IOWA STATE UNIVERSITY
Department of Computer Science

COM S 573: Machine Learning

Lecture 6: Logistic Regression

Linear Models

- Linear regression: predict a scalar
  - House price
  - Weight of a planet
- Linear perceptron: classifier
  - Predict an animal is a dog or not
  - Predict an image contains a square or not
- Logistic regression: classifier based on a probability
  - Predict how likely a team win
  - Predict how likely tomorrow is sunny

Linear Models

- Linear regression: predict a scalar

$$\hat{y}_i = \boldsymbol{w}^T \boldsymbol{x}_i$$

- Linear perceptron: predict $\{1, -1\}$

$$\hat{y}_i = sign(\boldsymbol{w}^T \boldsymbol{x}_i)$$

- Logistic regression: predict a probability

$$\hat{y}_i = sigmoid(\boldsymbol{w}^T \boldsymbol{x}_i)$$

Logistic Regression: Representation

- Logistic regression is is used to model the probability of a certain class or event.

$$\hat{y}_i = sigmoid(\boldsymbol{w}^T \boldsymbol{x}_i)$$

Linear regression

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

Logistic Regression: sigmoid

- Why sigmoid function?

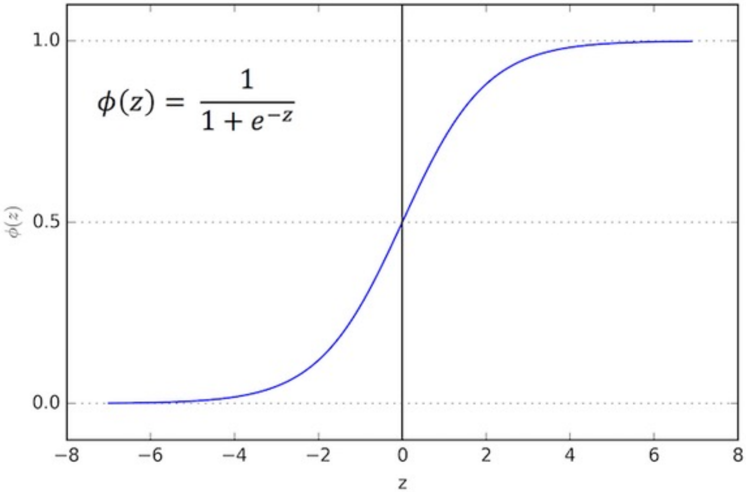
$$sigmoid(z) = \frac{1}{1+e^{-z}} = \frac{e^z}{e^z+1}$$

- Bounded between 0 and 1
  - Probability
- **Monotonically** increasing

$$x_i < x_j \rightarrow f(x_i) < f(x_j)$$

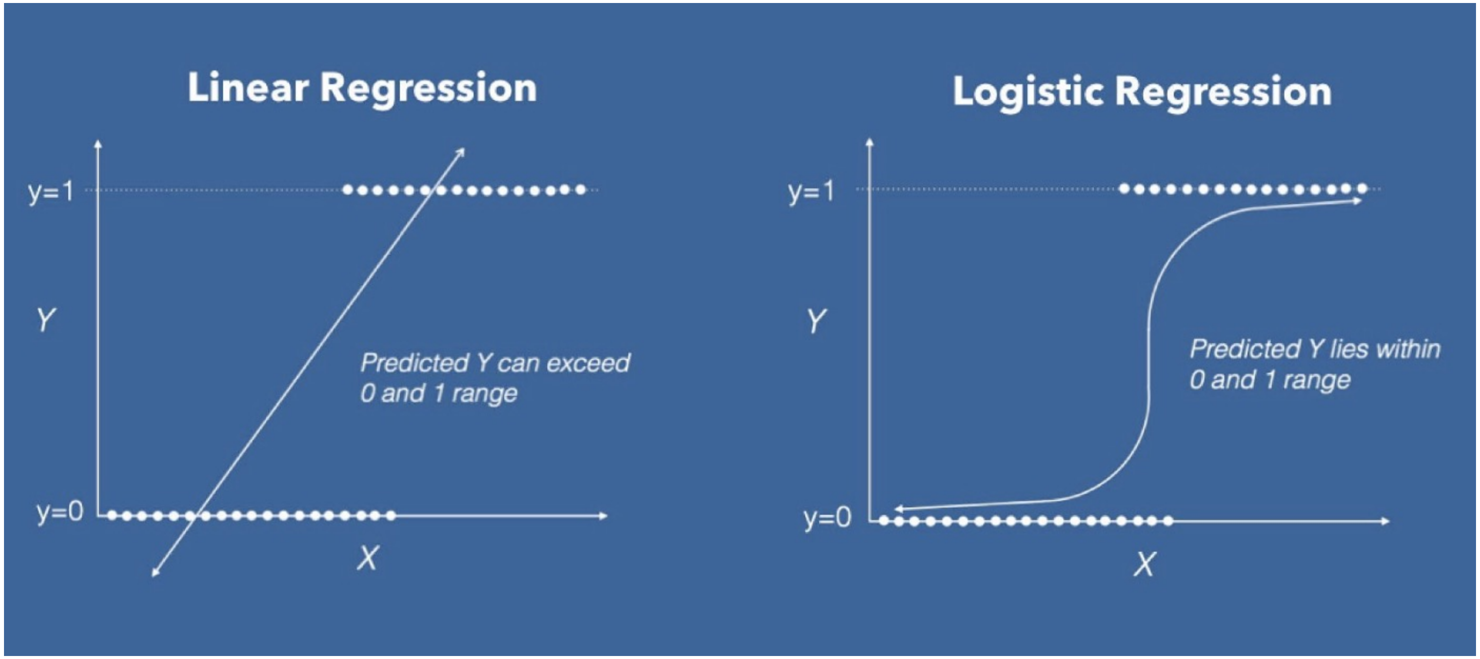
- Nice computational properties

$$f'(x_i) = f(x_i)(1 - f(x_i))$$



Logistic Regression: Representation

- Logistic regression is is used to model the probability of a certain class or event.



Logistic Regression: Representation

- For each training sample $< \boldsymbol{x}_i, y_i >$
- $\hat{y}_i = sigmoid(w_0 + w_1 x_{i,1} + w_2 x_{i,2} + \cdots + w_d x_{i,d})$
- Suppose $\boldsymbol{w} = [w_0, w_1, \cdots, w_d]^T$
- $\hat{y}_i = sigmoid(\boldsymbol{w}\boldsymbol{x}_i^T)$

Logistic Regression: Evaluation

Data likelihood for 1 training sample

$$p(y_n|\mathbf{x_n}, \mathbf{w}) = \left\{ \begin{array}{ll} \sigma(\mathbf{w}^T\mathbf{x_n}), & y_n = 1 \\ 1 - \sigma(\mathbf{w}^T\mathbf{x_n}), & y_n = 0 \end{array} \right\} = \left[\sigma(\mathbf{w}^T\mathbf{x_n})\right]^{y_n} \left[1 - \sigma(\mathbf{w}^T\mathbf{x_n})\right]^{1-y_n}$$

Data likelihood for all training data

$$L(\mathcal{D}|\mathbf{w}) = \prod_{n=1}^{N} p(y_n|\mathbf{x_n}, \mathbf{w}) = \prod_{n=1}^{N} \left[\sigma(\mathbf{w}^T\mathbf{x_n})\right]^{y_n} \left[1 - \sigma(\mathbf{w}^T\mathbf{x_n})\right]^{1-y_n}$$

Cross-entropy error (negative log-likelihood)

$$\begin{aligned} \mathcal{E}(\mathbf{w}) &= -\log L(\mathcal{D}|\mathbf{w}) \\ &= -\sum_{n=1}^{N} \left\{ y_n \log \left[\sigma(\mathbf{w}^T\mathbf{x_n})\right] + (1 - y_n) \log \left[1 - \sigma(\mathbf{w}^T\mathbf{x_n})\right] \right\} \end{aligned}$$

How to find the optimal w?

Logistic Regression: Optimization

Cross-entropy error (negative log-likelihood)

$$\mathcal{E}(\mathbf{w}) = -\sum_{n=1}^{N} \left\{ y_n \log \left[\sigma(\mathbf{w}^T \mathbf{x_n})\right] + (1 - y_n) \log \left[1 - \sigma(\mathbf{w}^T \mathbf{x_n})\right] \right\}$$

How to find the weights $\mathbf{w}$ of the logistic regression?

We can maximize data likelihood or minimize cross-entropy error

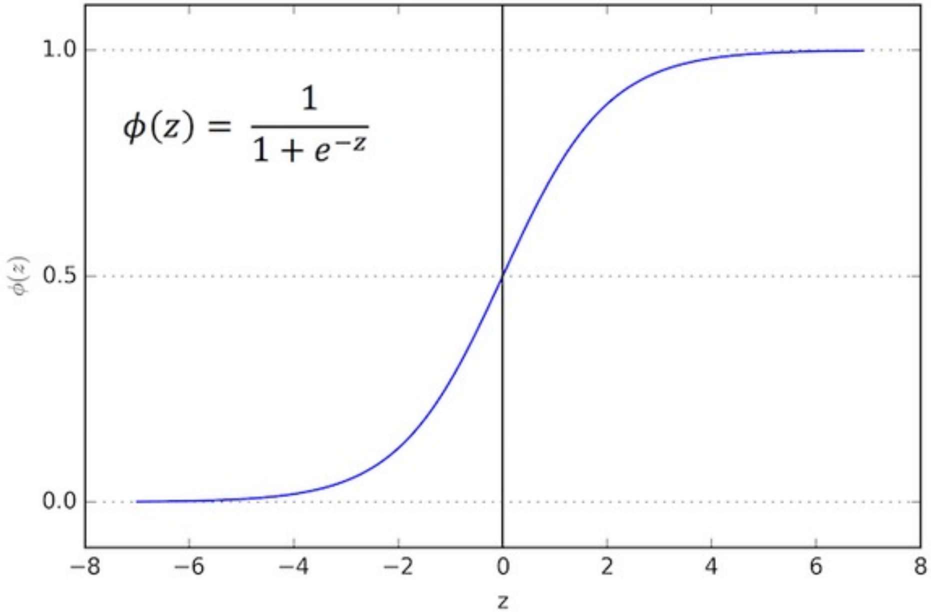
$$\mathbf{w}^* = \min_{\mathbf{w}} \mathcal{E}(\mathbf{w})$$

No closed-form solution $\rightarrow$ approximate methods, e.g. Gradient Descent.

$$\mathbf{w} := \mathbf{w} - \alpha(k) \cdot \nabla \mathcal{E}(\mathbf{w}), \quad \frac{\vartheta \mathcal{E}(\mathbf{w})}{\vartheta w_d} = \sum_{n=1}^{N} \underbrace{\left(\sigma(\mathbf{w}^T \mathbf{x_n}) - y_n\right)}_{\text{error}} x_{nd}$$

Logistic Regression: Question

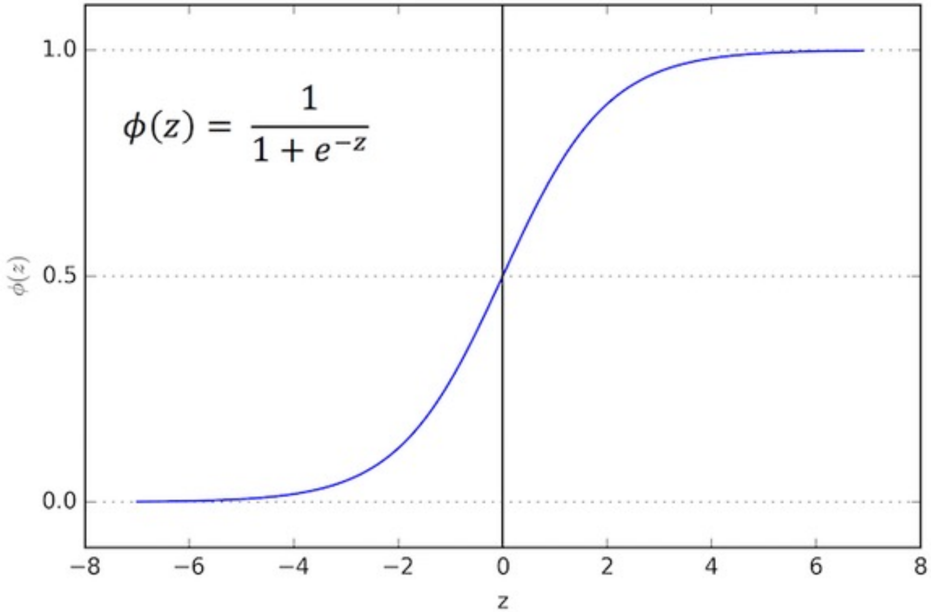
- Logistic regression is a linear classifier?



$$y_i = \begin{cases} 1 & \text{if } sigmoid(\boldsymbol{w}^T \boldsymbol{x}_i) >= 0.5 \\ -1 & \text{if } sigmoid(\boldsymbol{w}^T \boldsymbol{x}_i) < 0.5 \end{cases}$$

Logistic Regression: Question

- Logistic regression is a linear classifier?



$$y_i = \begin{cases} 1 & \text{if } sigmoid(\boldsymbol{w}^T \boldsymbol{x}_i) >= 0.5 \\ -1 & \text{if } sigmoid(\boldsymbol{w}^T \boldsymbol{x}_i) < 0.5 \end{cases}$$

- Yes, it is still a linear model.

Logistic Regression: Question

threshold

$$\hat{y}_i = sigmoid(\boldsymbol{w}^T \boldsymbol{x}_i) = \frac{1}{2}$$

$$\rightarrow \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}} = \frac{1}{2}$$

$$\rightarrow e^{-\boldsymbol{w}^T \boldsymbol{x}_i} = 1$$

$$\rightarrow \boldsymbol{w}^T \boldsymbol{x}_i = 0$$

Logistic regression is a linear classifier.

Logistic Regression: Question

- Logistic regression is a linear classifier.
- If change sigmoid to another function, still linear?

$$\hat{y}_i = sigmoid(\boldsymbol{w}^T \boldsymbol{x}_i)$$

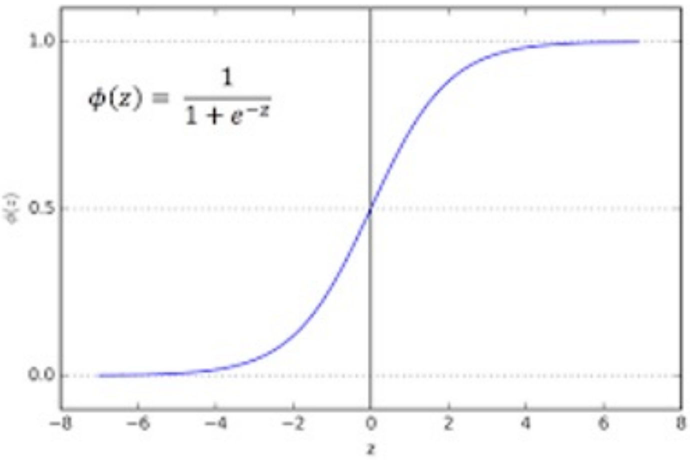
$$\Downarrow$$

$$\hat{y}_i = function(\boldsymbol{w}^T \boldsymbol{x}_i)$$

Logistic Regression: Question

- Logistic regression is a linear classifier, because sigmoid is a Monotonic function.

$$x_i < x_j \rightarrow f(x_i) < f(x_j)$$



Multi-Class Classification

- We are mostly dealing with binary classification
- How about multi-class classification?

Multi-Class Classification

- We are mostly dealing with binary classification
- How about multi-class classification?
- **Softmax**

$$y_1 = \boldsymbol{w_1}^T \boldsymbol{x}$$

$$y_2 = \boldsymbol{w_2}^T \boldsymbol{x}$$

$$\cdots$$

$$y_c = \boldsymbol{w_c}^T \boldsymbol{x}$$

$$z_j = \frac{e^{y_j}}{\sum_{i=1}^{c} e^{y_i}}$$

- What is the range of z_j ?

Multi-Class Classification

- We are mostly dealing with binary classification
- How about multi-class classification?
- **Softmax**

$$y_1 = \boldsymbol{w_1}^T \boldsymbol{x}$$

$$y_2 = \boldsymbol{w_2}^T \boldsymbol{x}$$

$$\cdots$$

$$y_c = \boldsymbol{w_c}^T \boldsymbol{x}$$

$$z_j = \frac{e^{y_j}}{\sum_{i=1}^{c} e^{y_i}}$$

- Choose the class with maximum value

Multi-Class Classification

- We are mostly dealing with binary classification
- How about multi-class classification?
- **Softmax**

$$y_1 = \boldsymbol{w_1}^T \boldsymbol{x}$$

$$y_2 = \boldsymbol{w_2}^T \boldsymbol{x}$$

$$\ldots$$

$$y_c = \boldsymbol{w_c}^T \boldsymbol{x}$$

$$z_j = \frac{e^{\boldsymbol{w}_j^T \boldsymbol{x}_j}}{\sum_{i=1}^{c} e^{\boldsymbol{w}_i^T \boldsymbol{x}_i}}$$

- What is the relationship between softmax and sigmoid?

$$\frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}}$$

Softmax VS. Sigmoid

- Relationship between softmax and sigmoid?
- Softmax can reduce to sigmoid when c = 2

$$z_j = \frac{e^{\boldsymbol{w}_j^T \boldsymbol{x}_j}}{\sum_{i=1}^{c} e^{\boldsymbol{w}_i^T \boldsymbol{x}_i}} \qquad \frac{1}{1 + e^{-\boldsymbol{w}^T \boldsymbol{x}_i}}$$

$\boxed{\text{HW1}}$