HW7-Notes

March 2020

$p(y \mid x; \theta) = h_\theta(x)^y (1 - h_\theta(x))^{1-y}$
$y = 1 \Rightarrow p(y = 1 \mid x; \theta) = h_\theta(x)^1 (1 - h_\theta(x))^0$
$y = 0 \Rightarrow p(y = 0 \mid x; \theta) = 1 - h_\theta(x)$
$L(\theta) = p(y = 0 \mid x; \theta) = \prod_i p(y^i \mid x^i; \theta) = \prod_i h_\theta(x^i)^{y^i} (1 - h_\theta(x^i))^{1-y^i})$
It is much easier to maximize the log likelihood.
$l(\theta) = logL(\theta) = \sum_i [y^i log h_\theta(x^i) + (1 - y^i) log(1 - h_\theta(x^i))]$
How to maximize it? Use (stochastic) gradient descent.
$\theta := \theta \bigoplus \alpha \nabla_\theta l(\theta)$
$\bigoplus$: along with the gradient direction.
$\frac{\partial}{\partial \theta_j} l(\theta) = \sum_{i=1}^n (y^i - h_\theta(x^i)) x_j^i$