CSC 411: Lecture 04: Logistic Regression

Raquel Urtasun & Rich Zemel

University of Toronto

Sep 23, 2015

Today

- Key Concepts:
  - Logistic Regression
  - Regularization
  - Cross validation

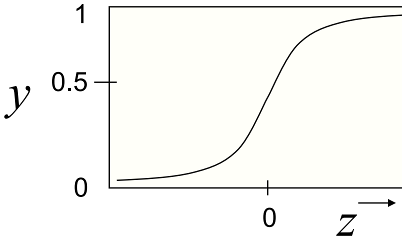
Logistic Regression

- An alternative: replace the $sign(\cdot)$ with the sigmoid or logistic function
- We assumed a particular functional form: sigmoid applied to a linear function of the data

$$y(\mathbf{x}) = \sigma\left(\mathbf{w}^T\mathbf{x} + w_0\right)$$

where the sigmoid is defined as

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$



- The output is a smooth function of the inputs and the weights

Logistic Regression

- We assumed a particular functional form: sigmoid applied to a linear function of the data

$$y(\mathbf{x}) = \sigma\left(\mathbf{w}^T\mathbf{x} + w_0\right)$$

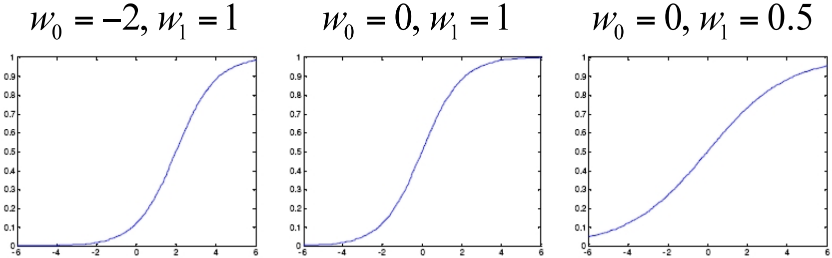
where the sigmoid is defined as

$$\sigma(z) = \frac{1}{1 + \exp(-z)}$$

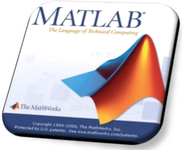
  - One parameter per data dimension (feature)
  - Features can be discrete or continuous
  - Output of the model: value $y \in [0, 1]$
  - This allows for gradient-based learning of the parameters: smoothed version of the $sign(\cdot)$

- Let's look at how modifying **w** changes the function shape
- 1D example:

$$y = \sigma\left(w_1 x + w_0\right)$$



- Demo



Probabilistic Interpretation

- If we have a value between 0 and 1, let's use it to model the posterior

$$p(C=0|\mathbf{x}) = \sigma(\mathbf{w}^T\mathbf{x} + w_0) \quad \text{with} \quad \sigma(z) = \frac{1}{1+\exp(-z)}$$

- Substituting we have

$$p(C=0|\mathbf{x}) = \frac{1}{1+\exp\left(-\mathbf{w}^T\mathbf{x} - w_0\right)}$$

- Supposed we have two classes, how can I compute $p(C=1|\mathbf{x})$?
- Use the marginalization property of probability

$$p(C=1|\mathbf{x}) + p(C=0|\mathbf{x}) = 1$$

- Thus (show matlab)

$$p(C=1|\mathbf{x}) = 1 - \frac{1}{1+\exp\left(-\mathbf{w}^T\mathbf{x} - w_0\right)} = \frac{\exp(-\mathbf{w}^T\mathbf{x} - w_0)}{1+\exp\left(-\mathbf{w}^T\mathbf{x} - w_0\right)}$$

Conditional likelihood

- Assume $t \in \{0, 1\}$, we can write the probability distribution of each of our training points $p(t^{(1)}, \cdots, t^{(N)} | \mathbf{x}^{(1)}, \cdots \mathbf{x}^{(N)})$
- Assuming that the training examples are sampled IID: independent and identically distributed

$$p(t^{(1)}, \cdots, t^{(N)} | \mathbf{x}^{(1)}, \cdots \mathbf{x}^{(N)}) = \prod_{i=1}^{N} p(t^{(i)} | \mathbf{x}^{(i)})$$

- We can write each probability as

$$\begin{aligned} p(t^{(i)} | \mathbf{x}^{(i)}) &= p(C = 1 | \mathbf{x}^{(i)})^{t^{(i)}} p(C = 0 | \mathbf{x}^{(i)})^{1 - t^{(i)}} \\ &= \left(1 - p(C = 0 | \mathbf{x}^{(i)})\right)^{t^{(i)}} p(C = 0 | \mathbf{x}^{(i)})^{1 - t^{(i)}} \end{aligned}$$

- We might want to learn the model, by maximizing the conditional likelihood

$$\max_{\mathbf{w}} \prod_{i=1}^{N} p(t^{(i)} | \mathbf{x}^{(i)})$$

- Convert this into a minimization so that we can write the loss function

Loss Function

$$p(t^{(1)}, \cdots, t^{(N)} | \mathbf{x}^{(1)}, \cdots \mathbf{x}^{(N)}) = \prod_{i=1}^{N} p(t^{(i)} | \mathbf{x}^{(i)})$$
$$= \prod_{i=1}^{N} \left(1 - p(C = 0 | \mathbf{x}^{(i)})\right)^{t^{(i)}} p(C = 0 | \mathbf{x}^{(i)})^{1 - t^{(i)}}$$

- It's convenient to take the logarithm and convert the maximization into minimization by changing the sign

$$\ell_{log}(\mathbf{w}) = -\sum_{i=1}^{N} t^{(i)} \log(1 - p(C = 0 | \mathbf{x}^{(i)}, \mathbf{w})) - \sum_{i=1}^{N} (1 - t^{(i)}) \log p(C = 0 | \mathbf{x}^{(i)}, \mathbf{w})$$

- Why is this equivalent to maximize the conditional likelihood?
- Is there a closed form solution?
- It's a convex function of $\mathbf{w}$. Can we get the global optimum?

Gradient Descent

$$\min_{\mathbf{w}} \ell(\mathbf{w}) = \min_{\mathbf{w}} \left\{ -\sum_{i=1}^{N} t^{(i)} \log(1 - p(C = 0|\mathbf{x}^{(i)}, \mathbf{w})) - \sum_{i=1}^{N} (1 - t^{(i)}) \log p(C = 0|\mathbf{x}^{(i)}, \mathbf{w}) \right\}$$

- Gradient descent: iterate and at each iteration compute steepest direction towards optimum, move in that direction, step-size λ

$$w_j^{(t+1)} \leftarrow w_j^{(t)} - \lambda \frac{\partial \ell(\mathbf{w})}{\partial w_j}$$

- But where is $\mathbf{w}$?

$$p(C = 0|\mathbf{x}) = \frac{1}{1 + \exp\left(-\mathbf{w}^T \mathbf{x} - w_0\right)} \qquad p(C = 1|\mathbf{x}) = \frac{\exp(-\mathbf{w}^T \mathbf{x} - w_0)}{1 + \exp\left(-\mathbf{w}^T \mathbf{x} - w_0\right)}$$

- You can write this in vector form

$$\bigtriangledown \ell(\mathbf{w}) = \left[\frac{\partial \ell(\mathbf{w})}{\partial w_0}, \cdots, \frac{\partial \ell(\mathbf{w})}{\partial w_k} \right]^T, \qquad \text{and} \qquad \triangle(\mathbf{w}) = -\lambda \bigtriangledown \ell(\mathbf{w})$$

Let's look at the updates

- The log likelihood is

$$\ell_{log-loss}(\mathbf{w}) = -\sum_{i=1}^{N} t^{(i)} \log p(C=1|\mathbf{x}^{(i)}, \mathbf{w}) - \sum_{i=1}^{N} (1-t^{(i)}) \log p(C=0|\mathbf{x}^{(i)}, \mathbf{w})$$

where the probabilities are

$$p(C=0|\mathbf{x}, \mathbf{w}) = \frac{1}{1+\exp(-z)} \qquad p(C=1|\mathbf{x}, \mathbf{w}) = \frac{\exp(-z)}{1+\exp(-z)}$$

and $z = \mathbf{w}^T\mathbf{x} + w_0$

- We can simplify

$$\begin{aligned}
\ell(\mathbf{w}) &= \sum_i t^{(i)} \log(1+\exp(-z^{(i)})) + \sum_i t^{(i)} z^{(i)} + \sum_i (1-t^{(i)}) \log(1+\exp(-z^{(i)})) \\
&= \sum_i \log(1+\exp(-z^{(i)})) + \sum_i t^{(i)} z^{(i)}
\end{aligned}$$

- Now it's easy to take derivatives

Updates

$$\ell(\mathbf{w}) = \sum_i t^{(i)} z^{(i)} + \sum_i \log(1 + \exp(-z^{(i)}))$$

- Now it's easy to take derivatives
- Remember $z = \mathbf{w}^T \mathbf{x} + w_0$

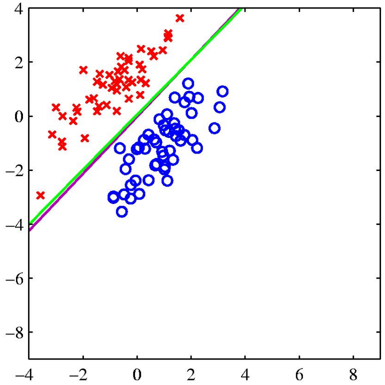
$$\frac{\partial \ell}{\partial w_j} = \sum_i t^{(i)} x_j^{(i)} - x_j^{(i)} \cdot \frac{\exp(-z^{(i)})}{1 + \exp(-z^{(i)})}$$

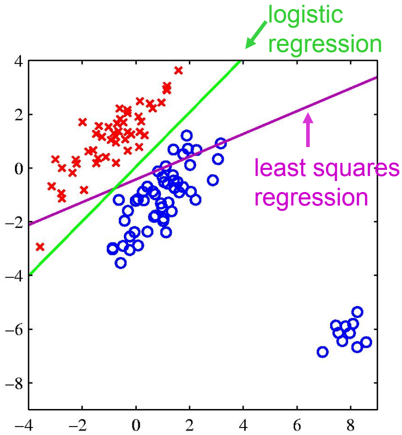
- What's $x_j^{(i)}$?
- And simplifying

$$\frac{\partial \ell}{\partial w_j} = \sum_i x_j^{(i)} \left(t^{(i)} - p(C = 1 | \mathbf{x}^{(i)}) \right)$$

- Don't get confused with indexes: j for the weight that we are updating and i for the training example
- Logistic regression has linear decision boundary

Logistic regression vs least squares





If the right answer is 1 and the model says 1.5, it loses, so it changes the boundary to avoid being "too correct" (tilts away from outliers)

33

Regularization

- We can also look at

$$p(\mathbf{w}|\{t\}, \{\mathbf{x}\}) \propto p(\{t\}|\{\mathbf{x}\}, \mathbf{w})\, p(\mathbf{w})$$

  with $\{t\} = (t^{(1)}, \cdots, t^{(N)})$, and $\{\mathbf{x}\} = (\mathbf{x}^{(1)}, \cdots, \mathbf{x}^{(N)})$

- We can define priors on parameters $\mathbf{w}$
- This is a form of regularization
- Helps avoid large weights and overfitting

$$\max_{\mathbf{w}} \log \left[p(\mathbf{w}) \prod_i p(t^{(i)}|\mathbf{x}^{(i)}, \mathbf{w}) \right]$$

- What's $p(\mathbf{w})$?

Regularized Logistic Regression

- For example, define prior: normal distribution, zero mean and identity covariance $p(\mathbf{w}) = \mathcal{N}(0, \alpha \mathbf{I})$
- This prior pushes parameters towards zero
- Including this prior the new gradient is

$$w_j^{(t+1)} \leftarrow w_j^{(t)} - \lambda \frac{\partial \ell(\mathbf{w})}{\partial w_j} - \lambda \alpha w_j^{(t)}$$

  where t here refers to iteration of the gradient descent
- How do we decide the best value of α?

Use of Validation Set

- We can divide the set of training examples into two disjoint sets: training and validation
- Use the first set (i.e., training) to estimate the weights **w** for different values of α
- Use the second set (i.e., validation) to estimate the best α, by evaluating how well the classifier does in this second set
- This test how well you generalized to unseen data
- The parameter α is the importance of the regularization, and it's a hyper-parameter

Cross-Validation

- Leave-p-out cross-validation:
  - We use p observations as the validation set and the remaining observations as the training set.
  - This is repeated on all ways to cut the original training set.
  - It requires $\mathcal{C}_n^p$ for a set of n examples
- Leave-1-out cross-validation: When $p = 1$, does not have this problem
- k-fold cross-validation:
  - The training set is randomly partitioned into k equal size subsamples.
  - Of the k subsamples, a single subsample is retained as the validation data for testing the model, and the remaining $k - 1$ subsamples are used as training data.
  - The cross-validation process is then repeated k times (the folds).
  - The k results from the folds can then be averaged (or otherwise combined) to produce a single estimation