2. Model Calibration

$h_\theta(x) \longrightarrow$ Why we might treat the output as a probability.

⇒ When the probabilities outputted by a model match empirical observation, the model is said to be well-calibrated.

↳ Logistic regression tends to output well calibrated probabilities.

(This is often not true with other classifiers such as Naive Bayes, or SVM)

⇒ Suppose we have a training set $\{(x^{(i)}, y^{(i)}) \mid x^{(i)} \in \mathbb{R}^{n+1} \text{ \& } y^{(i)} \in \{0,1\}\}$

⇒ Let $\theta \in \mathbb{R}^{n+1}$ be the maximum likelihood parameters learned after training a logistic regression model.

⇒ In order for the model to be considered well-calibrated, given any range of probabilities (a, b) such that $0 \leq a \leq b \leq 1$, and training examples $x^{(i)}$ where the model outputs $h_\theta(x^{(i)})$ fall in the range (a,b), the fraction of positives in that set of examples should be equal to the average of the model output for those examples.

$$\boxed{\frac{\sum_{i \in I_{a,b}} P(y^{(i)}=1 \mid x^{(i)}; \theta)}{|\{i \in I_{a,b}\}|} = \frac{\sum_{i \in I_{a,b}} \mathbb{I}\{y^{(i)} = 1\}}{|\{i \in I_{a,b}\}|}}$$

$$I_{a,b} = \{ i \mid i \in \{1, \dots m\}, h_\theta(x^{(i)}) \in (a,b) \}$$

$|S|$ = Size of Set S

(a) Show that the property holds true for

$$\left(\frac{\text{logistic regression}}{\text{model}} \right) \xrightarrow[\text{Over the range}]{} (a,b) = (0,1)$$

⇒ For Logistic regression:

$$\theta = \arg\max_\theta \, P(y^{(1)} \dots y^{(m)} \mid x^{(1)} \dots x^{(m)}; \theta)$$

$$\downarrow$$

$$\prod_{i=1}^{m} P(y^{(i)} \mid x^{(i)}; \theta) \qquad \left\{ h_\theta(x^{(i)}) = \frac{1}{1+e^{-\theta^T x^{(i)}}} \right\}$$

$$\downarrow$$

$$\prod_{i=1}^{m} h_\theta(x^{(i)})^{y^{(i)}} (1 - h_\theta(x^{(i)}))^{(1-y^{(i)})}$$

$$\theta = \arg\max_\theta \underbrace{\sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1-y^{(i)}) \log(1 - h_\theta(x^{(i)}))}_{l(\theta)}$$

⇒ To prove:

$$\frac{1}{m} \sum_{i=1}^{m} h_\theta(x^{(i)}) = \frac{1}{m} \sum_{i=1}^{m} \mathbb{I}\{y^{(i)} = 1\}$$

$$\sum_{i=1}^{m} h_\theta(x^{(i)}) = \sum_{i=1}^{m} \mathbb{I}\{y^{(i)} = 1\} = \sum_{i=1}^{m} y^{(i)}$$

$$\nabla l(\theta) = \sum_{i=1}^{m} (y^{(i)} - h_\theta(x^{(i)}))\theta$$

$\Rightarrow$ For the best θ $\nabla l(\theta) = 0$

$$\sum_{i=1}^{m} y^{(i)} = \sum_{i=1}^{m} h_\theta(x^{(i)})$$

(b) Both the answer are no.

(c) When a regularization $\lambda\|\theta\|$ is added, the equation becomes

$$\sum_{i=1}^{m} y^{(i)} = \sum_{i=1}^{m} h_\theta(x^{(i)}) + 2\lambda\theta_0$$

Where θ_0 is the parameter for the intercept. In general, we will not penalize this term, and in this case regularization will have no effect.