# Logistic regression & cross entropy Loss

## Setup: Logistic regression

Given data $\{(x_i, y_i)\}_i$, $x_i \in \mathbb{R}^n$, $y_i \in \{0, 1\}$

We will assume $y_i$ was generated according to distribution

$$f(x) = \mathbb{P}(y=1 \mid x)$$

Idea: $x$ is a data point & $y$ is corresponding label. $f(x)$ is going to be a smooth function denoting the probability of a positive label. We want to do regression in the sense that we want to find some function $f_\theta(x)$ to $f(x)$. However, we don't have access to $f$, only noisy samples $y \sim f(x)$.

In logistic regression, we assume the following parameterized form for $f_\theta$.

$$f_\theta(x) = \sigma(\theta^T x)$$

Note: No bias b/c this can actually be added by augmenting $x$ with a 1 at the end.

We will try to fit $f_\theta$ to $D = \{(x_i, y_i)\}_i$ by maximizing the likelihood of the data.

~~Terminology: If $f_\theta$ is a parametric class of probability distributions, then the $\theta$ (likelihood~~

Terminology: Let $\underbrace{P(y \mid x ; \theta)}_{:= f_\theta(x)}$ denote the probability (density) of $y$ given $x$ under the density induced by $\theta$. In an abuse of notation, the Likelihood function given data $D = \{(x_i, y_i)\}$ is

$$L(\theta) = P(\{y_i\} \mid \{x_i\} ; \theta)$$

$$= \prod_{i=1}^{n} P(y_i \mid x_i ; \theta)$$

b/c here will always assume independence

Note

$$P(y_i = 1 \mid x_i ; \theta) = \sigma(\theta^T x_i) = \frac{1}{1 + e^{-\theta^T x_i}}$$

$$P(y_i = y_i \mid x_i ; \theta) = \sigma(\theta^T x_i)^{y_i} \left(1 - \sigma(\theta^T x_i)\right)^{1 - y_i}$$

Let

$$LL(\theta) := \log L(\theta)$$ denote the log-likelihood.

Have

$$LL(\theta) = \ln \prod_{i=1}^{n} P(y_i \mid x_i ; \theta)$$

$$= \sum_{i=1}^{n} \ln P(y_i \mid x_i ; \theta)$$

$$= \sum_i y_i \sigma(\theta^T x_i) + (1 - y_i)\left(1 - \sigma(\theta^T x_i)\right)$$

Comments:

- You dont need to use the $\sigma(\theta^T x_i)(1-\sigma(\cdot))^{1-y_i}$ trick. Just split over positive & negative labels.
- In book they use $y_i \in \{-1, 1\}$ and use the loss

$$\sum_i \ln\left(1 - e^{-y_i \theta^T x_i}\right)$$

This is the same thing.

Lets generalize: Multiple Classes & General model

Let $f_n(x; \theta) = \mathbb{P}(y_i = n \mid x; \theta)$

Where "n" denotes a class number

Let $N_i = \#\{x_i \text{ occurs in } D\}$ $P(y = n \mid x) :=$ Empirical distribution of $y$ given $x$

For a single $x$ (for which there may be $N > 1$ indep labels = # samples $y$) $= \#\{y = n$ when $x$ was feature vector$\}$

$$L(\theta \mid x) = \prod_n f_n(x; \theta)^{N P_i(y=n \mid x)}$$

$$\frac{1}{N} LL(\theta \mid x) = \frac{1}{N} \ln \prod f_n(x; \theta)^{N P(y=n)}$$

$$= \sum_n \ln f_n(x; \theta)^{P(y=n)}$$

$$= \sum_n P(y=n) \ln f_n(x; \theta)$$

$$= H(P, f(x;\theta)) \quad \text{where } f \cdot (f_n)_n$$

For many $x$ (but typically only 1 label $y$ per $x$)

$$LL(\theta) = \log \prod_i L(\theta | x_i)$$

$$= \sum_i LL(\theta | x_i)$$

$$= \sum_i H(p, f(x; \theta))$$

↑ Typically an indicator on $y$