Machine Learning and Data Mining I: Lecture 5
Logistic Regression

Morgan McCarty

11 July 2023

1 Classification

Logistic regression is a classification algorithm. It functions by estimating the parameters of a Bernoulli distribution. The Bernoulli distribution is a discrete distribution with two possible outcomes, 0 and 1. The probability of 1 is p and the probability of 0 is $1 - p$. The Bernoulli distribution is a special case of the binomial distribution where $n = 1$.

- Data = $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ (i.e. a data set which has a set of features and a label for each data point)
- $x_i \in \mathbb{R}^m$ (i.e. x_i is a vector of m features)
- $y_i \in V$ where V is a discrete set of values (i.e. y_i is a label)
- For Logistic Regression, $V = \{0, 1\}$

1.1 Goal

- Learn a function $f : \mathbb{R}^m \to V$ ($f : x \to y$) that maps the features to the labels.
- What is the form of y?
- What is the underlying model?

2 Discriminant Function

For each class $i \in V$, we have a discriminant function $f_i(x)$. Where in Logistic Regression, $V = \{0, 1\}$, we have two discriminant functions $f_0(x)$ and $f_1(x)$. $f_i(x) \to \mathbb{R}$.

2.1 Prediction Rule

$y = \operatorname{argmax}_{i \in V} f_i(x)$

- The predicted class corresponds to the discriminant function with the highest score.
- This is a winner-take-all approach (i.e. we will have discrete predictions).

3 How Logistic Regression Works

- Binary classification: $y \in \{0, 1\}$
- The discriminant function for positive class:
  - Modeled by the Logistic Function (a.k.a. Sigmoid Function: $\sigma(x) = \frac{1}{1+e^{-x}}$):

$$f(z) = \frac{1}{1 + e^{-z}} \tag{1}$$

  - Maps $\mathbb{R} \to [0, 1]$
  - i.e. $P(y = 1|x) = f(w^T x)$

3.1 Parameterize the Logistic Function

$$\begin{aligned} f_1(x) &= P(y = 1|x) \\ &= \frac{1}{1 + e^{-w^T x}} \\ \therefore f_0(x) &= P(y = 0|x) \\ &= 1 - f_1(x) \\ &= \frac{e^{-w^T x}}{1 + e^{-w^T x}} \end{aligned}$$

3.2 Goal

Given training data: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, learn w that fits the parameterized logistic function to the data.

$$\begin{aligned} \text{Predict } y = 1 &: \\ f(w^T x) &\geq 0.5 \\ &\Rightarrow w^T x \geq 0 \end{aligned}$$

Otherwise Predict: $y = 0$

3.3 Different vs. Perceptron

Logistic Regression and Perceptron are both linear classifiers, but Logistic Regression is a probabilistic model while Perceptron is not. Additionally Logistic Regression functions on the sigmoid function while Perceptron functions on the step function.

3.4 Decision Boundary

Logistic Regression learns a decision boundary that is a hyperplane. At the decision boundary $f_1(x) = f_0(x)$ (i.e. $P(y = 1|x) = P(y = 0|x)$, both 0 and 1 are equally likely labels).

$$
\begin{aligned}
P(y = 1|x, w) &= P(y = 0|x, w) \\
\frac{P(y = 1|x, w)}{P(y = 0|x, w)} &= 1 \\
\ln\left(\frac{P(y = 1|x, w)}{P(y = 0|x, w)}\right) &= 0 \\
\ln\left(\frac{1}{e^{-w^T x}}\right) &= 0 \\
w^T x &= 0
\end{aligned}
$$

3.4.1 Questons

- How can we learn the optimal paramaters?
- What is the cost function?

4 Maximum Likelihood Estimation

MLE is a method of estimating the parameters of a statistical model given observations.

- Data: $D = \{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$ (i.e. a data set which has a set of features and a label for each data point)
- $x_i \in \mathbb{R}^m$ (i.e. x_i is a vector of m features)
- $y_i \in V$ where V is a discrete set of values (i.e. y_i is a label)
- Again this the same as what we saw earlier so $V = \{0, 1\}$ for Logistic Regression

4.1 Goal

- Fit the logistic function to the training data

$$
\begin{aligned}
f(x) &= P(y = 1|x) \\
&= \frac{1}{1 + e^{-w^T x}}
\end{aligned}
$$

- Find w that maximizes the likelihood of the data (i.e. the probability of the data given the parameters).

4.2 Generative Models

- Model the data generation process (i.e. how the data was generated, rather than just the data itself).
- For Logistic Regression:

- Model the data (features + labels) as being generated by repeated Bernoulli trials.

$$f(k;p) = \begin{cases} p & \text{if } k = 1 \\ 1-p & \text{if } k = 0 \end{cases}$$
$$f(k;p) = p^k(1-p)^{1-k} \text{ for } k \in \{0,1\}$$

4.3 Data Likelihood Function

- Class conditional Probabilities
  - Labels: $y \in \{0,1\}$

$$\begin{aligned} P(y=1|x) &= \sigma(w^T x) \\ &= \frac{1}{1+e^{-w^T x}} \\ P(y=0|x) &= 1 - P(y=1|x) \end{aligned}$$

  - Where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the logistic function (a.k.a. sigmoid function).
- Likelihood for a single observation:

$$\begin{aligned} P(y|x,w) &= Bernoulli(y|\sigma(w^T x)) \\ &= \sigma(w^T x)^y (1-\sigma(w^T x))^{1-y} \end{aligned}$$

- Likelihood for the entire data set:

$$P(y|x,w) = \prod_{i=1}^{n} \sigma(w^T x_i)^{y_i} (1-\sigma(w^T x_i))^{1-y_i}$$

  - Intuitively, this is the probability of the data given the parameters. We multiply the probabilities of each data point together because we assume that the data points are independent.
- Log Likelihood:

$$\begin{aligned} LL &= \ln P(y|x,w) \\ &= \sum_{i=1}^{n} \ln \sigma(w^T x_i)^{y_i} (1-\sigma(w^T x_i))^{1-y_i} \\ &= \sum_{i=1}^{n} y_i \ln \sigma(w^T x_i) + (1-y_i) \ln(1-\sigma(w^T x_i)) \end{aligned}$$

- Maximize the log likelihood:

$$\begin{aligned} w^* &= \underset{w}{\operatorname{argmax}}\, LL \\ &= \underset{w}{\operatorname{argmax}} \sum_{i=1}^{n} y_i \ln \sigma(w^T x_i) + (1-y_i) \ln(1-\sigma(w^T x_i)) \end{aligned}$$

  - This is equivalent to minimizing the negative log likelihood.

$$E(w) = -LL \tag{2}$$

- This is the cost function for Logistic Regression.

- Derivative of the cost function:
  - We first need to find the derivative of the sigmoid function.

$$\begin{aligned}
\sigma(x) &= \frac{1}{1+e^{-x}} \\
&= (1+e^{-x})^{-1} \\
&= (1+e^{-x})^{-2} \cdot e^{-x} \\
&= \sigma(x)^2 \cdot e^{-x} \\
\therefore \sigma'(x) &= \sigma(x)^2 \cdot e^{-x} \\
&= \sigma(x)(1-\sigma(x))
\end{aligned}$$

  - Therefore $\sigma'(w^T x) = \sigma(w^T x)(1 - \sigma(w^T x))$
  - Now we can find the derivative of the cost function.

$$\frac{\partial E(w)}{\partial w_j} = \Sigma_{i=1}^{N}(\sigma w^T x_i - y_i)x_{ij}$$

- It is important to note that there is no closed form solution for w^*. As such we have to use an iterative method to find w^* (e.g. gradient descent).

- Maximizing the log likelihood is equivalent to minimizing the negative log likelihood. So the negative log likelihood is the cost function for Logistic Regression.

$$E(w) = -\Sigma_{i=1}^{N}(y_i \ln \sigma(w^T x_i) + (1-y_i)\ln(1-\sigma(w^T x_i))) \tag{3}$$

4.4 Gradient Descent

- Initialize w to a zero vector. $w = [0, 0, \ldots, 0]^T$

- While the cost function is not minimized (i.e. $E(w) > 0$, not converged to zero):

$$\begin{aligned}
&w^k \leftarrow w^{k-1} - \eta \nabla E(w^{k-1}) \\
&w^k \leftarrow w^{k-1} - \eta X^T(0 - Y)
\end{aligned}$$

- Convergence criteria:
  - $E(w^{k-1}) - E(w^k) < \epsilon$
  - The reduction in the cost function is less than some threshold ϵ.
  - Or we have reached a maximum number of iterations.

4.5 Shortcomings

One of the biggest shortcomings of Logistic Regression is the possibility of overfitting, especially when the data is linearly separable.

4.5.1 Regularization

- We can use regularization to prevent overfitting.
- This is the same as what we used in Linear Regression.
- We add a regularization term to the cost function.

$$E(w) = -\Sigma_{i=1}^{N}(y_i \ln \sigma(w^T x_i) + (1 - y_i) \ln(1 - \sigma(w^T x_i))) + \frac{\lambda}{2} w^T w \tag{4}$$

  - Where λ is the regularization parameter.
  - λ is a hyperparameter of the model.
  - $\lambda \geq 0$

4.6 Cross Entropy Loss

- The cost function for Logistic Regression is also known as the cross entropy loss function.
- It can be done for either label: 0 or 1.