EECS 551 Discussion 12

Zongyu Li

University of Michigan

December 3, 2021

Today's Agenda

- Introduction to logistic regression
- Task 6

Logistic Regression

- Logistic function, aka sigmoid function, is

$$f(x) = \frac{1}{1 + e^{-x}}.$$

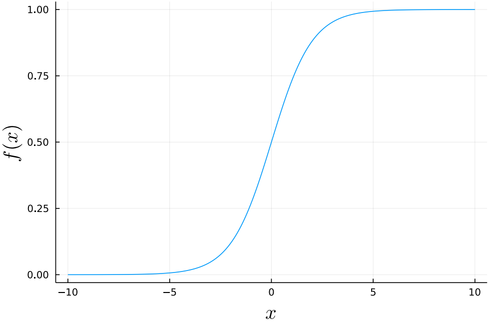
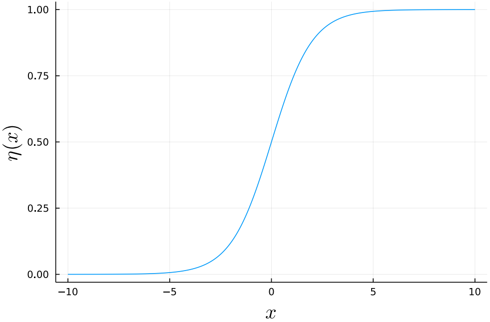


Figure: Logistic function.

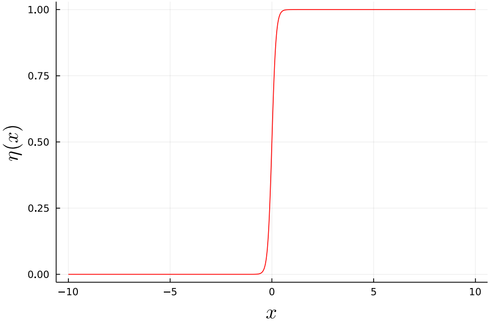
Logistic Regression

- Weight and bias terms?

$$\eta(x) = \frac{1}{1 + e^{-(wx+b)}}.$$



(a) $w = 1, b = 0$

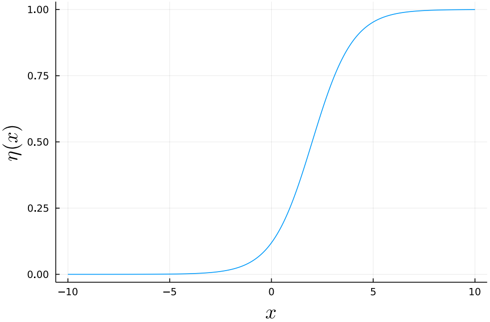


(b) $w = 10, b = 0$

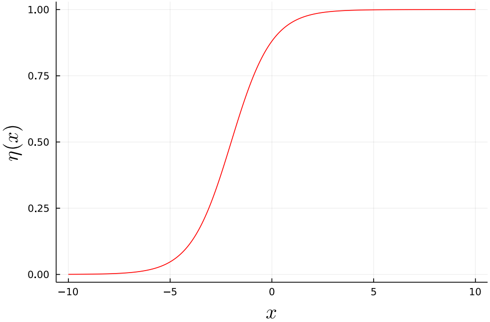
Logistic Regression

- Weight and bias terms?

$$\eta(x) = \frac{1}{1 + e^{-(wx+b)}}.$$



(a) $w = 1, b = -2$



(b) $w = 1, b = 2$

Logistic Regression

- Interpretation?
- $\eta(x)$ builds a continuous relationship between probability and variable x. For example, raining and humidity, disease and age, etc.
- The probability that it will be rainy tomorrow is a continuous, monotonically increasing function of today's humidity.
- The probability that having Alzheimer's disease is a continuous, monotonically increasing function of age.
- Sounds logistic?
- We can use it for binary classification!
- If $\eta(x) \geq 0.5$, we believe something is true, otherwise is false.

Logistic Regression

- Mathematically, consider a binary classification problem with labels $\boldsymbol{y} \in \{-1, 1\}$, our (Bayes) classifier is

$$f(\boldsymbol{x}) \triangleq \begin{cases} 1, & \eta(\boldsymbol{x};\boldsymbol{\theta}) \geq 0.5 \\ -1, & \text{otherwise} \end{cases},$$

  where

$$\eta(\boldsymbol{x};\boldsymbol{\theta}) \triangleq \frac{1}{1+e^{-(\boldsymbol{w}'\boldsymbol{x}+b)}}, \quad \boldsymbol{w}, \boldsymbol{x} \in \mathbb{R}^N, b \in \mathbb{R}, \boldsymbol{\theta} \triangleq \{\boldsymbol{w}, b\}.$$

- $\eta(\boldsymbol{x},\boldsymbol{\theta})$ models the conditional probability for label "1"

$$p(Y = 1|\boldsymbol{x},\boldsymbol{\theta}) = \eta(\boldsymbol{x};\boldsymbol{\theta}).$$

Logistic Regression

- Given data $\{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots, (\boldsymbol{x}_N, y_N)\}$, our goal is to find a set of parameters $\boldsymbol{\theta} = \{\boldsymbol{w}, b\}$ to maximize the conditional probability

$$p(Y_1 = y_1, Y_2 = y_2, \ldots, Y_N = y_N | \boldsymbol{X}, \boldsymbol{\theta}),$$

  where Y_i are random samples, $\boldsymbol{X} = \{\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_N\}$.
- We assume the conditional independence of labels given $\boldsymbol{X}$ and $\boldsymbol{\theta}$.
- Then the objective becomes

$$\underset{\boldsymbol{\theta}}{\text{maximize}} \quad \prod_{i=1}^{N} p(Y_i = y_i | \boldsymbol{x}_i, \boldsymbol{\theta}).$$

- Question: how to represent $p(Y_i = y_i | \boldsymbol{x}_i, \boldsymbol{\theta})$ using $\eta(\boldsymbol{x}_i; \boldsymbol{\theta})$?

Logistic Regression

- We know
$$p(Y=1|\boldsymbol{x}_i,\boldsymbol{\theta}) = \eta(\boldsymbol{x}_i,\boldsymbol{\theta}) = \frac{1}{1+e^{-(\boldsymbol{w}'\boldsymbol{x}_i+b)}},$$
  and
$$p(Y=-1|\boldsymbol{x}_i,\boldsymbol{\theta}) = 1-\eta(\boldsymbol{x}_i,\boldsymbol{\theta}) = \frac{e^{-(\boldsymbol{w}'\boldsymbol{x}_i+b)}}{1+e^{-(\boldsymbol{w}'\boldsymbol{x}_i+b)}} = \frac{1}{1+e^{(\boldsymbol{w}'\boldsymbol{x}_i+b)}}.$$
- Combining these two cases, we have
$$p(Y_i=y_i|\boldsymbol{x}_i,\boldsymbol{\theta}) = \frac{1}{1+e^{-y_i(\boldsymbol{w}'\boldsymbol{x}_i+b)}}$$
- Hence, our objective becomes
$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmax}} \prod_{i=1}^{N} \frac{1}{1+e^{-y_i(\boldsymbol{w}'\boldsymbol{x}_i+b)}}.$$

Logistic Regression

- Taking the negative log-likelihood yields

$$\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \sum_{i=1}^{N} \log\left(1 + e^{-y_i(\boldsymbol{w}'\boldsymbol{x}_i + b)}\right).$$

- Define $h(z) \triangleq \log(1 + e^{-z})$, and set $b = 0$, we reached the form in lecture notes

$$\hat{\boldsymbol{w}} = \underset{\boldsymbol{w}}{\operatorname{argmin}}\, \mathbf{1}'\, h_{.}(\boldsymbol{A}\boldsymbol{w}),$$

where

$$\boldsymbol{A} \triangleq \begin{bmatrix} y_1\boldsymbol{x}_1' \\ y_2\boldsymbol{x}_2' \\ \cdot \\ \cdot \\ y_N\boldsymbol{x}_N' \end{bmatrix}.$$

Logistic Regression

For the logistic loss, the cost function is not quadratic, but it does have a Lipschitz continuous gradient. For gradient-based optimization, we need the cost function gradient:

$$\underbrace{\nabla \Psi(\boldsymbol{x})}_{\text{in } \mathbb{F}^N} = \nabla \left(\mathbf{1}_M' \, h.(\boldsymbol{A}\boldsymbol{x}) + \beta \frac{1}{2} \|\boldsymbol{x}\|_2^2 \right) = \left(\sum_{m=1}^M \nabla h(\boldsymbol{A}_{m,:}\boldsymbol{x}) \right) + \beta \boldsymbol{x}$$

$$= \left(\sum_{m=1}^M \boldsymbol{A}_{m,:}' \, \dot{h}(\boldsymbol{A}_{m,:}\boldsymbol{x}) \right) + \beta \boldsymbol{x} = \boldsymbol{A}' \dot{h}.(\boldsymbol{A}\boldsymbol{x}) + \beta \boldsymbol{x}. \tag{8.12}$$

The cost function **Hessian matrix** is:

$$\nabla^2 \Psi(\boldsymbol{x}) = \nabla^T \nabla \Psi(\boldsymbol{x}) = \nabla^T \left(\sum_{m=1}^M \boldsymbol{A}_{m,:}' \, \dot{h}(\boldsymbol{A}_{m,:}\boldsymbol{x}) + \beta \boldsymbol{x} \right) = \sum_{m=1}^M \boldsymbol{A}_{m,:}' \, \ddot{h}(\boldsymbol{A}_{m,:}\boldsymbol{x}) \boldsymbol{A}_{m,:} + \beta \boldsymbol{I}$$

$$= \boldsymbol{A}' \operatorname{Diag}\underbrace{\left\{ \ddot{h}.(\boldsymbol{A}\boldsymbol{x}) \right\}}_{\geq 0} \boldsymbol{A} + \beta \boldsymbol{I} = \underbrace{\underbrace{\boldsymbol{A}' \boldsymbol{D}(\boldsymbol{x}) \boldsymbol{A}}_{\succeq 0} + \beta \underbrace{\boldsymbol{I}}_{\succ 0}}_{\succ 0}, \qquad \boldsymbol{D}(\boldsymbol{x}) \triangleq \operatorname{Diag}\left\{ \ddot{h}.(\boldsymbol{A}\boldsymbol{x}) \right\} \succeq 0$$

27. Ψ is a **strictly convex** function when ψ is the **logistic** loss function and $\beta > 0$. (?)

A: True B: False ??

Logistic Regression

To apply GD to the cost function (8.11), we need a Lipschitz constant for its gradient.

Next we describe two different ways of deriving a bound.

Method 1.
Start with the gradient expression (8.12):

$$\begin{aligned}
\left\|\nabla \Psi(\boldsymbol{x}) - \nabla \Psi(\boldsymbol{z})\right\|_2 &= \left\|\boldsymbol{A}'\dot{h}.(\boldsymbol{A}\boldsymbol{x}) + \beta \boldsymbol{x} - \boldsymbol{A}'\dot{h}.(\boldsymbol{A}\boldsymbol{z}) - \beta \boldsymbol{z}\right\|_2 \\
&= \left\|\boldsymbol{A}'(\dot{h}.(\boldsymbol{A}\boldsymbol{x}) - \dot{h}.(\boldsymbol{A}\boldsymbol{z})) + \beta(\boldsymbol{x} - \boldsymbol{z})\right\|_2 \\
&\leq \left\|\boldsymbol{A}'\right\|_2 \left\|\dot{h}.(\boldsymbol{A}\boldsymbol{x}) - \dot{h}.(\boldsymbol{A}\boldsymbol{z})\right\|_2 + \beta \left\|\boldsymbol{x} - \boldsymbol{z}\right\|_2 \\
&\leq \left\|\boldsymbol{A}'\right\|_2 L_{\dot{h}} \left\|\boldsymbol{A}\boldsymbol{x} - \boldsymbol{A}\boldsymbol{z}\right\|_2 + \beta \left\|\boldsymbol{x} - \boldsymbol{z}\right\|_2 \\
&\leq L_{\dot{h}} \left\|\boldsymbol{A}'\right\|_2 \left\|\boldsymbol{A}\right\| \left\|\boldsymbol{x} - \boldsymbol{z}\right\|_2 + \beta \left\|\boldsymbol{x} - \boldsymbol{z}\right\|_2 = \left(\left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2 L_{\dot{h}} + \beta\right) \left\|\boldsymbol{x} - \boldsymbol{z}\right\|_2
\end{aligned}$$

$$\implies L_{\nabla \Psi} = L_{\dot{h}} \left\|\boldsymbol{A}'\boldsymbol{A}\right\|_2 + \beta.$$

For the second inequality we used the fact that $\dot{h}$ is Lipschitz:

$$\left\|\dot{h}.(\boldsymbol{s}) - \dot{h}.(\boldsymbol{t})\right\|_2^2 = \sum_m \left|\dot{h}(s_m) - \dot{h}(t_m)\right|^2 \leq \sum_m L_{\dot{h}}^2 \left|s_m - t_m\right|^2 = L_{\dot{h}}^2 \left\|\boldsymbol{s} - \boldsymbol{t}\right\|_2^2.$$