Logistic Regression

Hongchang Gao

Spring 2024

Supervised Learning

- Supervised Learning Methods:

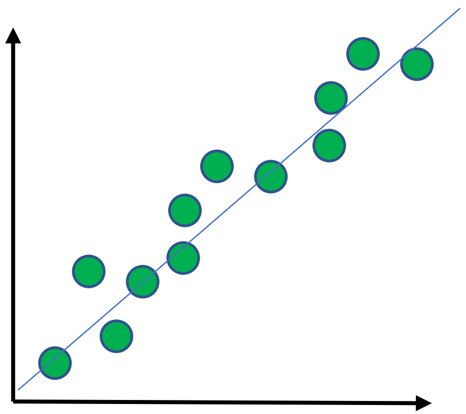
Given n samples: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

Learn a mapping function $x_i \xrightarrow{f(x)} y_i$

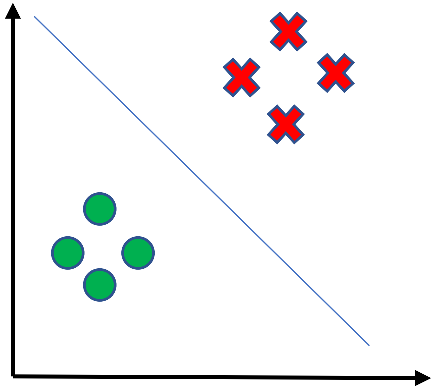
Y is continuous: regression

Y is discrete/categorical: classification

Regression

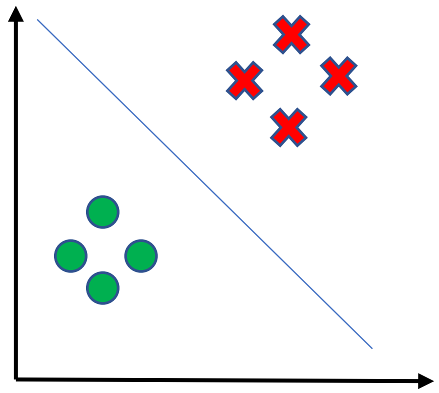


classification

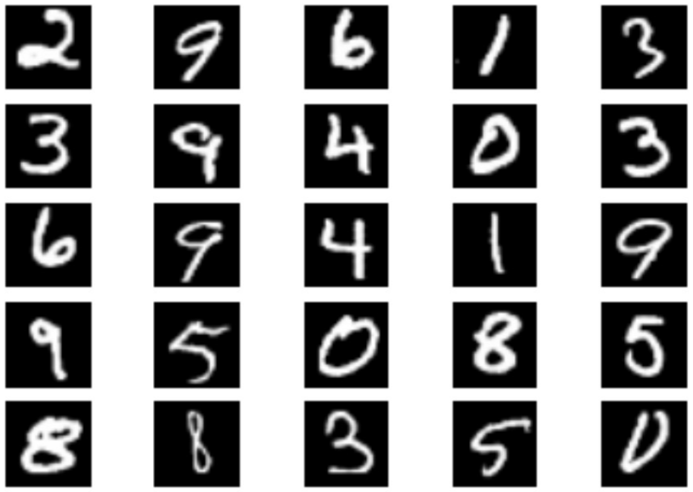


Classification

- Binary Classification
  - Only two classes
  - E.g. fraud detection



- Multi-class Classification
  - Multiple classes



Binary Classification

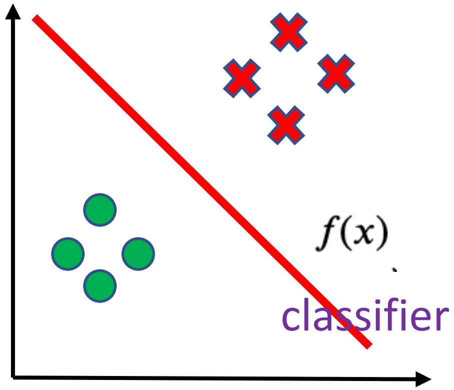
- Binary classification
  - Learn a classifier to separate positive samples from negative samples
    - Positive sample: its label is 1
    - Negative sample: its label is 0

Given n samples: $\{(x_1, y_1), (x_2, y_2), \cdots, (x_n, y_n)\}$

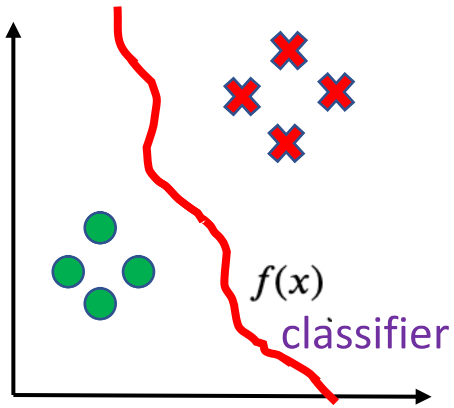
Learn a mapping function $x_i \xrightarrow{f(x)} \begin{cases} 0 \\ 1 \end{cases}$

Binary Classification

- Linear classifier
- Non-linear classifier



Linear



Non-linear

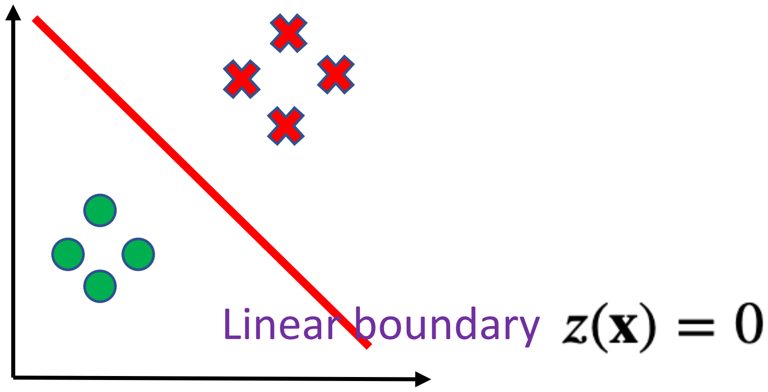
Linear Classifier

- Linear classifier

$$z(\mathbf{x}) = \mathbf{w}^T\mathbf{x} = w_0 + w_1x_1 + w_2x_2 + \cdots + w_dx_d$$

  - $z(\mathbf{x}) = 0$ specifies a linear boundary, separating the space into two half-spaces
  - A reasonable decision rule

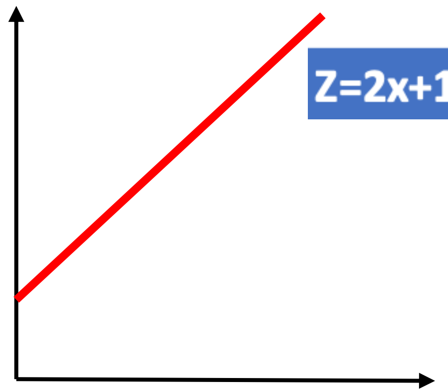
$$\hat{y} = \begin{cases} 1, & z(\mathbf{x}) > 0 \\ 0, & z(\mathbf{x}) < 0 \end{cases}$$



Linear Classifier

- Example

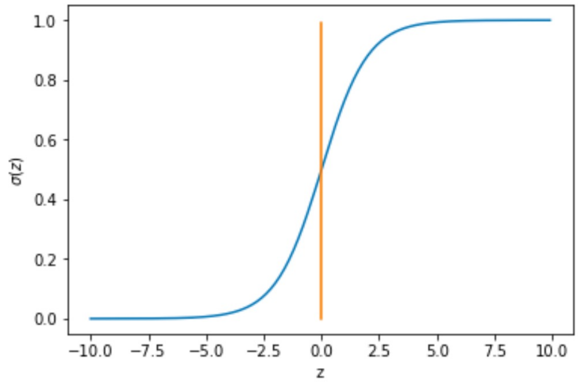
x	Z=2x+1	Predicted label
1	3	1
2	5	1
-2	-3	0
-1	-1	0



Logistic Regression

- How to let the classifier:
  - output 1 for positive samples,
  - output 0 for negative samples?
- Sigmoid function
  - $z > 0$, the function value is close to 1 => the predicted label is 1
  - $z < 0$, the function value is close to 0 => the predicted label is 0
  - $z = 0$, the function value is 0.5

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



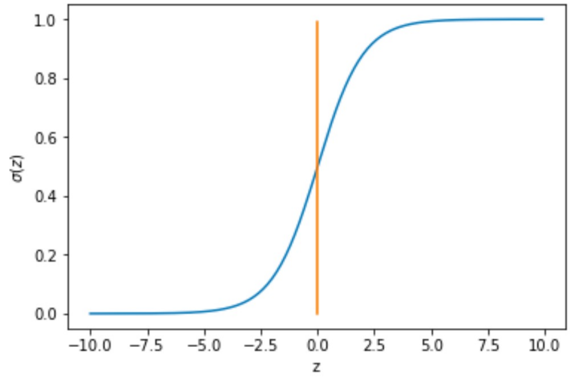
Logistic Regression

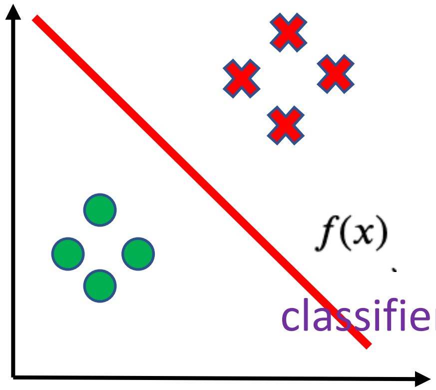
- How to let the classifier output 1 or 0 for positive or negative samples?

$$f(\mathbf{x}_i) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}_i)}$$

$$z_i = \mathbf{w}^T\mathbf{x}_i = w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d}$$ Linear transformation

$$\sigma(z_i) = \frac{1}{1+\exp(-z_i)}$$ Sigmoid function





Logistic Regression

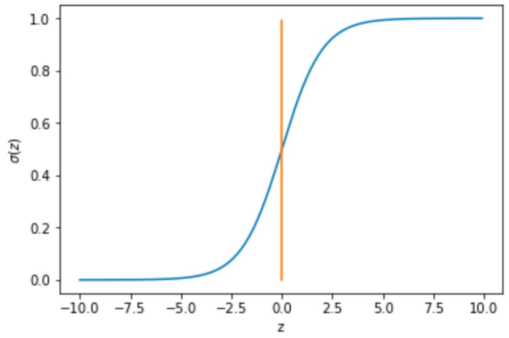
- Example

$$f(\mathbf{x}_i) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x}_i)}$$

x	Z=2x+1	F(x)
1	3	0.95
2	5	0.99
-2	-3	0.04
-1	-1	0.26

Probabilistic Interpretation

- The output of logistic regression models the class probability
  - The probability that a sample belongs to class 1



$$p(y=1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{\exp(\mathbf{w}^T\mathbf{x})}{1+\exp(\mathbf{w}^T\mathbf{x})}$$

  - The probability that a sample belongs to class 0

$$p(y=0|\mathbf{x}) = 1 - p(y=1|\mathbf{x}) = 1 - \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{\exp(-\mathbf{w}^T\mathbf{x})}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x})}$$

Probabilistic Interpretation

- Example

$$p(y=1|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x}) = \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{\exp(\mathbf{w}^T\mathbf{x})}{1+\exp(\mathbf{w}^T\mathbf{x})}$$

$$p(y=0|\mathbf{x}) = 1 - p(y=1|\mathbf{x}) = 1 - \frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{\exp(-\mathbf{w}^T\mathbf{x})}{1+\exp(-\mathbf{w}^T\mathbf{x})} = \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x})}$$

x	Z=2x+1	F(x)	P(y=1|x)	P(y=0|x)
1	3	0.95	0.95	0.05
2	5	0.99	0.99	0.01
-2	-3	0.04	0.04	0.96
-1	-1	0.26	0.26	0.74

$$\frac{1}{1+\exp(-\mathbf{w}^T\mathbf{x})}$$

How to optimize it?

- Likelihood function
  - Measure how well a model fits the data
  - The higher the likelihood function, the better the model fits the data

$$L(\theta) = p(x_1; \theta)p(x_2; \theta) \cdots p(x_n; \theta)$$

$$p(y_i|\mathbf{x}_i) = \begin{cases} p(1|\mathbf{x}_i), & y_i = 1 \\ 1 - p(1|\mathbf{x}_i), & y_i = 0 \end{cases}$$

$$\Downarrow$$

$$p(y_i|\mathbf{x}_i) = p(1|\mathbf{x}_i)^{y_i}(1 - p(1|\mathbf{x}_i))^{(1-y_i)}$$

$$\Downarrow$$

$$\max_{\mathbf{w}} \textstyle\prod_{i=1}^{n} p(1|\mathbf{x}_i)^{y_i}(1 - p(1|\mathbf{x}_i))^{(1-y_i)}$$

- Maximize the likelihood function

$$\max_{\mathbf{w}} \textstyle\prod_{i=1}^{n} p(y_i|\mathbf{x}_i)$$

How to optimize it?

- Example

$$\prod_{i=1}^{n} p(1|\mathbf{x}_i)^{y_i}(1 - p(1|\mathbf{x}_i))^{(1-y_i)} \xrightarrow{\text{Z is known}} 0.95*0.99*0.96*0.74$$

y	x	Z=2x+1	F(x)	P(y=1|x)	P(y=0|x)
1	1	3	0.95	0.95	0.05
1	2	5	0.99	0.99	0.01
0	-2	-3	0.04	0.04	0.96
0	-1	-1	0.26	0.26	0.74

$$\prod_{i=1}^{n} p(1|\mathbf{x}_i)^{y_i}(1 - p(1|\mathbf{x}_i))^{(1-y_i)} \xrightarrow{\text{Z is unknown}} ?$$

How to optimize it?

- Minimize negative log-likelihood function

$$\min_{\mathbf{w}} -\log \prod_{i=1}^{n} p_i^{y_i}(1-p_i)^{(1-y_i)} \Rightarrow \min_{\mathbf{w}} -\sum_{i=1}^{n}\{y_i \log p_i + (1-y_i)\log(1-p_i)\}$$

Cross-entropy loss function

$$\sum_{i=1}^{n}\{y_i \log p_i + (1-y_i)\log(1-p_i)\}$$

$$= \sum_{i=1}^{n}\{y_i \log \frac{\exp(\mathbf{w}^T\mathbf{x}_i)}{1+\exp(\mathbf{w}^T\mathbf{x}_i)} + (1-y_i)\log \frac{1}{1+\exp(\mathbf{w}^T\mathbf{x}_i)}\}$$

$$= \sum_{i=1}^{n}\{y_i\mathbf{w}^T\mathbf{x}_i - y_i\log(1+\exp(\mathbf{w}^T\mathbf{x}_i)) - (1-y_i)\log(1+\exp(\mathbf{w}^T\mathbf{x}_i))\}$$

$$= \sum_{i=1}^{n}\{y_i\mathbf{w}^T\mathbf{x}_i - \log(1+\exp(\mathbf{w}^T\mathbf{x}_i))\}$$

How to optimize it?

- Use the gradient descent method to learn the model parameter w

$$\boxed{\min_{\mathbf{w}} \sum_{i=1}^{n} \{\log(1 + \exp(\mathbf{w}^T \mathbf{x}_i)) - y_i \mathbf{w}^T \mathbf{x}_i\}}$$

Loss function

  - Gradient

$$\frac{\partial L}{\partial \mathbf{w}} = \sum_{i=1}^{n} \mathbf{x}_i (y_i - \frac{\mathbf{w}^T \mathbf{x}_i}{1+\exp(\mathbf{w}^T \mathbf{x}_i)})$$

  - Gradient descent

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \sum_{i=1}^{n} \mathbf{x}_i (y_i - \frac{\mathbf{w}_k^T \mathbf{x}_i}{1+\exp(\mathbf{w}_k^T \mathbf{x}_i)})$$

Overfitting

- To avoid the overfitting issue, we add a regularization term

$$\min_{\mathbf{w}} \sum_{i=1}^{n} \{\log(1 + \exp(\mathbf{w}^T \mathbf{x}_i)) - y_i \mathbf{w}^T \mathbf{x}_i\} + \lambda \|\mathbf{w}\|_2^2$$

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \eta \textstyle\sum_{i=1}^{n} \mathbf{x}_i \left(y_i - \frac{\mathbf{w}_k^T \mathbf{x}_i}{1+\exp(\mathbf{w}_k^T \mathbf{x}_i)}\right) - 2\eta\lambda \mathbf{w}_k$$