cases}$$

Substitute in the inequality constraints to get the following expression

$$12a_1 + 15a_2 + w_0 - 1 = -\frac{6}{13} < 0$$

Which violates the constraint!

Example 1——Solve the primary problem of SVM

- Step 3: only keep a_1, a_2, w_0

$$\begin{cases} w_1 = 3a_1 + 4a_2 - a_3 = 2a_1 + 3a_2 \\ w_2 = 3a_1 + 3a_2 - a_3 = 2a_1 + 2a_2 \\ a_3 = a_1 + a_2 \\ a_1(12a_1 + 15a_2 + w_0 - 1) = 0 \\ a_2(14a_1 + 18a_2 + w_0 - 1) = 0 \\ (a_1 + a_2)(-4a_1 - 5a_2 - w_0 - 1) = 0 \\ 12a_1 + 15a_2 + w_0 - 1 \geq 0 \\ 14a_1 + 18a_2 + w_0 - 1 \geq 0 \\ -4a_1 - 5a_2 - w_0 - 1 \geq 0 \\ a_1 \geq 0, a_2 \geq 0, a_3 \geq 0 \end{cases}$$

(4) If $a_1 > 0, a_2 > 0$, then

$$\begin{cases} 12a_1 + 15a_2 + w_0 - 1 = 0 \\ 14a_1 + 18a_2 + w_0 - 1 = 0 \\ -4a_1 - 5a_2 - w_0 - 1 = 0 \end{cases}$$

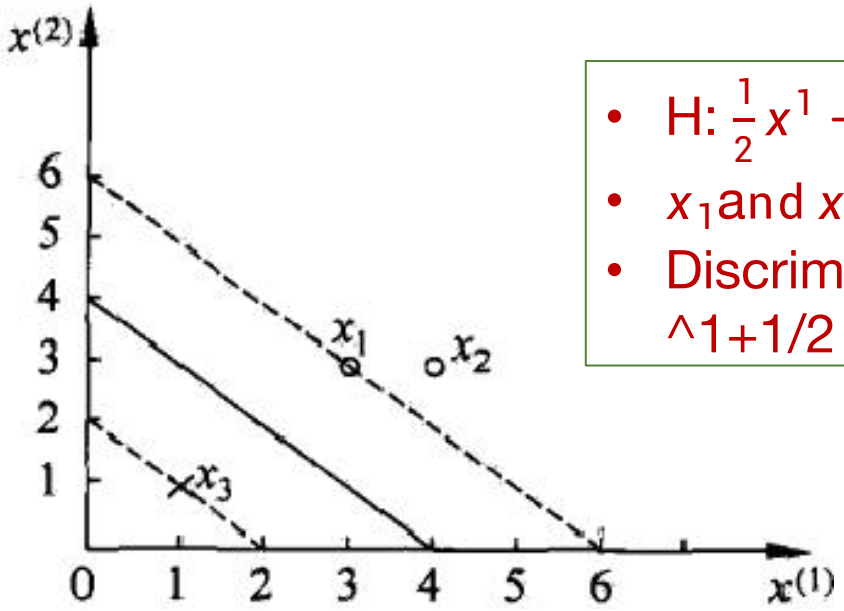
We will have

$$\begin{cases} a_1 = \frac{3}{2} \\ a_2 = -1 < 0 \\ a_3 = \frac{1}{2} \\ w_0 = -2 \\ w_1 = 0 \\ w_2 = 1 \end{cases}$$

Which violates the constraint!

Example 1——Solve the primary problem of SVM

- Problem: Given a training set size of 3, in which (x_1, y_1) and (x_2, y_2) are positive samples, and (x_3, y_3) is negative sample.$x_1 = (3; 3), y_1 = +1, x_2 = (4; 3), y_2 = +1; x_3 = (1; 1), \quad y_3 = -1$. Solve the optimal separating hyperplane H with maximum margin.



- H: $\frac{1}{2}x^1 + \frac{1}{2}x^2 - 2 = 0$
- $x_1$and x_3 are support vectors: $y_i g(\boldsymbol{x}_i) = 1$
- Discriminant Function: $g(x)$=sign(1/2 x ^1+1/2 x^2–2)

- m inequality constraints means 2^m cases!!

Primary Problem → Duel Problem

SVM-Duel Problem

- **The primary problem of SVM**

$$\min_{\boldsymbol{w},w_0} \frac{1}{2}\|\boldsymbol{w}\|^2$$

$$s.t.\ y_i(\boldsymbol{w}^T\boldsymbol{x}_i + w_0) \geq 1,\ i = 1,2,\ldots,m$$

- Lagrange function: Lagrange multiplier $\boldsymbol{a} = (a_1, a_2, \ldots, a_m)^T$:

$$L(\boldsymbol{w}, w_0, \boldsymbol{a}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} a_i(1 - y_i(\boldsymbol{w}^T x_i + w_0))$$

- **Primary problem →Duel problem (maxi-mini problem)**

$$\max_{\boldsymbol{a}} \min_{\boldsymbol{w},w_0} L(\boldsymbol{w}, w_0, \boldsymbol{a})$$

- To get the solution of SVM duel problem:

➢ Solve the minimum of $L(\boldsymbol{w}, b, \boldsymbol{a})$ on $\boldsymbol{w}, w_0$: $\min_{\boldsymbol{w},w_0} L(\boldsymbol{w}, w_0, \boldsymbol{a})$

➢ Solve the maximum of $\min_{\boldsymbol{w},w_0} L(\boldsymbol{w}, w_0, \boldsymbol{a})$ on $\boldsymbol{a}$: $\max_{\boldsymbol{a}} \min_{\boldsymbol{w},w_0} L(\boldsymbol{w}, w_0, \boldsymbol{a})$

SVM-Duel Problem

$$L(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \frac{1}{2}\|\boldsymbol{w}\|^2 + \sum_{i=1}^{m} \alpha_i (1 - y_i(\boldsymbol{w}^T \boldsymbol{x}_i + w_0))$$
$$= \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + \sum_{i=1}^{m} \alpha_i - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{w}^T \boldsymbol{x}_i - \sum_{i=1}^{m} \alpha_i y_i w_0$$

- (1) Solve $\min_{\boldsymbol{w}, w_0} L(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$, set the partial of $L(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$ on $\boldsymbol{w}$ and w_0 as 0

$$\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{w} - \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i = 0, \text{ then } \boldsymbol{w} = \sum_{i=1}^{m} \alpha_i y_i \boldsymbol{x}_i$$

$$\frac{\partial L}{\partial w_0} = -\sum_{i=1}^{m} \alpha_i y_i = 0, \text{ then } \sum_{i=1}^{m} \alpha_i y_i = 0$$

Then,

$$\min_{\boldsymbol{w}, w_0} L(\boldsymbol{w}, w_0, \boldsymbol{\alpha}) = \sum_{i=1}^{m} \alpha_i - \frac{1}{2}\sum_{i=1}^{m}\sum_{j=1}^{m} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i{}^T \boldsymbol{x}_j$$

SVM-Duel Problem

- (2) Solve the maximum of $\min_{\boldsymbol{w},w_0} L(\boldsymbol{w}, w_0, \boldsymbol{a})$ on $\boldsymbol{a}$

$$\max_{\boldsymbol{a}} \min_{\boldsymbol{w},w_0} L(\boldsymbol{w}, w_0, \boldsymbol{a}) = \max_{\boldsymbol{a}} \left(\sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i \, a_j y_i y_j \boldsymbol{x}_i{}^T \boldsymbol{x}_j \right)$$

$$s.t. \sum_{i=1}^{m} a_i y_i = 0, \; a_i \geq 0, \; i = 1,2,\ldots,m$$

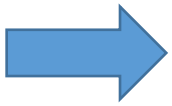
Substitute a_i after solving to get $\boldsymbol{w}$ and w_0, $w_0 = y_j - \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i^T \boldsymbol{x}_j$

$$g(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{x} + w_0 = \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + w_0$$

SVM-Duel Problem——Solution Sparsity

- Final SVM: $g(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0 = \sum_{i=1}^{m} a_i y_i \boldsymbol{x}_i^T \boldsymbol{x} + w_0$
- **SVM satisfies KKT(Karush-Kuhn-Tucker) conditions:**

$$\begin{cases} a_i \geq 0 \\ y_i g(\boldsymbol{x}_i) \geq 1 \\ a_i(y_i g(\boldsymbol{x}_i) - 1) = 0 \end{cases}$$



For any sample $(\boldsymbol{x}_i, y_i)$, there must exist $a_i = 0$ or $y_i g(\boldsymbol{x}_i) = 1$

- If $a_i = 0$, then $y_i g(\boldsymbol{x}_i) > 1$, $(\boldsymbol{x}_i, y_i)$ does not affect SVM $g(\boldsymbol{x})$.
- If $a_i > 0$, then $y_i g(\boldsymbol{x}_i) = 1$, $(\boldsymbol{x}_i, y_i)$ is on the boundary hyperplane H1 or H2, which is the support vector.

Solution Sparsity of SVM: After training, most of the training samples are not reserved. That is, the final SVM only concerns support vectors which is small amount.

For duel problem ,we only need to solve support vectors and corresponding multiplier a.

Example 2——Solve the duel problem of SVM

- Problem: Given a training set size of 3, in which (x_1, y_1) and (x_2, y_2) are positive samples, and (x_3, y_3) is negative sample.$x_1 = (3; 3), y_1 = +1, x_2 = (4; 3), y_2 = +1; x_3 = (1; 1), \quad y_3 = -1$. Solve the linearly separable SVM.

Example 2——Solve the duel problem of SVM

- SVM Primary Problem

$$\min_{\boldsymbol{w},b} \frac{1}{2}(w_1^2 + w_1^2)$$
$$s.t. \begin{cases} 3w_1 + 3w_2 + w_0 \geq 1 \\ 4w_1 + 3w_2 + w_0 \geq 1 \\ -w_1 - w_2 + w_0 \geq 1 \end{cases}$$

$$\max_{\boldsymbol{\alpha}} \min_{\boldsymbol{w},w_0} L(\boldsymbol{w}, w_0, \boldsymbol{\alpha})$$
$$= \max_{\boldsymbol{\alpha}} \left(\sum_{i=1}^{m} a_i - \frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} a_i\, a_j y_i y_j \boldsymbol{x}_i{}^T \boldsymbol{x}_j \right)$$
$$s.t. \sum_{i=1}^{m} a_i y_i = 0,\ a_i \geq 0,\ i = 1,2,\ldots,m$$

- Step1: transform to SVM duel problem

$$\max_{a_1,a_2,a_3} \left(\sum_{i=1}^{3} a_i - \frac{1}{2} \sum_{i=1}^{3} \sum_{j=1}^{3} a_i\, a_j y_i y_j \boldsymbol{x}_i{}^T \boldsymbol{x}_j \right)$$
$$s.t. \begin{cases} a_1 + a_2 - a_3 = 0 \\ a_1 \geq 0, a_2 \geq 0, a_3 \geq 0 \end{cases}$$

$$\min_{a_1,a_2,a_3} \left(\frac{1}{2}(18a_1^2 + 25a_2^2 + 2a_3^2 + 42a_1a_2 - 12a_1a_3 - 14a_2a_3) - a_1 - a_2 - a_3 \right)$$
$$s.t. \begin{cases} a_1 + a_2 - a_3 = 0 \\ a_1 \geq 0, a_2 \geq 0, a_3 \geq 0 \end{cases}$$

Example 2——Solve the duel problem of SVM

- Step 2: Substitute constraints $a_3 = a_1 + a_2$, then the objective function is:

$$s(a_1, a_2) = 4a_1^2 + \frac{13}{2}a_2^2 + 10a_1a_2 - 2a_1 - 2a_2$$
$$s.t.\ a_1 \geq 0, a_2 \geq 0, a_3 \geq 0$$

Solve the partial of $s(a_1, a_2)$ on a_1, a_2, and set as 0:

$$\begin{cases} \frac{\partial s}{\partial a_1} = 8a_1 + 10a_2 - 2 = 0 \\ \frac{\partial s}{\partial a_2} = 13a_2 + 10a_1 - 2 = 0 \end{cases}$$

then $\begin{cases} a_1 = \frac{3}{2} \\ a_2 = -1 < 0, \\ a_3 = \frac{1}{2} \end{cases}$

This violates constraints! We will find the minimum on boundary value of a_i.

Example 2——Solve the duel problem of SVM

- Step 3: Solve vector $\boldsymbol{w}$ with KKT conditions

(1) When $a_1 = 0$,

$s(0, a_2) = \frac{13}{2} a_2^2 - 2a_2$,

Set $\frac{\partial s}{\partial a_2} = 13a_2 - 2 = 0$, then $a_2 = \frac{2}{13}$, $s_{min} = -\frac{2}{13}$

(2) When $a_2 = 0$,

$s(a_1, 0) = 4a_1^2 - 2a_1$,

Set $\frac{\partial s}{\partial a_1} = 8a_1 - 2 = 0$, then $a_1 = \frac{1}{4}$, $s_{min} = -\frac{1}{4}$

Thus, when $a_1 = \frac{1}{4}, a_2 = 0, a_3 = \frac{1}{4}$, $s_{min} = -\frac{1}{4}$

And $\boldsymbol{w} = \sum_{i=1}^{3} a_i y_i \boldsymbol{x}_i = a_1 y_1 \boldsymbol{x}_1 + a_3 y_3 \boldsymbol{x}_3 = (\frac{1}{2}, \frac{1}{2})$

Example 2——Solve the duel problem of SVM

- Step 4: Solve w_0 with KKT conditions

Since $a_1 = \frac{1}{4} > 0$, the corresponding sample $\boldsymbol{x}_1$ is the support vector, then $y_1 g(\boldsymbol{x}_1) = 1$ and $w_0 = -2$.

- Hyperplane H (SVM)

$$\frac{1}{2}x^1 + \frac{1}{2}x^2 - 2 = 0$$

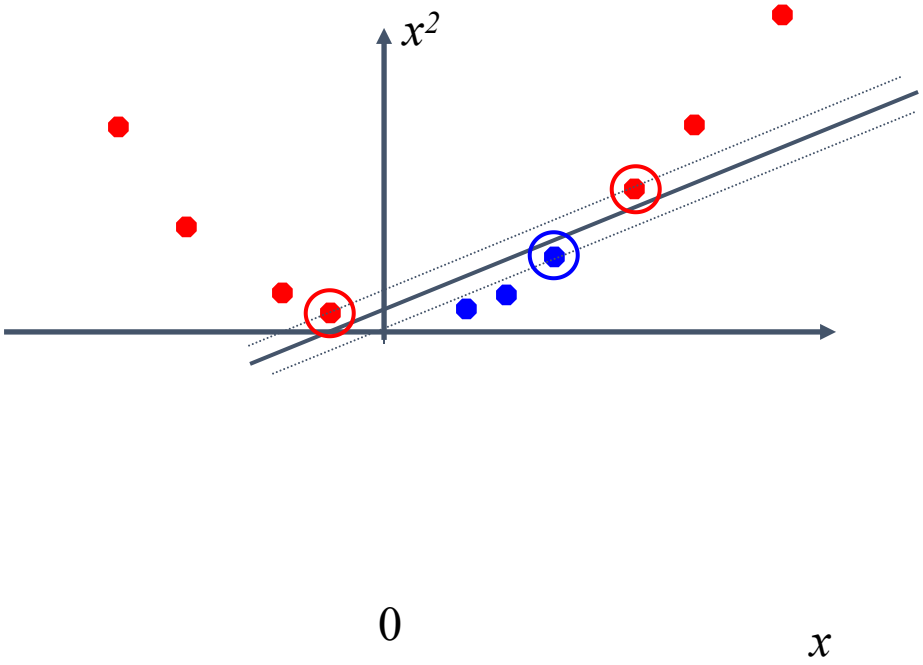
- For new sample, discriminant function is

$$g(\boldsymbol{x}) = \text{sign}\left(\frac{1}{2}x^1 + \frac{1}{2}x^2 - 2\right)$$

Classification

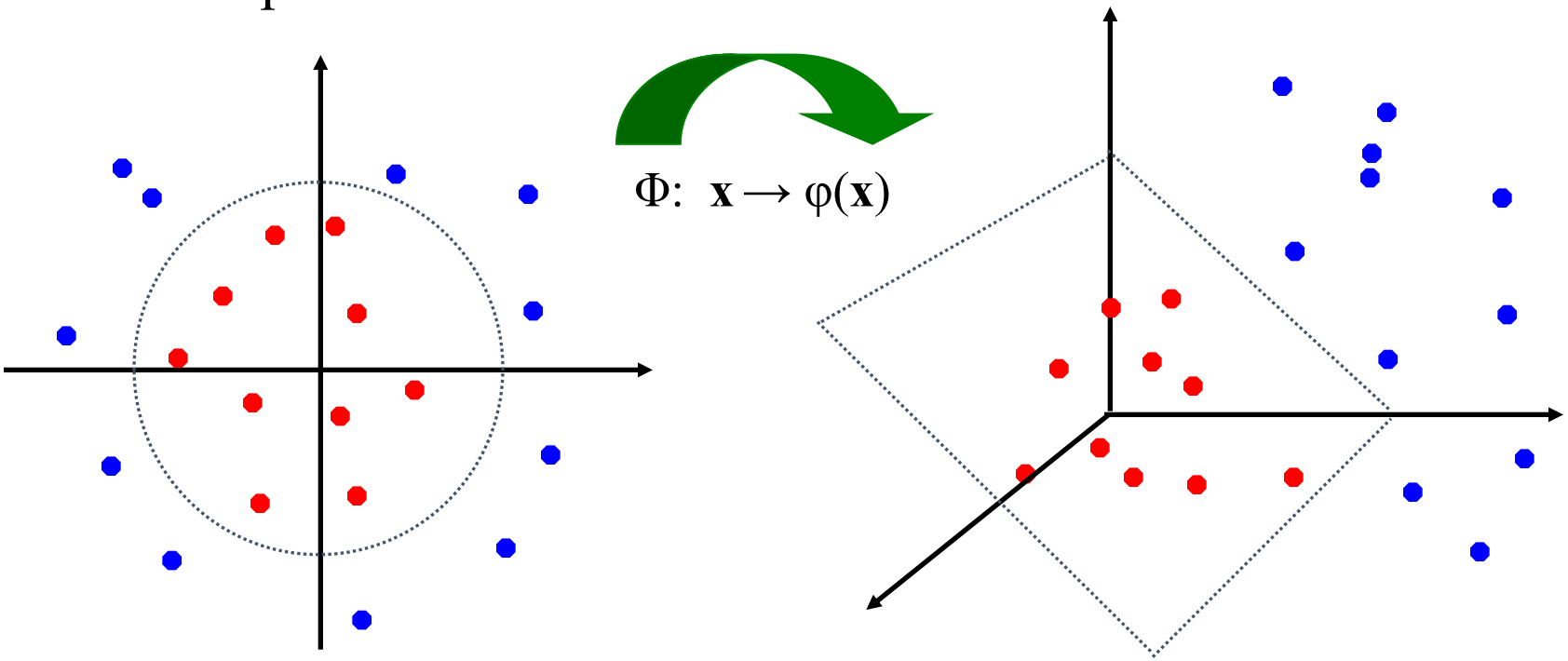
- NON-linear SVMs
  - what are we going to do if the dataset is just too hard?

  - How about… mapping data to a higher-dimensional space:



Classification

- NON-linear SVMs
  - General idea: the original input space can always be mapped to some higher-dimensional feature space where the training set is separable:



A *kernel function* is some function that corresponds to an inner product in some expanded feature space.

Classification

Weakness of SVM

- It is sensitive to noise
- It only considers two classes

- how to do multi-class classification with SVM?

Suppose there are m different classes,

1) OVA-SVM : (m SVMs in total)

- SVM_1 learns "Output==1" vs "Output != 1"
- SVM_2 learns "Output==2" vs "Output != 2"
- :
- SVM_m learns "Output==m" vs "Output != m"

2) OVO-SVM: (m(m-1)/2 SVMs in total)

- SVM_{12} learns "Output==1" vs "Output == 2"
- SVM_{13} learns "Output==1" vs "Output ==3"
- :
- $SVM_{m(m-1)}$ learns "Output==m" vs "Output == m-1"

Classification

- Other classifiers
  - Decision trees
  - Sparse representation classifier
  - Neural networks
  - ……

Conclusion

- Linear Regression
- Logistic Regression
- Classification
  - Distance-based algorithms
  - Linear classifiers
  - Other classifiers
- ……