Cross Entropy, Regularized Logistic Regression

Submit a PDF of your answers to Canvas

1. Consider two distributions $\boldsymbol{p}$ and $\boldsymbol{q}$ over a joint support with $\boldsymbol{p} = [0.75\ \ 0.125\ \ 0.125\ \ 0]^T$ and $\boldsymbol{q} = [0.25\ \ 0.25\ \ 0.25\ \ 0.25]^T$.

Jensen

a) Compute the cross entropy $H(\boldsymbol{p},\boldsymbol{q})$ in bits. $= \sum_{x \in X} -p(x)\log_2 q(x) = 4$

b) Show that in general, $H(\boldsymbol{p},\boldsymbol{q}) \geq H(\boldsymbol{p})$. $H(p) - H(p,q) = \sum_{x \in X} p(x)\log_2 \frac{q(x)}{p(x)} \leq \log_2\left[\sum_{x \in X} \cancel{p(x)}\frac{q(x)}{\cancel{p(x)}}\right]$ $= \log_2\left[\sum_{x \in X} q(x)\right]$ ≤ 0

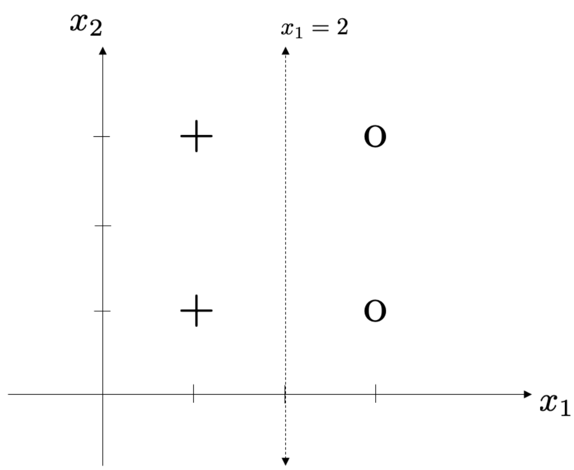
c) Under what conditions does $H(\boldsymbol{p},\boldsymbol{q}) = D(\boldsymbol{p}||\boldsymbol{q})$?
$H(p,q) = D(p||q) + H(p) \rightarrow$ they're equal if $H(P) = 0$

d) Compute the softmax of $\boldsymbol{p}$. Your answer should be a vector of length 4.

e) Compute the softmax of $\boldsymbol{q}$. Your answer should be a vector of length 4.

softmax$(p) = [0.39\ \ 0.21\ \ 0.21\ \ 0.18]$ softmax$(q) = (0.25\ \ 0.25\ \ 0.25\ \ 0.25)$

2. This problem continues a previous activity. Consider the four data points shown below. The data points at $(1,1)$ and $(1,3)$ belong to the class labeled $y = 1$, while the data points at $(3,1)$ and $(3,3)$ belong to the class labeled $y = -1$.



A decision boundary at $x_1 = 2$ can be expressed as the set of points that satisfy $\boldsymbol{x}^T\boldsymbol{w} = 0$, where

$$\boldsymbol{x} = \begin{bmatrix} 1 \\ x_1 \\ x_2 \end{bmatrix} \text{ and } \boldsymbol{w} = \begin{bmatrix} w_0 \\ w_1 \\ w_2 \end{bmatrix}$$

Last time we concluded that the set of points $\boldsymbol{x}^T\boldsymbol{w} = 0$ with

$$\boldsymbol{w} = \begin{bmatrix} w_0 \\ -\frac{w_0}{2} \\ 0 \end{bmatrix}$$

correspond to the vertical line at $x_1 = 2$. The starting point for motivating binary logistic regression is to assume that $y \in \{-1, 1\}$ is a Bernoulli random variable with bias that depends on $\boldsymbol{x}$:

$$p(y = 1|\boldsymbol{x}) = \frac{1}{1 + e^{-\boldsymbol{x}^T\boldsymbol{w}}}$$

and

$$p(y = -1|\boldsymbol{x}) = \frac{1}{1 + e^{\boldsymbol{x}^T\boldsymbol{w}}}$$

Last time we concluded the log-likelihood $\log L(\boldsymbol{w}) = \log p(y_1, y_2, y_3, y_4)$ given the data in the figure resulted in a negative log likelihood of

$$\text{nnl}(\boldsymbol{w}) = -\log L(\boldsymbol{w}) = 4\log\left(1 + e^{-w_0/2}\right).$$

Since the data was separable, this resulted in an unstable solution, which increased as $w_0 \to \infty$.

a) One way to ensure a unique solution is to add a regularizer term to the negative log likelihood. Find the solution to

$$\arg\min_{\boldsymbol{w}} \text{nll}(\boldsymbol{w}) + \frac{1}{10}||\boldsymbol{w}||_1$$

under the constraint that the decision boundary corresponds to the vertical line in the picture.

b) Sketch a picture of the corresponding logistic surface.

a) Here, we minimize $4\log(1+e^{-W_0/2}) + \frac{1}{10}||W||_1$

$= 4\log(1+e^{-W_0/2}) + \frac{1}{10}\left(\frac{3}{2}|W_0|\right)$

the derivitive suggests $W_0 = -2\ln\left(\frac{3}{37}\right) = -5.02$

b)