Logistic Regression

CS 242

April 16, 2024

We are given training data $\mathbf{x}^{(1)} \dots \mathbf{x}^{(N)}$, with class labels $y_1 \dots y_N$. We will represent the class label as positive or negative one, which gives a simple form for the classifier's output probability that applies to both class labels.

$$\begin{aligned} y &\in \{-1, 1\} \\ \sigma(z) &= \frac{1}{1 + e^{-z}} \\ P(y = 1 \mid \mathbf{x}; \mathbf{w}) &= \sigma(\mathbf{w} \cdot \mathbf{x}) \\ P(y = -1 \mid \mathbf{x}; \mathbf{w}) &= 1 - \sigma(\mathbf{w} \cdot \mathbf{x}) \\ &= \sigma(-\mathbf{w} \cdot \mathbf{x}) \\ P(y \mid \mathbf{x}; \mathbf{w}) &= \sigma(y\, \mathbf{w} \cdot \mathbf{x}) \end{aligned}$$

The stochastic gradient descent algorithm for training consists of making a step in the direction of the data point $\mathbf{x}^{(n)}$, multiplied by the class label y_n, when we make an error, with a learning rate α.

For $n = 1 \dots N$

$$\mathbf{w} \leftarrow \mathbf{w} + \alpha \left(1 - P(y_n \mid \mathbf{x}^{(n)}; \mathbf{w})\right)\ y_n \mathbf{x}^{(n)}$$

The algorithm above is derived from taking the total likelihood of the training data. We aim to minimize loss L, which is defined as the negative log likelihood of the training data.

$$\begin{aligned} L(\mathbf{w}) &= -\log P(y_1 \dots y_N \mid \mathbf{x}^{(1)} \dots \mathbf{x}^{(N)}; \mathbf{w}) \\ &= -\log \prod_{n=1}^{N} P(y_n \mid \mathbf{x}^{(n)}; \mathbf{w}) \\ &= \sum_n -\log \sigma(y_n \mathbf{w} \cdot \mathbf{x}^{(n)}) \end{aligned}$$

This objective function above consists of a sum over examples in the train-

ing data of a term L_n corresponding to the nth data point.

$$L(\mathbf{w}) = \sum_n L_n$$
$$L_n = -\log \sigma(y_n \mathbf{w} \cdot \mathbf{x}^{(n)})$$

The SGD algorithm makes an update according to the gradient of one point's L_n.

$$\begin{aligned}\frac{\partial L_n}{\partial \mathbf{w}} &= \frac{\partial}{\partial \mathbf{w}} \left(-\log \sigma(y_n \mathbf{w} \cdot \mathbf{x}^{(n)}) \right) \\ &= \left(1 - \sigma(y_n \mathbf{w} \cdot \mathbf{x}^{(n)}) \right) \; y_n \mathbf{x}^{(n)}\end{aligned}$$