Classification

- Is a discrete categorization of an output variable, e.g. $y \in \{0, 1\}$
  - Can also have multiple classes, so some set $S \mid \text{len}(S) > 2$
- Linear regression -> threshold classifier output
  - E.g. if $h_\theta(x) \geq 0.5$ do $y = 1$ else $y = 0$
  - However is not effective when there are > 2 clusters -> DNU
- Classification - $y = 0$ or $y = 1$
  - $0 \leq h_\theta(x) \leq 1$
  - Binary classification

Hypothesis Representation

- Need $0 \leq h_\theta(x) \leq 1$
- $h\theta_0(x) = g(\theta^T x)$ where $g = \frac{1}{1+e^{-z}}$ is the sigmoid = logistical function and $z = \theta^T x$
  - $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$
- Interpretation
  - $h_\theta(x)$ is estimated probability that $y = 1$ on input x
  - $h_\theta(x) = P(y = 1 | x; \theta)$ is probability of $y = 1$ given x parameterized by θ
    - Due to total sum probability -> $P(y = 0|x; \theta) = 1 - P(y = 1|x; \theta)$

Decision Boundary

- Prediction boundary - If $h_\theta(x) \geq 0.5$ do $y = 1$ else $y = 0$
  - Thus $y = 0 \implies \theta^T x < 0$ and $y = 1 \implies \theta^T x \geq 0$
  - Graph the equation $\theta_0 + \theta_1 x_1 + ... + \theta_n x_n \geq 0$ (higher order planes of $\mathbb{R}^{n+1}$)
- Nonlinear decision boundaries
  - Have polynomial terms in features
  - Ex. $-1 + x_1^2 + x_2^2 \geq 0 \implies$ unit circle

Logistic cost function

- Setup/prime
  - Training set $S = \left\{ \left(x^{(1)}, y^{(1)}\right), \left(x^{(2)}, y^{(2)}\right), \cdots, \left(x^{(m)}, y^{(m)}\right) \right\}$
    - m examples
  - A feature vector $x \in \begin{bmatrix} x_0 \\ x_1 \\ \vdots \\ x_n \end{bmatrix}$
  - $h_\theta(x) = \frac{1}{1+e^{-\theta^T x}}$
- Cost $\left(h_\theta\left(x\right), y\right) = \frac{1}{2}\left(h_\theta\left(x\right) - y\right)^2$
  - Is non-convex - many local extrema which may hinder gradient descent
- Cost function for logistic regression

$$\boxed{\text{Cost}\left(h_\theta(x), y\right) = \begin{cases} -\log\left(h_\theta(x)\right) \text{ if } y = 1 \\ -\log\left(1 - h_\theta(x)\right) \text{ if } y = 0 \end{cases}}$$

- Cost function behavior
  - Cost is 0 if $y = 1, h_\theta(x) = 1$
  - As $h_\theta(x) \to 0$, Cost $\to \infty$
  - If $h_\theta(x) = 0$ **but** $y = 1$, then learning algorithm penalized heavily