Discussion 4
Logistic Regression and Coordinate Descent
Machine Learning, Spring 2019

1 Interpretation of logistic regression

Given a dataset $\{\mathbf{x}_i, y_i\}_{i=1}^{n}$, $y_i \in \{\pm 1\}$ the logistic regression model is defined by

$$p(y \mid \mathbf{x};\, \boldsymbol{\theta}) = \sigma(y\boldsymbol{\theta}^\top \mathbf{x}),$$

where σ is the logistic sigmoid function defined by

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

The *log odds* of $y = 1$ conditioned on $\mathbf{x}$ is defined as

$$\log \frac{p(+1 \mid \mathbf{x};\, \boldsymbol{\theta})}{p(-1 \mid \mathbf{x};\, \boldsymbol{\theta})}.$$

(a) Prove that the log odds is equal to the simple expression $\boldsymbol{\theta}^\top \mathbf{x}$.

(b) In light of (a), give an interpretation for each $\boldsymbol{\theta}_i$. For example, if we increase the i-th component of $\mathbf{x}$ while holding the others constant, what effect does this have on the log odds.

2 On the loss of logistic function

After reading a paper on a logistic regression students in ML class divided in two groups: Alpha and Beta.

Alpha group claims that given dataset $\{\mathbf{x}_i, y_i\}_{i=1}^n$, the loss for logistic regression (parametrized by weight vector θ) is

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\left[-y_i\theta^\top \mathbf{x}_i + \log\left(1 + e^{\theta^\top \mathbf{x}_i}\right)\right]$$

while Beta group is sure that the loss is

$$\mathcal{L}(\theta) = \frac{1}{n}\sum_{i=1}^{n}\log\left(1 + e^{-y_i\theta^\top \mathbf{x}_i}\right)$$

Help students to find a correct solution. Which loss is correct and why?

Consider $y \in \{0, 1\}$**.**

- Step 1. Write down $\mathbf{P}(y_i = a|\theta, \mathbf{x}_i)$, where a is all possible values that y takes.

- Step 2. Write down the cross-entropy loss (negative log-likelihood).

Consider $y \in \{-1, 1\}$**.**

- Step 3. Write down $\mathbf{P}(y_i = a|\theta, \mathbf{x}_i)$, where a is all possible values that y takes.

- Step 4. Write down the cross-entropy loss (negative log-likelihood).

2.1 Which loss is correct?

3 ℓ_2-regularization and Coordinate Descent

Consider a dataset $\mathbf{x}_i \in \mathbb{R}^D$, $y_i \in \{-1, 1\}$, $i = 1, ..., N$ and a negative log-likelihood of logistic regression

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{N} \sum_{i=1}^{N} \log \left(1 + e^{-y_i \boldsymbol{\theta}^\top \mathbf{x}_i}\right)$$

We add an ℓ_2 regularization on parameter $\boldsymbol{\theta}$, and therefore get the regularized loss function, $\lambda > 0$

$$\tilde{\mathcal{L}}(\boldsymbol{\theta}) = \mathcal{L}(\boldsymbol{\theta}) + \frac{\lambda}{2} \boldsymbol{\theta}^\top \boldsymbol{\theta}$$

1. Given that $\mathcal{L}(\boldsymbol{\theta})$ is a convex function prove that $\tilde{\mathcal{L}}(\boldsymbol{\theta})$ is also convex.

   Useful lemmas and definitions

   (a) A function $f : R^n \to R$ is convex if dom f is a convex set and if for all x_1, $x_2 \in$ dom f, and α with $0 \leq \alpha \leq 1$, we have $f(\alpha x_1 + (1 - \alpha) x_2) \leq \alpha f(x_1) + (1 - \alpha) f(x_2)$. f is concave if $-f$ is convex.

   (b) A continuous, twice differentiable function of several variables f is convex if and only if its Hessian matrix of second partial derivatives is positive semidefinite on the interior of the convex set $\nabla f \succeq 0$.

   (c) If f and g are convex functions, then function $h(x) = f(x) + g(x)$ is also convex.

2. Suppose that our data is linearly separable, therefore there exists some parameter vector $\boldsymbol{\theta}_*$ such that $y_i \boldsymbol{\theta}_*^\top \mathbf{x}_i \geq \delta > 0$ for all i, where $\delta > 0$. Prove that there exists some sequence $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2, \boldsymbol{\theta}_3, \ldots$ such that

   $$\lim_{i \to \infty} \mathcal{L}(\boldsymbol{\theta}_i) = 0.$$

   Hint: consider a sequence of $k\boldsymbol{\theta}_*$, where k is some scalar

3. What does 2 imply about the range of parameter $\boldsymbol{\theta}$ when loss is minimized? Given that what is the advantage of a ℓ_2-regularization for logistic regression?

4. Suppose we use coordinate descent to optimize the regularized loss function above. Is there a closed-form update for the j-th coordinate of $\boldsymbol{\theta}$?

5. Now let's try coordinate descent on the quadratic approximation to the objective function. Derive the quadratic approximation by Taylor expansion, and then try to find a closed-form update for the j-th coordinate of $\boldsymbol{\theta}$.

   (a) Step 1. Denote $p(\mathbf{x}) = \left(1 + e^{-\boldsymbol{\theta}^\top \mathbf{x}}\right)^{-1}$, thus
   if $y = 1$, $\left(1 + e^{-y_i \boldsymbol{\theta}^\top \mathbf{x}}\right)^{-1} = p(\mathbf{x})$; if $y = -1$, $\left(1 + e^{-y_i \boldsymbol{\theta}^\top \mathbf{x}}\right)^{-1} = 1 - p(\mathbf{x})$.

   (b) Step 2. Perform Taylor expansion of $\mathcal{L}(\boldsymbol{\theta})$ up to second degree

   (c) Step 3. Denote $w_i = p(\mathbf{x}_i)(1 - p(\mathbf{x}_i))$, $z_i = \frac{(y_i+1)/2 - p(\mathbf{x}_i)}{w_i}$

   (d) Step 4. Using the quadratic approximation of $\mathcal{L}(\boldsymbol{\theta})$ derive the closed-form of best step-size for the regularized loss.