CS 189/289A Introduction to Machine Learning
Fall 2021 Jennifer Listgarten and Jitendra Malik

DIS3

1 Logistic Regression

Assume that we have n i.i.d. data points $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$, where each y_i is a binary label in $\{0, 1\}$. We model the posterior probability as a Bernoulli distribution and the probability for each class is the sigmoid function, i.e., $p(y|\mathbf{x}; \mathbf{w}) = q^y(1-q)^{1-y}$, where $q = s(\mathbf{w}^\top \mathbf{x})$ and $s(\zeta) = \frac{1}{1+e^{-\zeta}}$ is the sigmoid function. Write out the likelihood and log likelihood functions. Comment on whether it is possible to find a closed form maximum likelihood estimate of $\mathbf{w}$, and describe an alternate approach.

2 Initialization of Weights for Backpropagation

Assume a fully-connected 1-hidden-layer network. Denote the dimensionalities of the input, hidden, and output layers as $d^{(0)}$, $d^{(1)}$, and $d^{(2)}$. That is, the input (which we will denote with a superscript (0)) has dimensions $x_1^{(0)}, \ldots, x_{d^{(0)}}^{(0)}$. Let g denote the activation function applied at each layer. As defined in lecture, let $S_j^{(l)} = \sum_{i=1}^{d^{(l-1)}} w_{ij}^{(l)} x_i^{(l-1)}$ be the weighted input to node j in layer l, and let $\delta_j^{(l)} = \frac{\partial \ell}{\partial S_j^{(l)}}$ be the partial derivative of the final loss ℓ with respect to $S_j^{(l)}$.

Recall that backpropagation is simply an efficient method to compute the gradient of the loss function so we can use it with gradient descent. These methods require the parameters to be initialized to some value. In logistic regression we were able to initialize all weights as 0.

(a) Imagine that we initialize the values of our weights to be some constant w. After performing the forward pass, what is the value of $x_j^{(1)}$ in terms of the elements of $\{x_i^{(0)} : i = 1, \ldots, d^{(0)}\}$? What is the relationship between each $x_j^{(1)}$?

(b) After the backward pass of backpropagation, what is the relation between the members of the set $\{\delta_i^{(1)} : i = 1, \ldots, d^{(1)}\}$, assuming we have calculated $\{\delta_j^{(2)} : j = 1, \ldots, d^{(2)}\}$?

(c) For a reasonable loss function, can we say the same about each $\delta_i^{(2)}$?

(d) *Even though* $w_{ij}^{(2)}$ is different for each j (but for a fixed j, it is the same for each i), this pattern continues. Why?

(e) To solve this problem, we randomly initialize our weights. This is called symmetry breaking. Why are we able to set our weights to 0 for logistic regression?