Approximating Class Posteriors with a Model Function Whose Parameters are Estimated via MLE for the 2-class Case

assume iid

Given dataset $D = \{(x_1, l_1), \ldots, (x_N, l_N)\}$

and model $h(x;\theta)$ to approximate $p(L=1|x)$, and $1-h(x;\theta)$ to approximate $p(L=0|x)$, use MLE to optimize θ with D.

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmax}}\ p(D|\theta) = \underset{\theta}{\operatorname{argmax}}\ \ln p(D|\theta)$$

$$= \underset{\theta}{\operatorname{argmax}}\ \ln\left[p(x_1, l_1, \ldots, x_N, l_N|\theta)\right]$$

Assume iid samples

$$= \underset{\theta}{\operatorname{argmax}}\ \ln\left[\prod_{n=1}^{N} p(x_n, l_n|\theta)\right]$$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \ln p(x_n, l_n|\theta)$$

Bayes Rule

$$= \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \ln\left[p(l_n|x_n, \theta)\, p(x_n|\theta)\right]$$

Assume $x_n \perp\!\!\!\perp \theta$

$$= \underset{\theta}{\operatorname{argmax}} \sum_{n=1}^{N} \ln p(l_n|x_n, \theta) + \sum_{n=1}^{N} \ln p(x_n|\cancel{\theta})$$

Assuming $\{x_1, \dots, x_N\}$ are independent from θ makes the second term $\sum_{n=1}^{N} \ln p(x_n|\theta) = \sum_{n=1}^{N} \ln p(x_n)$ and this term does <u>not</u> depend on θ anymore, so it is an additive constant. Therefore,

$$\hat{\theta}_{ML} = \arg\max_{\theta} \sum_{n=1}^{N} \ln p(\ell_n | x_n \theta)$$

Each $\ell_n \in \{0, 1\}$, takes one of 2 values so there are two cases to consider

if $\ell_n = 1$, $p(\ell_n | x_n \theta) \approx h(x_n; \theta)$

if $\ell_n = 0$, $p(\ell_n | x_n \theta) \approx 1 - h(x_n; \theta)$

Let $\ln p(\ell_n | x_n \theta) =$ $\ell_n \ln h(x_n; \theta) + (1-\ell_n) \ln (1 - h(x_n; \theta))$

Substituting this:

$$\hat{\theta}_{ML} = \arg\max_{\theta} \sum_{n=1}^{N} \Big[\underbrace{\ell_n \ln h(x_n; \theta)}_{\text{contributes for } x_n \text{ if } \ell_n = 1} + \underbrace{(1-\ell_n) \ln (1 - h(x_n; \theta))}_{\text{contributes for } x_n \text{ if } \ell_n = 0} \Big]$$

Multiplying with $-1/N$:

$$\hat{\theta}_{ML} = \underset{\theta}{\operatorname{argmin}} \frac{-1}{N} \sum_{n=1}^{N} \left[l_n \ln h(x_n;\theta) + (1-l_n)\ln(1-h(x_n;\theta)) \right]$$

Now let's consider logistic-generalized-linear models:

$$h(x;\cancel{\theta}\,w) = \frac{1}{1+e^{-w^T b(x)}}$$

where $b(x) = \begin{bmatrix} b_0(x) \\ b_1(x) \\ \vdots \\ b_M(x) \end{bmatrix}$, typically with $b_0(x)=1$

Logistic-linear model: $b(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}$ $\quad w = \begin{bmatrix} w_0 \\ w_x \end{bmatrix}$

e.g. $x \in \mathbb{R}^3 \Rightarrow b(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \end{bmatrix}$ $\quad w = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \\ w_3 \end{bmatrix}$

Logistic-quadratic-polynomial model: $b(x) = \begin{bmatrix} 1 \\ x \\ q(x) \end{bmatrix}$

all monic quadratic polynomials

e.g. $x \in \mathbb{R}^2 \Rightarrow b(x) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_1^2 \\ x_1 x_2 \\ x_2^2 \end{bmatrix}$

Choose $\dim w = \dim b(x)$