Applied Machine Learning

Linear Models (3)

Computer Science, Fall 2022

Instructor: Xuhong Zhang

Logistic Regression

- Our previous classification models are more focused on just predicting the class, now we are interested in giving the probability of the instance being that class, e.g. $p(y|x)$
- Some basic probability rules:

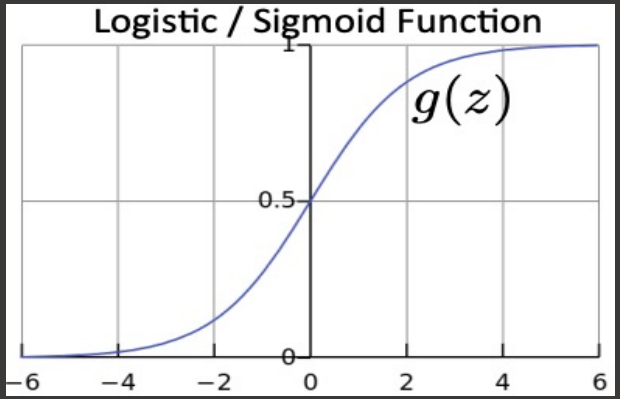
$$0 \leq p(event) \leq 1$$
$$p(event) + p(\neg event) = 1$$

Logistic Function

- Takes a probabilistic approach to learn discriminative functions (i.e., a classifier)

- We assume a function $h_w(\mathrm{x})$ should give $p(y = 1|x; w)$:
  - Want $0 \leq h_w(x) \leq 1$

- Logistic regression model:



$$h_w(x) = g(w^T x)$$

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

$$h_w(\mathrm{x}) = \frac{1}{1 + e^{-w^T x}}$$

Logistic Regression

- **Logistic Regression** is an approach to learn functions of the form $f: X \longrightarrow Y$, or $P(y|\mathbf{x})$ in the case where Y is discrete-valued, and $\mathrm{x} = \langle x_1, \ldots, x_D \rangle$ is any vector containing discrete or continuous variables.
- When Y is Boolean, then the model assumed by Logistic Regression is

$$h_w(\mathbf{x}) = P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^{D} w_i x_i)}$$

and

$$P(y = 0|\mathbf{x}) = \frac{\exp(w_0 + \sum_{i=1}^{D} w_i x_i)}{1 + \exp(w_0 + \sum_{i=1}^{D} w_i x_i)}$$

Interpretation of Hypothesis Output

- $h_w(x) = \text{estimated } p(y = 1|x; w)$
- Example: Cancer diagnosis from tumor size

$$x = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 1 \\ \text{tumor size} \\ \text{patient age} \end{bmatrix}$$

Then we have

$$h_w(x) = 0.7$$

Tell patient that 70% chance of tumor being malignant

$P(y = 0|x; w) = 1 - p(y = 1|x; w)$, so the tumor being benign is 30%

Another Interpretation

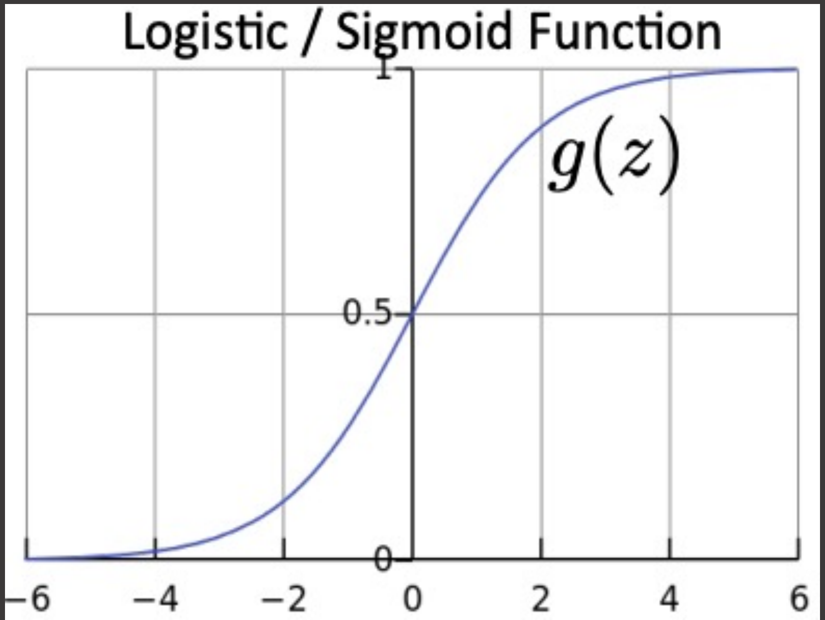
- Let's take the log of the odds of $y = 1$
  - The odds in favor of an event is the quantity $p/(1-p)$, where p is the probability of the event
- So

$$\log \boxed{\frac{p(y=1|\mathbf{x};w)}{p(y=0|\mathbf{x};w)}} = w_0 + w_1 x_1 + \cdots + w_D x_D$$

odds of $y = 1$

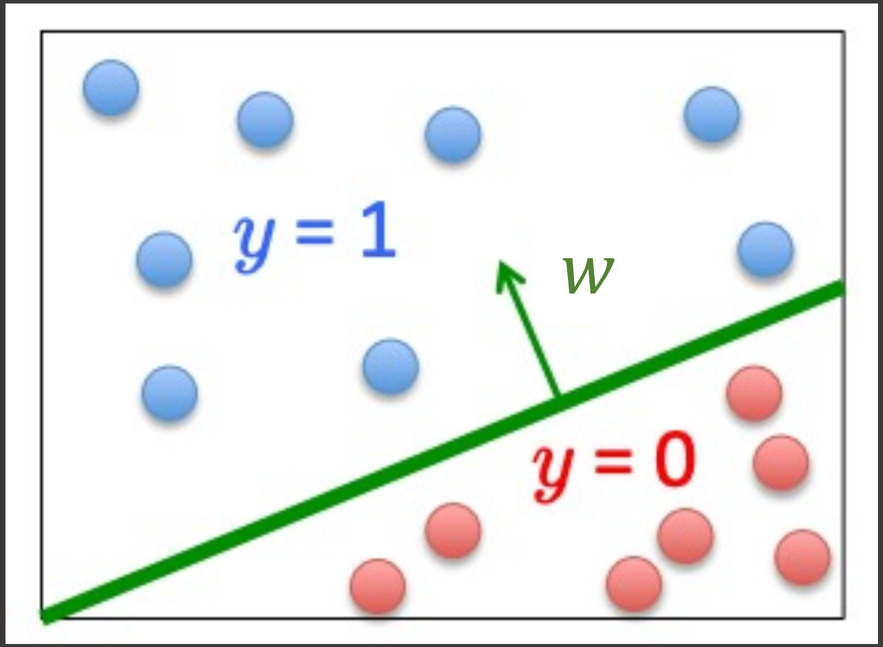
- In other words, logistic regression assumes that the log odds is a linear function of x

Logistic Regression



$$h_w(x) = g(w^T x)$$
$$g(z) = \frac{1}{1 + e^{-z}}$$

Predict $y = 1$ if $h_\theta(x) \geq 0.5$
Otherwise $y = 0$



$w^T x$ should be large negative for negative instances

$w^T x$ should be large positive values for positive instances

Logistic Regression

- Given $\{(x^{(1)}, y^{(1)}), (x^{(2)}, y^{(2)}), \dots, (x^{(n)}, y^{(n)})\}$ where $x^{(i)} \in \mathbb{R}^d, y^{(i)} \in \{0,1\}$

- Model: $h_w(x) = g(w^T x)$

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

$$w = \begin{bmatrix} w_0 \\ w_1 \\ \vdots \\ w_D \end{bmatrix} \qquad x^T = \begin{bmatrix} 1 & x_1 & \cdots & x_D \end{bmatrix}$$

Logistic Regression Objective Function

- We can't just use squared loss as in linear regression

$$J(\theta) = \frac{1}{2n}\sum_{i=1}^{n}(\varphi\left(x^{(i)}\right) - y^{(i)})^2$$

- If we use the logistic regression model, it will result in a non-convex optimization

- Deriving the cost function via ***Maximum Likelihood Estimation***

Maximum Likelihood Estimation

- Likelihood of data is given by: $l(w) = \prod_{i=1}^{n} p(y^i|\mathbf{x}^i; w)$

- We are looking for the w that maximizes the likelihood

$$w_{\mathrm{MLE}} = \underset{w}{\operatorname{argmax}}\, l(w) = \underset{w}{\operatorname{argmax}} \prod_{i=1}^{n} p(y^i|\mathbf{x}^i; w)$$

- Take the log

$$\begin{aligned} w_{\mathrm{MLE}} &= \underset{w}{\operatorname{argmax}} \log \prod_{i=1}^{n} p(y^i|\mathbf{x}^i; w) \\ &= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} \log p(y^i|\mathbf{x}^i; w) \end{aligned}$$

Maximum Likelihood Estimation

- Expand as follows

$$w_{\text{MLE}} = \operatorname{argmax} \sum_{i=1}^{n} \log p(y^i | \mathbf{x}^i; w)$$

$$= \operatorname{argmax} \sum_{i=1}^{n} [y^i \log p(y^i = 1 | \mathbf{x}^i; w) + (1 - y^i) \log(1 - p\big(y^i = 1 \big| \mathbf{x}^i; w\big))]$$

- Substitute in model, and take negative to yield

Logistic regression objective

$$\min_w J(w)$$

$$J(w) = -\sum_{i=1}^{n} [y^i \log \varphi_w(x^i) + (1 - y^i) \log(1 - \varphi_w(x^i))]$$

Intuition about the objective

$$J(w) = -\sum_{i=1}^{n} [y^i \log \varphi_w(x^i) + (1 - y^i) \log(1 - \varphi_w(x^i))]$$

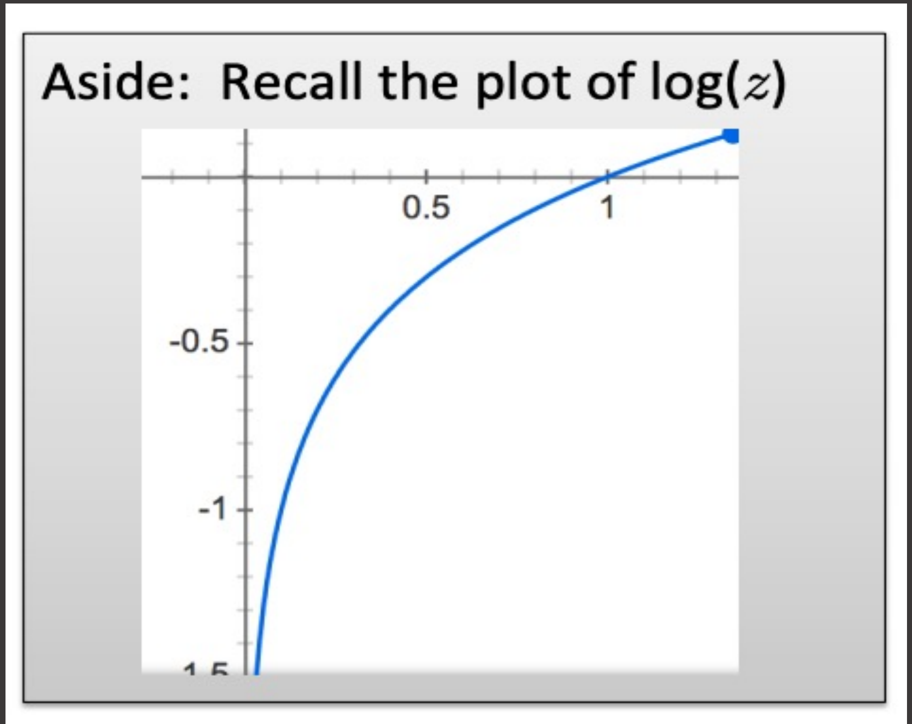
- Cost of a single instance:

$$\text{cost}\big(y(x^i), y^i\big) = \begin{cases} -\log\big(y(x^i)\big). & if\ y^i = 1 \\ -\log\big(1 - y(x^i)\big). & if\ y^i = 0 \end{cases}$$

  - Can re-write objective function as: $J(w) = \sum_{i=1}^{n} \text{cost}\big(y(x^i), y^i\big)$
  - Compare to linear regression: $J(w) = \frac{1}{N}\sum_{i=1}^{N}(y^i - y(x^i))^2$

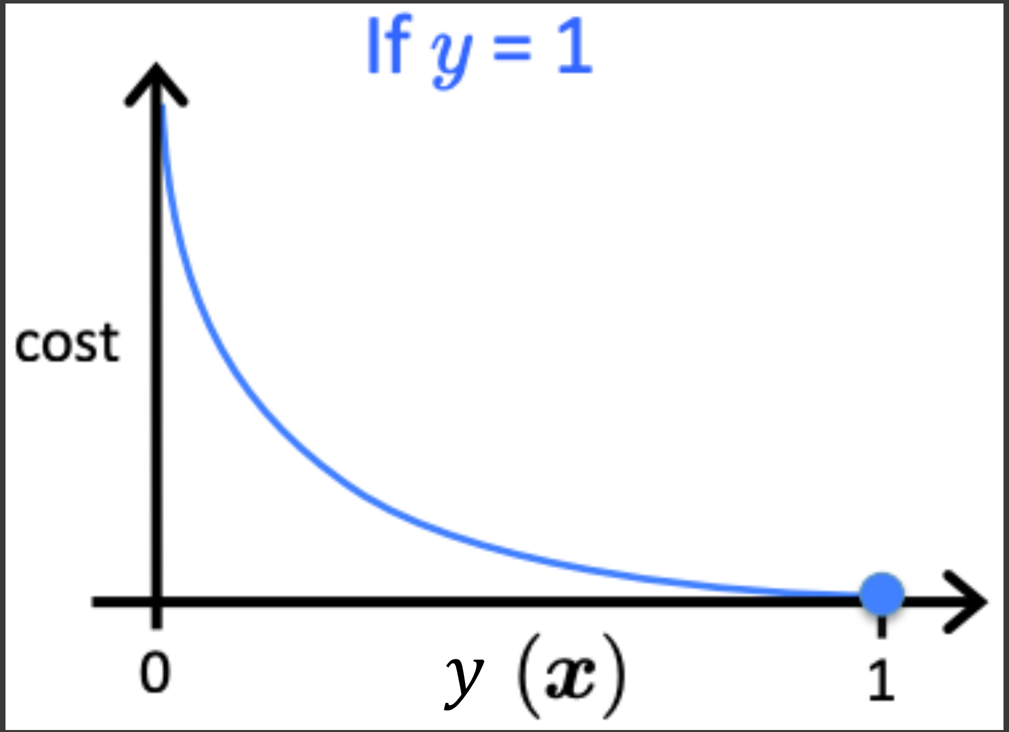
Intuition about the objective

- $\text{cost}\left(y(x^i), y^i\right) = \begin{cases} -\log\left(y(x^i)\right). & if\ y^i = 1 \\ -\log\left(1 - y(x^i)\right). & if\ y^i = 0 \end{cases}$



Intuition about the objective

- $\text{cost}\left(y(x^i), y^i\right) = \begin{cases} \boxed{-\log\left(y(x^i)\right). \quad if\ y^i = 1} \\ -\log\left(1 - y(x^i)\right). \quad if\ y^i = 0 \end{cases}$

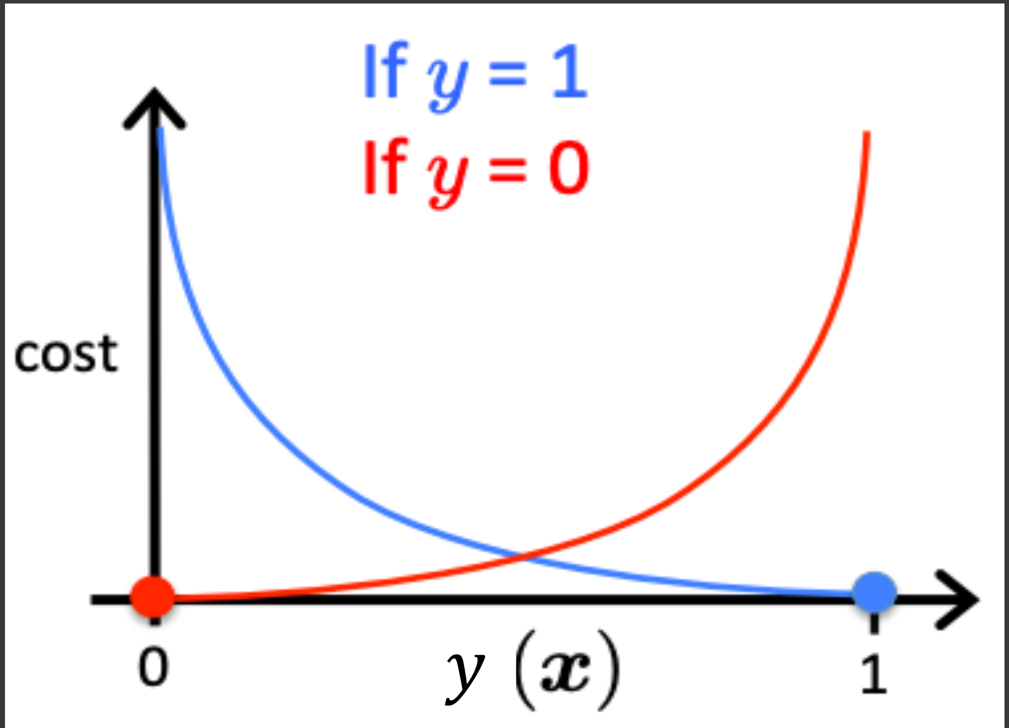


If $y = 1$

- Cost = 0 if prediction is correct
- As $y_w(x) \to 1, \text{cost} \to 0$
- Captures intuition that larger mistake should get larger penalties
  - ✓ e.g., predict $y_w(x^i) = 0,$ but $y^i = 1$

Intuition about the objective

- $\text{cost}\left(y(x^i), y^i\right) = \begin{cases} -\log\left(y(x^i)\right). & if\ y^i = 1 \\ \boxed{-\log\left(1 - y(x^i)\right). \quad if\ y^i = 0} \end{cases}$



If $y = 0$

- Cost = 0 if prediction is correct
- As $1 - y_w(x) \rightarrow 1, \text{cost} \rightarrow 0$
- Captures intuition that larger mistake should get larger penalties
  - ✓ e.g., predict $y_w(x^i) = 0$, but $y^i = 1$

Logistic Regression

- To classify any given x, we generally want to assign the value y_k that maximizes $P(y = y_k|\mathbf{x})$.
  - We assign the label $\mathrm{y} = 0$ if the following condition holds:

$$1 < \frac{P(y = 0|\mathbf{x})}{P(y = 1|\mathbf{x})}$$

  - Substituting the equations from previous slide, we have

$$1 < \exp(w_0 + \sum_{i=1}^{D} w_i x_D)$$

Logistic Regression

- Taking the natural log of both sides we then have a linear classification rule that assigns label $Y = 0$ if X:

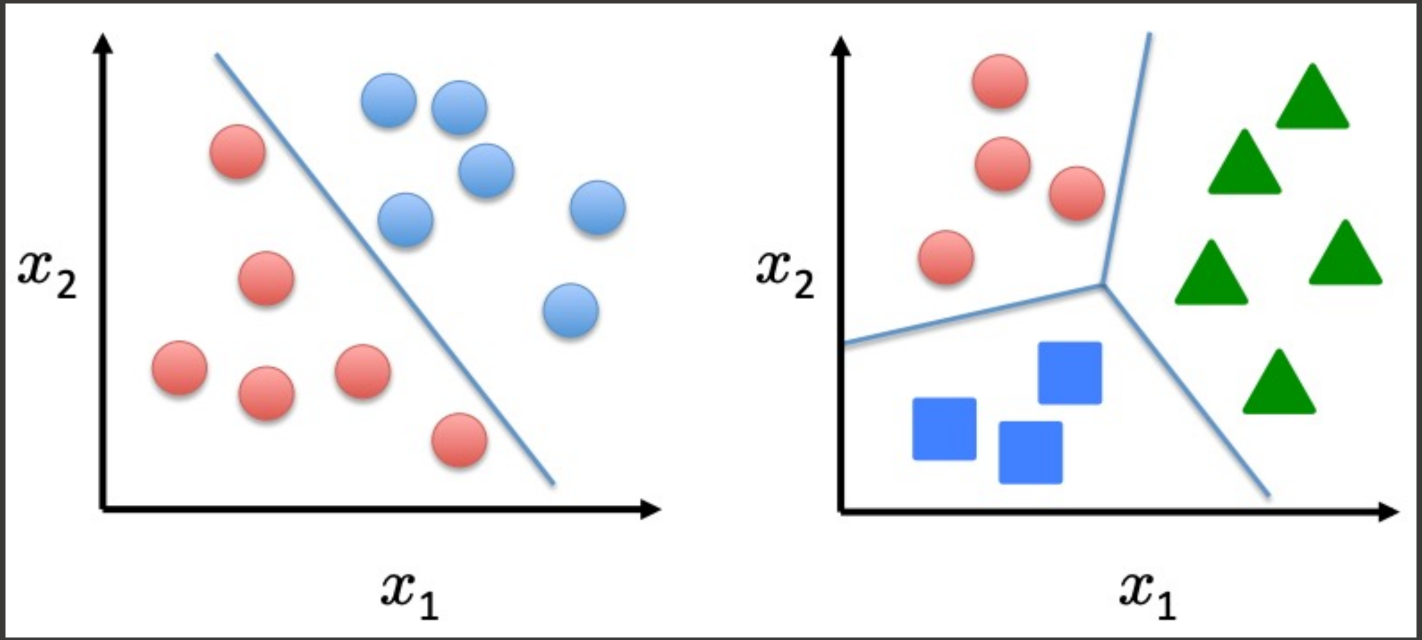
$$0 < w_0 + \sum_{i=1}^{D} w_i x_i$$

and assigns $Y = 1$ otherwise.

- The parametric form of $P(Y|X)$ used by Logistic Regression is precisely the form implied by the assumptions of a Gaussian Naïve Bayes classifier.
- Therefore, we can view Logistic Regression as a closely related alternative to GNB, though the two can produce different results in many cases.

Multi-Class Classification

- Multi-Class Classification applications:
    - Healthy/cold/flu/pneumonia
    - Frog/car/cat/human



Multi-Class Classification

- For 2 classes

$$y_w(x) = \frac{1}{1 + \exp(-w^T\mathbf{x})} = \frac{\exp(w^T\mathbf{x})}{\underbrace{\boxed{1}}_{\text{Weight assigned to y=0}} + \underbrace{\boxed{\exp(w^T\mathbf{x})}}_{\text{Weight assigned to y=1}}}$$

- For multi-classes with C classes $\{1, \dots, K\}$

$$p(y = k | x; w_1, \dots, w_K) = \frac{\exp(w_k^T\mathbf{x})}{\sum_k \exp(w_k^T\mathbf{x})}$$

Softmax function

Multi-Class Classification

- We can use $\frac{\exp(w_k^T \mathbf{x})}{\sum_{k=1}^{K} \exp(w_k^T \mathbf{x})}$ as the model for class K

- Gradient descent simultaneously updates all parameters for all models

- Predict class label as the most probable model

$$\max_{K} \frac{\exp(w_k^T \mathbf{x})}{\sum_{k=1}^{K} \exp(w_k^T \mathbf{x})}$$