Alternative to logistic regression (12 pts)

In class, we discussed the logistic regression model for binary classification problem. Here, we consider an alternative model. We have a training set $\{(\boldsymbol{x}_n, y_n)\}_{n=1}^N$ where $\boldsymbol{x}_n \in \mathbb{R}^{D+1}$ and $y_n \in \{0, 1\}$. Like in logistic regression, we will construct a probabilistic model for the probability that y_n belongs to class 0 or 1, given $\boldsymbol{x}_n$ and the model parameters, $\boldsymbol{\theta}_0$ and $\boldsymbol{\theta}_1$ ($\boldsymbol{\theta}_0, \boldsymbol{\theta}_1 \in \mathbb{R}^{D+1}$). More specifically, we model the target y_n as:

$$\begin{aligned} p(y_n = 0 | \boldsymbol{x}_n; \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= C e^{\boldsymbol{\theta}_0^\top \boldsymbol{x}} \\ p(y_n = 1 | \boldsymbol{x}_n; \boldsymbol{\theta}_0, \boldsymbol{\theta}_1) &= C e^{\boldsymbol{\theta}_1^\top \boldsymbol{x}} \end{aligned} \tag{1}$$

(a) (2 pts) Find the value of C that makes Equation 1 a valid probability distribution for y_n.

(1) $1 = \int_{-\infty}^{\infty} f(x)\,dx = \int_0^1 C e^{\theta_0^T x}\,dx = C \int_0^1 e^u \frac{du}{\theta_0^T} = \left(\frac{e^{\theta_0^T x}}{\theta_0^T}\right) C + \text{const}$

let $u = \theta_0^T x \Rightarrow du = \theta_0^T dx \Rightarrow dx = \frac{du}{\theta_0^T}$

$\Rightarrow 1 = C \frac{e^{\theta_0^T x}}{\theta_0^T}\Big|_0^1 = C\left(\frac{e^{\theta_0^T * (1)}}{\theta_0^T} - \frac{e^{\theta_0^T * 0}}{\theta_0^T}\right) =$

$1 = C * \left(\frac{e^{\theta_0^T} - 1}{\theta_0^T}\right) \Rightarrow C = \frac{\theta_0^T}{e^{\theta_0^T} - 1} = \frac{\theta_1^T}{e^{\theta_1^T} - 1}$

(b) (5 pts) Write the log likelihood of the parameters for a single instance $(\boldsymbol{x}_n, y_n)$: $l(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1) = \log p(y_n|\boldsymbol{x}_n; \boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$. Express your answer in terms of $y_n, \boldsymbol{x}_n, \boldsymbol{\theta}_0, \boldsymbol{\theta}_1$. (Hint: use the notation in class where for a Bernoulli random variable Y_n that takes values 1 with probability θ and 0 with probability $1 - \theta$, we can write its probability mass function: $p(y_n) = \theta^{y_n}(1-\theta)^{1-y_n}$).

(c) (2 pts) Write the log likelihood of the parameters $\mathcal{LL}(\boldsymbol{\theta}_0, \boldsymbol{\theta}_1)$ for the full training data.

(d) (3 pts) Recall that in logistic regression, we model the target y_n as :

$$\begin{aligned} p\left(y_n = 0 | \boldsymbol{x}_n; \boldsymbol{\theta}\right) &= 1 - \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) \\ p\left(y_n = 1 | \boldsymbol{x}_n; \boldsymbol{\theta}\right) &= \sigma(\boldsymbol{\theta}^\top \boldsymbol{x}) \end{aligned}$$

Here $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function. Show that our new model (Equation 1) is exactly the same as the logistic regression model with $\boldsymbol{\theta} = \boldsymbol{\theta}_1 - \boldsymbol{\theta}_0$.

Identities

Probability density/mass functions for some distributions

$$\text{Normal} \quad : \quad P(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\text{Multinomial} \quad : \quad P(\boldsymbol{x}; \boldsymbol{\pi}) = \prod_{k=1}^{K} {\pi_k}^{x_k}$$

$\boldsymbol{x}$ is a length K vector with exactly one entry equal to 1 and all other entries equal to 0

$$\text{Poisson} \quad : \quad P(x; \lambda) = \frac{\lambda^x \exp(-\lambda)}{x!}$$

Matrix calculus

Here $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{b} \in \mathbb{R}^n, \boldsymbol{A} \in \mathbb{R}^{n \times n}$. $\boldsymbol{A}$ is symmetric.

$$\nabla \boldsymbol{x}^{\mathrm{T}} \boldsymbol{A} \boldsymbol{x} = 2\boldsymbol{A}\boldsymbol{x}, \qquad \nabla \boldsymbol{b}^{\mathrm{T}} \boldsymbol{x} = \boldsymbol{b}$$

Entropy

The entropy $H(X)$ of a Bernoulli random variable $X \sim \mathit{Bernoulli}(p)$ for different values of p:

p	$H(X)$
$\frac{1}{2}$	1
$\frac{1}{3}$	0.92
$\frac{1}{4}$	0.81
$\frac{1}{5}$	0.73
$\frac{2}{5}$	0.97