Assignment 3: Classification, Logistic Regression, and Gradient Descent

Machine Learning

Fall 2019

Learning Objectives

- Learn about the framing of the classification problem in machine learning.
- Learn about the logistic regression algorithm.
- Learn about gradient descent for optimization.
- Some C&E topic.

Prior Knowledge Utilized

- Supervised learning problem framing.
- Training / testing splits.

Recall: Supervised Learning Problem Setup

We are given a training set of datapoints, $(\mathbf{x_1}, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_n, y_n)$ where each $\mathbf{x_i}$ represents an element of an input space (e.g., a d-dimensional feature vector) and each y_i represents an element of an output space (e.g., a scalar target value). Our goal is to determine a function $\hat{f}$ that maps from the input space to the output space.

We assume there is a loss function, ℓ, that determines the amount of loss that a particular prediction $\hat{y}_i$ incurs due to a mismatch with the actual output y_i. The best possible model, $\hat{f}^\star$, is the one that minimizes these losses over the training set. This notion can be expressed with the following equation.

$$\hat{f}^\star = \arg\min_{\hat{f}} \sum_{i=1}^{n} \ell\left(\hat{f}(\mathbf{x_i}), y_i\right) \tag{1}$$

1 The Classification Problem

2 Perceptron?

TODO: this might fit better after discussing gradient descent. I can also see just punting is since it is probably not really that important.

3 Mathematical Foundations

3.1 Probability

3.2 Logistic function

The logistic function turns out to be very useful for modeling the probability that some event occurs. TODO.

Exercise 1

In this exercise you will be working to better understand some of the properties of the logistic function. Remember, the logistic function, σ, is defined as:

$$\sigma(x) = \frac{1}{1+e^{-x}} \ . \tag{2}$$

(a) Do some thought exercises on the logistic function. Limiting cases, etc. TODO.

(b) Show that $\sigma(-x) = 1 - \sigma(x)$.

☆ Solution

$$\begin{aligned}
\sigma(-x) &= \frac{1}{1+e^{x}} && (3)\\
&= \frac{e^{-x}}{e^{-x}+1} \quad \text{multiply by top and bottom by } e^{-x} && (4)\\
\sigma(-x) - 1 &= \frac{e^{-x}}{e^{-x}+1} - \frac{1+e^{-x}}{1+e^{-x}} \quad \text{subtract } -1 \text{ on both sides} && (5)\\
&= \frac{-1}{1+e^{-x}} && (6)\\
&= -\sigma(x) && (7)\\
\sigma(-x) &= 1 - \sigma(x) && (8)
\end{aligned}$$

(c) Show that the derivative of the logistic function $\frac{d}{dx}\sigma(x) = \sigma(x)(1-\sigma(x))$

☆ Solution

Two solutions for the price of 1!

Solution 1:

$$\frac{d}{dx}\sigma(x) = -e^{-x}\sigma(x)^2 \qquad \text{apply quotient rule} \qquad (9)$$

$$= \sigma(x)\left(\frac{-e^{-x}}{1+e^{-x}}\right) \qquad \text{expand out one of the } \sigma(x)\text{'s} \qquad (10)$$

$$= \sigma(x)\left(\frac{-1}{e^{x}+1}\right) \qquad \text{multiply top and bottom by } e^x \qquad (11)$$

$$= \sigma(x)(-\sigma(-x)) \qquad \text{substitute for } \sigma(-x) \qquad (12)$$

$$= \sigma(x)(\sigma(x)-1) \qquad \text{apply } \sigma(-x) = 1-\sigma(x) \qquad (13)$$

Solution 2:

$$\frac{d}{dx}\sigma(x) = \frac{-e^{-x}}{(1+e^{-x})^2} \qquad \text{apply quotient rule} \qquad (14)$$

$$= \frac{-e^{-x}}{1+2e^{-x}+e^{-2x}} \qquad \text{expand the bottom} \qquad (15)$$

$$= \frac{-1}{e^{x}+2+e^{-x}} \qquad \text{multiply top and bottom by } e^x \qquad (16)$$

$$= \frac{-1}{(1+e^{x})(1+e^{-x})} \qquad \text{factor} \qquad (17)$$

$$= -\sigma(x)\sigma(-x) \qquad \text{decompose using definition of } \sigma(x) \qquad (18)$$

$$= -\sigma(x)(1-\sigma(x)) \qquad \text{apply } \sigma(-x) = 1-\sigma(x) \qquad (19)$$

$$= \sigma(x)(\sigma(x)-1) \qquad \text{distribute the } -1 \qquad (20)$$

(d) The log odds of an event occurring is defined as

$$\ln\left(\frac{p(\text{event occurs})}{p(\text{event does not occur})}\right) = \ln\left(\frac{p(\text{event occurs})}{1-p(\text{event does occur})}\right) . \qquad (21)$$

If we assume that $p(\text{event occurs}) = \sigma(x)$, show that the log odds of the event occurring is equal to x.

☆ Solution

$$\ln\left(\frac{p(\text{event occurs})}{p(\text{event does not occur})}\right) = \ln\left(\frac{\sigma(x)}{1-\sigma(x)}\right) \tag{22}$$
$$= \ln\left(\frac{\sigma(x)}{\sigma(-x)}\right) \tag{23}$$
$$= \ln\left(\frac{1+e^x}{1+e^{-x}}\right) \tag{24}$$
$$= \ln\left(e^x \frac{1+e^x}{e^x(1+e^{-x})}\right) \tag{25}$$
$$= x + \ln\left(\frac{1+e^x}{e^x+1}\right) \tag{26}$$
$$= x \tag{27}$$

3.3 Log-loss

One of the components of our supervised learning problem framing is the loss function ℓ. Recall that this function takes as input the true output value, y, and a predicted output value, $\hat{y}$, and returns the loss that the model incurs for any potential mismatch between the values. When we were working with linear regression, we sought to minimize the sum of squared errors and consequently used the loss function $\ell(y, \hat{y}) = (y - \hat{y})^2$.

4 Top-down View of Logistic Regression

5 Gradient Descent

5.1 Chain Rule for Gradients

5.2 Visualization

6 Algorithm Derivation

Todo: this is easier with the identities of the derivative of a logistic function.

$$\mathbf{w}^\star = \arg\min_{\mathbf{w}} e(\mathbf{w}) \tag{28}$$

$$e(\mathbf{w}) = \sum_{i=1}^{n} y_i \log \frac{1}{1+e^{-\mathbf{w}^\top \mathbf{x_i}}} + (1-y_i) \log \frac{1}{1+e^{\mathbf{w}^\top \mathbf{x_i}}} \tag{29}$$

$$= \arg\min_{\mathbf{w}} \sum_{i=1}^{n} -y_i \log \left(1+e^{-\mathbf{w}^\top \mathbf{x_i}}\right) - (1-y_i) \log \left(1+e^{\mathbf{w}^\top \mathbf{x_i}}\right) \tag{30}$$

$$\nabla e(\mathbf{w}) = \sum_{i=1}^{n} \frac{y_i \mathbf{x_i}}{1+e^{-\mathbf{w}^\top \mathbf{x_i}}} - \frac{(1-y_i)\,\mathbf{x_i}}{1+e^{\mathbf{w}^\top \mathbf{x_i}}} \tag{31}$$

$$= \sum_{i=1}^{n} \mathbf{x_i} \left(\frac{y_i}{1+e^{-\mathbf{w}^\top \mathbf{x_i}}} - \frac{(1-y_i)}{1+e^{\mathbf{w}^\top \mathbf{x_i}}} \right) \tag{32}$$