Math 7243

Machine Learning and Statistical Learning Theory

Section 3. Logistic Regression

Instructor: He Wang
Department of Mathematics
Northeastern University

Review :

1. Linear regression
2. Locally weighted linear regression
3. Ridge Regression
4. Lasso Regression

Training Data:

features labels

$$D = \{(\vec{x}^{(1)}, y^{(1)}), \ldots, (\vec{x}^{(n)}, y^{(n)})\}$$

Assumption: Linear Regression

$$h(\vec{x}) = \vec{\theta}^T \vec{x} = \theta_0 + \theta_1 x_1 + \cdots + \theta_d x_d$$

Today:

1. Logistic Regression (binary)
2. Softmax Regression (multiclass)

$\vec{x} = \begin{bmatrix} x_0 \\ x_1 \\ x_2 \end{bmatrix}$

- Example: Data of students sleep time, study time, and pass/fail.

If we know the test scores, we can use linear regression to predict the test scores.

x_1 slept	x_2 studied	Y	Score z
7.63	7.40	1	90
2.03	3.93	0	50
3.82	0.72	0	40
7.15	3.89	1	80
6.47	8.19	1	90
...	...	...	...

$P(Y=1 \mid x_1=7, x_2=7) = ?$

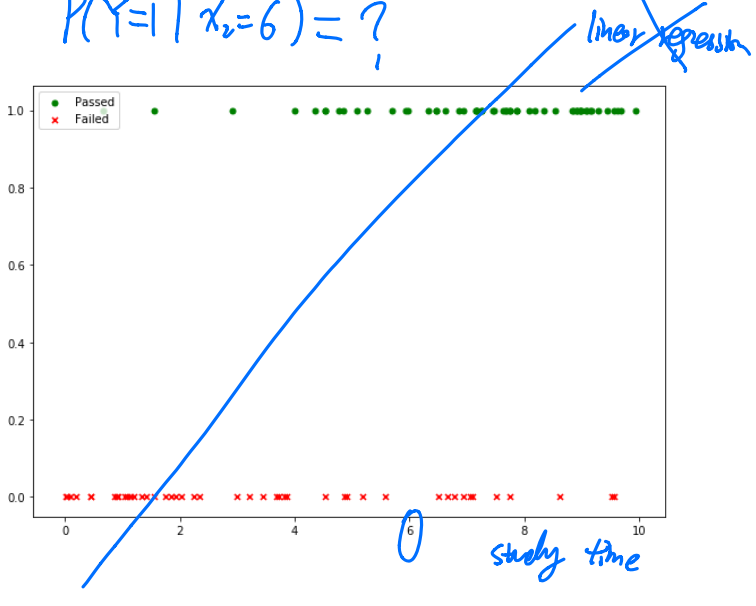
$P(Y=0 \mid x_1=2, x_2=2) = ?$

- Logistic regression

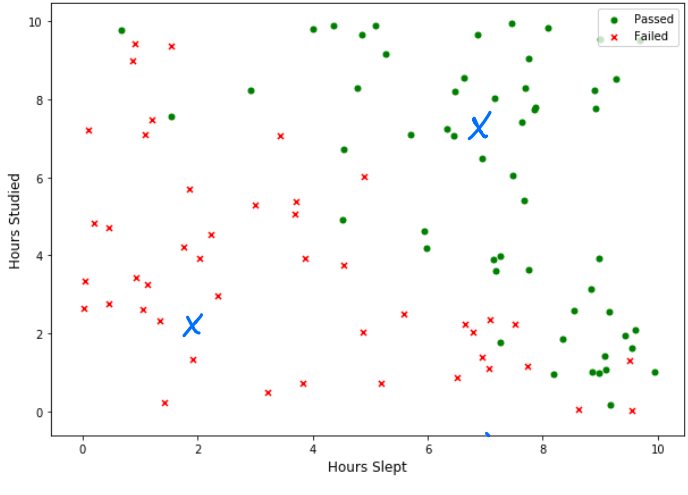
$h(\vec{x}) = P(Y=1 \mid \vec{x})$

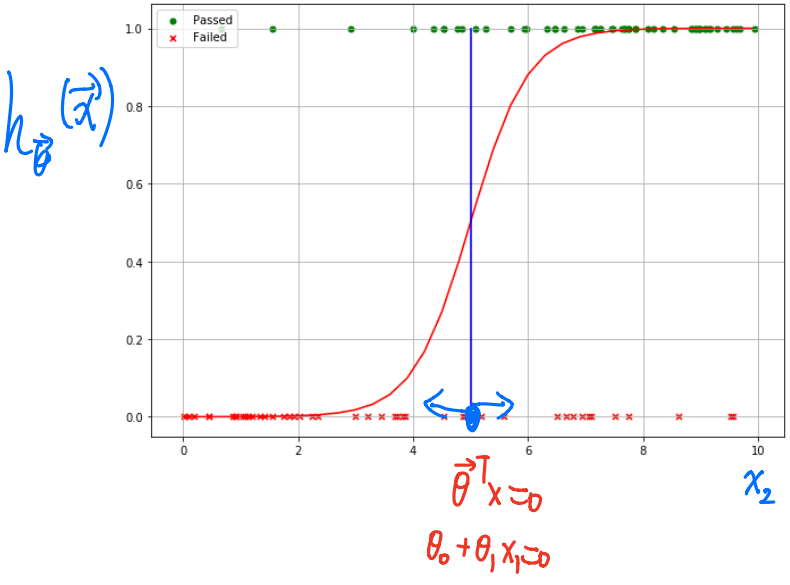
Logistic regression is a classification algorithm, used to predict probabilities based on given set of independent variables.

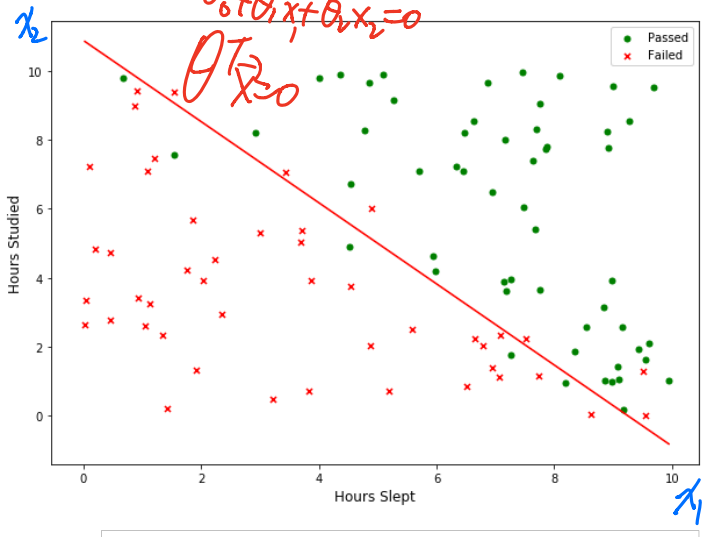
$P(Y=1 \mid x_2=6) = ?$



passed Y=1, failed Y=0



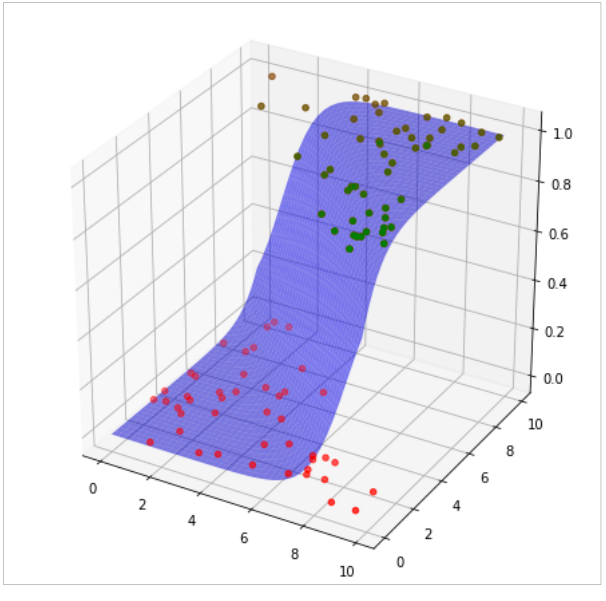




$h(\vec{x})=as\ for\ \vec{x}\in\{\vec{x}\in\mathbb{R}^d \mid f(\vec{x})=0\}$

- Decision Boundary $\theta^T\vec{x}=f(\vec{x})=0$

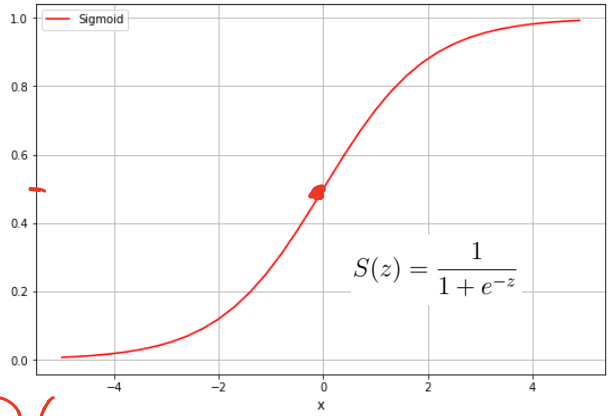
logistic regression prediction function returns a probability between 0 and 1, in order to predict which class this data belongs we need to set a threshold.



- Sigmoid function.

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

$S(0)=0.5$



The sigmoid function maps any real value into a value in [0,1].

Hypothesis:
Goal: $h_{\vec{\theta}}(\vec{x}) = S(\vec{\theta}^T\vec{x}) = \frac{1}{1+e^{-\vec{\theta}^T\vec{x}}}$ = probability $P(y=1 \mid \vec{x})$

class Prediction:

$$y(\vec{x}) = \begin{cases} 1\text{-pos} & \text{if } h(\vec{x}) \geq 0.5 \\ 0\text{-fail} & \text{if } h(\vec{x}) < 0.5 \end{cases}$$

Boundary $\vec{\theta}^T\vec{x}=0$

Remark: $\log\left(\frac{h(\vec{x})}{1-h(\vec{x})}\right) = \vec{\theta}^T\vec{x}$

$= \log\left(\frac{P(y=1 \mid \vec{x})}{P(y=0 \mid \vec{x})}\right)$

Other functions

① Step function $g(z) = \begin{cases} 0 & z<0 \\ 1 & z \geq 0 \end{cases}$

② Sign function $g(z) = \begin{cases} -1 & z<0 \\ 1 & z>0 \end{cases}$

③ Tanh $g(z) = \frac{e^z - e^{-z}}{e^z + e^{-z}}$.

$y \in \{0, 1\}$

Assumption: $S(\theta^T \vec{x}) = \frac{1}{1+e^{-\theta^T \vec{x}}}$

$$\begin{cases} P(y=1|\vec{x};\theta) = h_\theta(\vec{x}) \\ P(y=0|\vec{x};\theta) = 1 - h_\theta(\vec{x}) \end{cases} \Rightarrow \quad P(y|\vec{x};\theta) = h_\theta(\vec{x})^y \left(1 - h_\theta(\vec{x})\right)^{1-y}$$

or -1

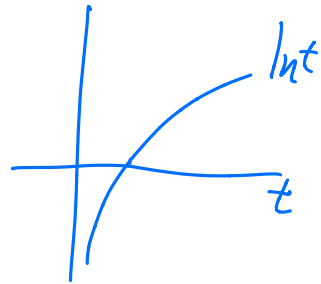
maximize

Likelihood function:

$$L(\theta) = P(\vec{y} \mid X; \theta)$$

$$= \prod_{i=1}^{n} P(y^{(i)} \mid x^{(i)}; \theta) \qquad \text{(i.i.d)}$$

$$= \prod_{i=1}^{n} h(x^{(i)})^{y^{(i)}} (1 - h(x^{(i)}))^{1-y^{(i)}}$$



maximize log likelihood:

$$l(\theta) = \ln(L(\theta))$$

$$= \sum_{i=1}^{n} \left(y^{(i)} \underbrace{\ln h(x^{(i)})}_{<0} + (1-y^{(i)}) \underbrace{\ln(1 - h(x^{(i)}))}_{<0} \right)$$

$(x^{(i)}, y^{(i)})$

log cost for one observation:

$$\text{cost}(\theta) = \begin{cases} -\ln(h(x^{(i)})) & \text{if } y^{(i)} = 1 \\ -\ln(1 - h(x^{(i)})) & \text{if } y^{(i)} = 0 \end{cases}$$

$$= -\left(y^{(i)} \ln(h(x^{(i)})) + (1-y^{(i)}) \ln(1 - h(x^{(i)})) \right)$$

minimize log cost function:

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

$$h(x^{(i)}) = S(\theta^T x^{(i)})$$

$$S'(z) = S(z)(1-S(z))$$

$$\frac{\partial h(x^{(i)})}{\partial \theta_j} = S(z)(1-S(z)) \cdot x_j^{(i)}$$

- Gradient: $\nabla J(\theta) = \begin{bmatrix} \frac{\partial J(\theta)}{\partial \theta_0} \\ \vdots \\ \frac{\partial J(\theta)}{\partial \theta_d} \end{bmatrix}$

$$\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{n}\sum_{i=1}^{n}\left(y^{(i)} \frac{1}{S(z)} S(z)(1-S(z)) \cdot x_j^{(i)} - (1-y^{(i)}) \frac{1}{1-S(z)} S(z)(1-S(z)) \cdot x_j^{(i)}\right)$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left(S(\theta^T x^{(i)}) - y^{(i)}\right) x_j^{(i)} = \frac{1}{n}\sum_{i=1}^{n}\left(h(x^{(i)}) - y^{(i)}\right) x_j^{(i)}$$

- Gradient descent: $\vec{\theta}^{next} = \vec{\theta} - \alpha \nabla J(\vec{\theta})$
- Newton's method: $\vec{\theta}^{next} = \vec{\theta} - H^{-1} \nabla J(\vec{\theta})$

Hessian $H_{jk} = \frac{\partial^2 J}{\partial \theta_j \partial \theta_k} = \frac{1}{n}\sum_{i=1}^{n} h(x^{(i)})(1-h(x^{(i)})) x_j^{(i)} x_k^{(i)}$

Question: If $y \in \{-1, 1\}$, assume $\begin{cases} P(y=1 \mid \vec{x}, \vec{\theta}) = h_\theta(\vec{x}) \\ P(y=-1 \mid \vec{x}, \vec{\theta}) = 1 - h_\theta(\vec{x}) \end{cases}$

find $\nabla J(\theta)$.

- $P(y \mid \vec{x}; \theta) = h_\theta(\vec{x})^{\frac{1+y}{2}} \left(1 - h_\theta(\vec{x})\right)^{\frac{1-y}{2}}$

  better $\rightarrow h_\theta(y\vec{x})$

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

  $= \frac{1}{n} \sum_{i=1}^{n} \ln\left(1 + e^{-y^{(i)}\theta^T x^{(i)}}\right)$

- Gradient ∇J:

  $\frac{\partial J(\theta)}{\partial \theta_j} = -\frac{1}{n} \sum_{i=1}^{n} \left(1 - h\left(y^{(i)}x^{(i)}\right)\right) y^{(i)} x_j^{(i)}$

Gradient descent: $\theta^{\text{next}} := \theta - \alpha \nabla J(\theta)$

To use Newton's method $\theta^{\text{next}} := \theta - H^{-1}\nabla J$

- Hessian matrix H

  $H_{jk} = \frac{\partial^2 J}{\partial \theta_j \partial \theta_k}$

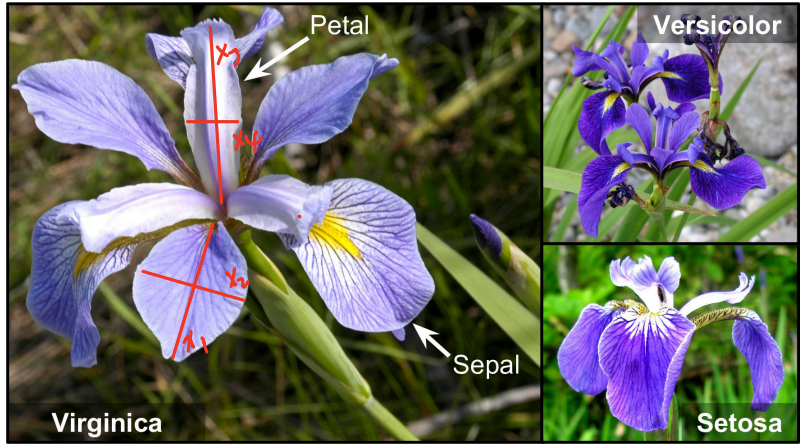
  $= \frac{1}{n} \sum_{i=1}^{n} h\left(y^{(i)}x^{(i)}\right)\left(1 - h\left(y^{(i)}x^{(i)}\right)\right) y^{(i)} x_j^{(i)} y^{(i)} x_k^{(i)}$

  $= \frac{1}{n} \sum_{i=1}^{n} h\left(y^{(i)}x^{(i)}\right)\left(1 - h\left(y^{(i)}x^{(i)}\right)\right) x_j^{(i)} x_k^{(i)}$

- **Softmax Regression** (Multinomial Logistic Regression)

  - Flowers of three iris plant species:

The famous Iris database, first used by Sir R.A. Fisher(1936), is best known database to be found in the pattern recognition literature. It contains the **sepal** and **petal length** and **width** of 150 iris flowers of three different species: Iris-Setosa, Iris-Versicolor, and Iris-Virginica.



$\vec{x} \in \mathbb{R}^4$

Data features: sepal length, sepal width, petal length, petal width. (x_1 x_2 x_3 x_4)

$y \in \mathbb{Z}_2^3 = \{0,1\}^3$

Classes: 0-Iris-Setosa, 1-Iris-Versicolour, 2-Iris-Virginica $y \in \{0, 1, 2\}$ or $\{\bullet, \blacksquare, \blacktriangle\}$

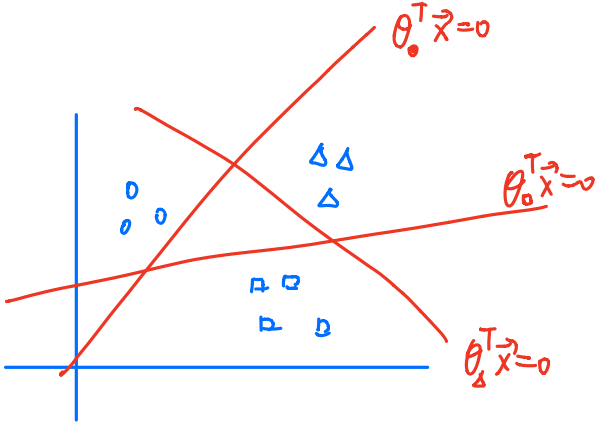
$$y \in \left\{ \begin{bmatrix}1\\0\\0\end{bmatrix} \begin{bmatrix}0\\1\\0\end{bmatrix} \begin{bmatrix}0\\0\\1\end{bmatrix} \right\} \quad \text{or} \quad \left\{ \begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1\\0\end{bmatrix}, \begin{bmatrix}0\\1\end{bmatrix} \right\}$$

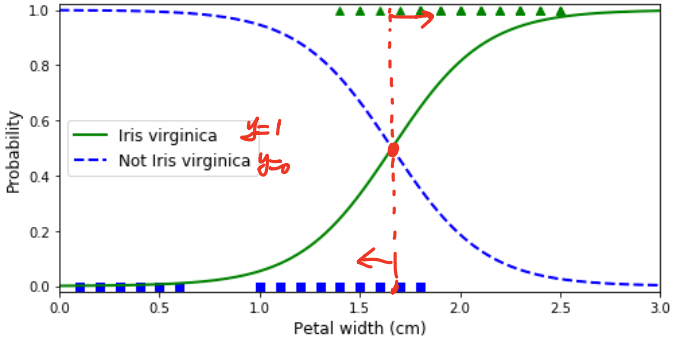
$\bullet \quad \blacksquare \quad \blacktriangle$

x_1	x_2	x_3	x_4	y
[5.1,	3.5,	1.4,	0.2],	0
[4.9,	3. ,	1.4,	0.2],	0
[4.7,	3.2,	1.3,	0.2],	0
[4.6,	3.1,	1.5,	0.2],	⋮
[5. ,	3.6,	1.4,	0.2],	⋮
[5.4,	3.9,	1.7,	0.4],	
[4.6,	3.4,	1.4,	0.3],	
[5. ,	3.4,	1.5,	0.2],	
[4.4,	2.9,	1.4,	0.2]	2
...				⋮

Want $\vec{\theta}_\bullet, \vec{\theta}_\blacksquare, \vec{\theta}_\blacktriangle \in \mathbb{R}^4$

$$\theta = \underset{4\times 3}{\left[\vec{\theta}_\bullet \; \vec{\theta}_\blacksquare \; \vec{\theta}_\blacktriangle\right]}$$

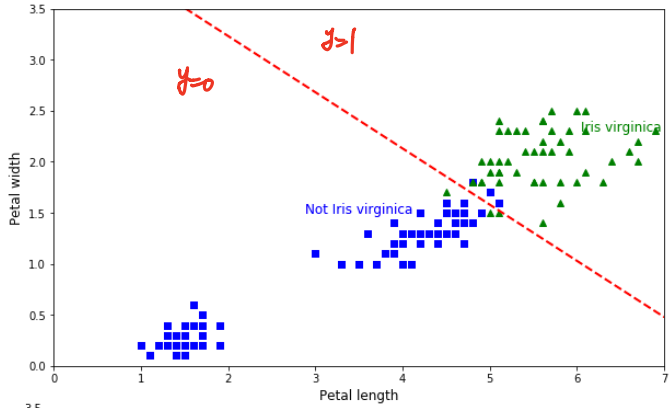


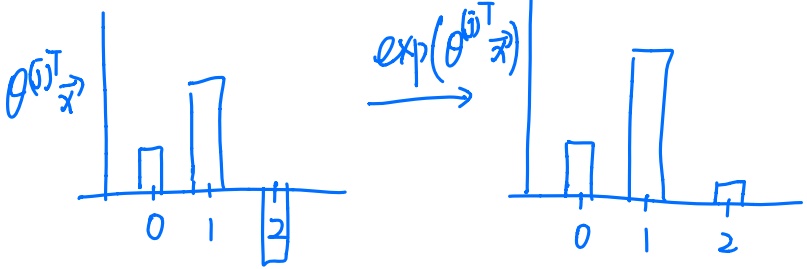


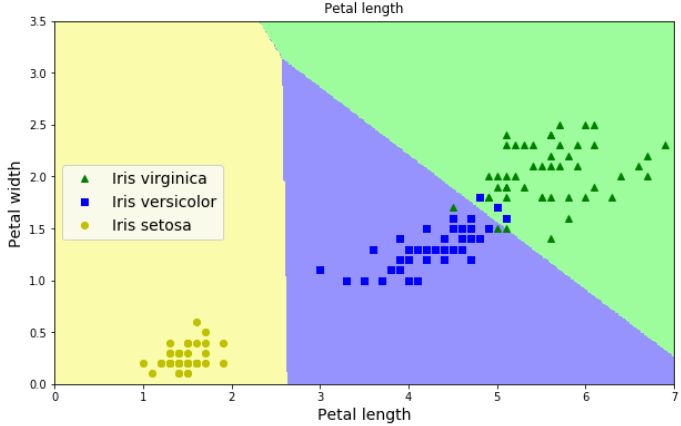
Hypothesis:

$$h_\theta(\vec{x}) = \begin{bmatrix} P(y=0|\vec{x};\theta) \\ P(y=1|\vec{x};\theta) \\ P(y=2|\vec{x};\theta) \end{bmatrix} = \frac{1}{\sum_{j=0}^{2} \exp(\theta^{(j)T}\vec{x})} \begin{bmatrix} \exp(\theta^{(0)T}\vec{x}) \\ \exp(\theta^{(1)T}\vec{x}) \\ \exp(\theta^{(2)T}\vec{x}) \end{bmatrix}$$

↑ distribution vector







$$\rightarrow \frac{\exp(\theta^{(i)T}\vec{x})}{\sum_{j=0}^{2}\exp(\theta^{(j)T}\vec{x})}$$

log. Cost Function: (Cross entropy cost function)

$$H(p,q) = -\sum p(x)\log(q(x))$$

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\Big(1\{y^{(i)}=0\}\ln\big(P(y^{(i)}=0\mid\vec{x}^{(i)};\theta)\big) + 1\{y^{(i)}=1\}\ln\big(P(y^{(i)}=1\mid x^{(i)};\theta)\big) + 1\{y^{(i)}=2\}\ln\big(P(y^{(i)}=2\mid x^{(i)};\theta)\big)$$

$$= -\frac{1}{n}\sum_{i=1}^{n}\sum_{k=0}^{2}\left(1\{y^{(i)}=k\}\ln\left(\frac{\exp(\theta^{(k)T}\vec{x}^{(i)})}{\sum_{j=0}^{K}\exp(\theta^{(j)T}\vec{x}^{(i)})}\right)\right)$$

$$h(\vec{x}^{(i)})^{(k)}$$

For the case of binary classification

$$h_\theta(\vec{x}) = \frac{1}{\exp(\theta^{(1)T}\vec{x}) + \exp(\theta^{(2)T}\vec{x})}\begin{bmatrix}\exp(\theta^{(1)T}\vec{x})\\ \exp(\theta^{(2)T}\vec{x})\end{bmatrix}$$

$$= \begin{bmatrix}\frac{1}{1+\exp((\theta^{(1)T}-\theta^{(2)T})\vec{x})}\\ 1-\frac{1}{1+\exp((\theta^{(1)T}-\theta^{(2)T})\vec{x})}\end{bmatrix}$$

Set $\theta = \theta^{(1)} - \theta^{(2)}$

$$= \begin{bmatrix}\frac{1}{1+\exp(-\theta^T\vec{x})}\\ 1-\frac{1}{1+\exp(-\theta^T\vec{x})}\end{bmatrix}$$

where $1\{\ \}$ is the "indicator function" $\begin{cases} 1\{\text{true}\} = 1 \\ 1\{\text{false}\} = 0 \end{cases}$

- Use gradient descent or Newton's method to find $\theta^{(0)}, \theta^{(1)}, \theta^{(2)} \ldots$
  ? H is symbol.
- Gradient $\nabla_{\theta^{(k)}} J(\theta) = \frac{1}{n} \sum_{i=1}^{n} \left(h(x^{(i)})^{(k)} - 1(y^{(i)} = k) \right) x^{(i)}$