Regularized Logistic Regression is Strictly Convex

Jason D. M. Rennie
jrennie@csail.mit.edu

January 9, 2005

Abstract

We show that Logistic Regression and Softmax are convex.

1 Binary LR

Let $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$, $\vec{x}_i \in \mathbb{R}^d$, be a set of examples. Let $\vec{y} = \{y_1, \ldots, y_n\}$, $y_i \in \{-1, +1\}$, be a corresponding set of labels. Logistic Regression learns parameters[1] $\vec{w} \in \mathbb{R}^d$ so as to minimize

$$-\log P(\vec{y}|X, \vec{w}) = \sum_{i=1}^{n} \log\left(1 + \exp(-y_i \vec{w}^T \vec{x}_i)\right). \tag{1}$$

To show that the LR objective is convex, we consider the partial derivatives. Define $g(z) = \frac{1}{1+e^{-z}}$. Note that $1 - g(z) = \frac{e^{-z}}{1+e^{-z}}$ and $\frac{\partial g(z)}{\partial z} = -g(z)(1 - g(z))$.

$$\frac{\partial \log P(\vec{y}|X, \vec{w})}{\partial w_j} = -\sum_{i=1}^{n} y_i x_{ij}(1 - g(y_i \vec{w}^T \vec{x}_i)) \tag{2}$$

$$\frac{\partial^2 \log P(\vec{y}|X, \vec{w})}{\partial w_j \partial w_k} = \sum_{i=1}^{n} y_i^2 x_{ij} x_{ik} g(y_i \vec{w}^T \vec{x}_i)(1 - g(y_i \vec{w}^T \vec{x}_i)) \tag{3}$$

To show that the objective is convex, we first show that the Hessian (the matrix of second derivatives) is positive semi-definite (PSD). A matrix, M, is PSD iff $\vec{a}^T M \vec{a} \geq 0$ for all vectors $\vec{a}$. Let ∇^2 be the Hessian for our objective. Define $P_i := g(y_i \vec{w}^T \vec{x}_i)(1 - g(y_i \vec{w}^T \vec{x}_i))$ and $\rho_{ij} = x_{ij}\sqrt{P_i}$. Then,

$$\vec{a}^T \nabla^2 \vec{a} = \sum_{i=1}^{n} \sum_{j=1}^{d} \sum_{k=1}^{d} a_j a_k x_{ij} x_{ik} P_i, \tag{4}$$

$$= \sum_{i=1}^{n} \vec{a}^T \vec{\rho}_i \vec{\rho}_i^T \vec{a} \geq 0, \tag{5}$$

[1] In terms of the two vector formulation, $\vec{w} = W_+ - W_-$.

Note that $\vec{a}^T \vec{\rho}_i \vec{\rho}_i^T \vec{a} = (\vec{a}^T \vec{\rho}_i)^2 \geq 0$. Hence, the Hessian is PSD. Theorem 2.6.1 of Cover and Thomas (1991) gives us that an objective with a PSD Hessian is convex. If we add an L2 regularizer, $C\vec{w}^T\vec{w}$, to the objective, then the Hessian is positive definite and hence the objective is strictly convex.

2 Two-weight LR

Let $X = \{\vec{x}_1, \ldots, \vec{x}_n\}$, $\vec{x}_i \in \mathbb{R}^d$, be a set of examples. Let $\vec{y} = \{y_1, \ldots, y_n\}$, $y_i \in \{-1, +1\}$, be a corresponding set of labels. Logistic Regression learns parameters $W_- \in \mathbb{R}^d$ and $W_+ \in \mathbb{R}^d$ so as to minimize

$$-\log P(\vec{y}|X, W) = \sum_{i=1}^{n} \log\left(\exp(W_+\vec{x}_i) + \exp(W_-\vec{x}_i)\right) - \sum_{i=1}^{n} W_{y_i}\vec{x}_i. \quad (6)$$

To show that the LR objective is convex, we consider the partial derivatives. Define $Z_i := \exp(W_+\vec{x}_i) + \exp(W_-\vec{x}_i)$. Define $P_{+i} = \exp(W_+\vec{x}_i)/Z_i$ and $P_{-i} = \exp(W_-\vec{x}_i)/Z_i$.

$$\frac{\partial \log P(\vec{y}|X, W)}{\partial W_{uj}} = \sum_{i=1}^{n} x_{ij} P_u - \sum_{i|y_i=u} x_{ij} \quad (7)$$

$$\frac{\partial^2 \log P(\vec{y}|X, W)}{\partial W_{uj} \partial W_{vk}} = \delta_{u=v} \sum_{i=1}^{n} x_{ij} x_{ik} P_{ui} - \sum_{i=1}^{n} x_{ij} x_{ik} P_{ui} P_{vi} \quad (8)$$

$$= (-1)^{\delta_{u=v}+1} \sum_{i=1}^{n} x_{ij} x_{ik} P_{+i} P_{-i} \quad (9)$$

To show that the objective is convex, we first show that the Hessian (the matrix of second derivatives) is positive semi-definite (PSD). A matrix, M, is PSD iff $\vec{a}^T M \vec{a} \geq 0$ for all vectors $\vec{a}$. Let ∇^2 be the Hessian for our objective. Define $\rho_{iuj} := u x_{ij} \sqrt{P_{+i} P_{-i}}$.

$$\vec{a}^T \nabla^2 \vec{a} = \sum_{i=1}^{n} \sum_{j,k,u,v} (-1)^{\delta_{j=u}+1} a_{uj} a_{vk} x_{ij} x_{ik} P_{+i} P_{-i} \quad (10)$$

$$= \sum_{i=1}^{n} \vec{a}^T \vec{\rho}_i \vec{\rho}_i^T \vec{a} \geq 0, \quad (11)$$

Note that $\vec{a}^T \vec{\rho}_i \vec{\rho}_i^T \vec{a} = (\vec{a}^T \vec{\rho}_i)^2 \geq 0$. Hence, the Hessian is PSD. Theorem 2.6.1 of Cover and Thomas (1991) gives us that an objective with a PSD Hessian is convex. If we add an L2 regularizer, $C(W_- W_-^T + W_+ W_+^T)$, to the objective, then the Hessian is positive definite and hence the objective is strictly convex.

Note that we abuse notation by collapsing two indices into a single vector, e.g. $\vec{a} = (a_{-1}, a_{-2}, \ldots, a_{-d}, a_{+1}, \ldots, a_{+d})$. Similar for ρ.

3 Softmax

Next, we show that the multiclass generalization of LR, commonly known as "softmax," is convex. Let $\vec{y} = \{y_1, \ldots, y_n\}$, $y_i \in \{1, \ldots, m\}$, be the set of multi-class labels. Softmax learns parameters $W \in \mathbb{R}^{m \times d}$ so as to minimize

$$-\log P(\vec{y}|X, W) = \sum_{i=1}^{n} \left[\log \left(\sum_{u=1}^{m} \exp(W_u \vec{x}_i) \right) - W_{y_i} \vec{x}_i \right]. \tag{12}$$

We use W_u (W_{y_i}) to denote the u^{th} ($y_i{}^{\text{th}}$) row of W. To show that the Softmax objective is convex, we consider the the partial derivatives. Define $Z_i = \sum_{u=1}^{m} \exp(W_u \vec{x}_i)$ and $P_{iu} = \exp(W_u \vec{x}_i)/Z_i$. Note that

$$\frac{\partial P_{iu}}{\partial W_{vk}} = x_{ik} P_{iu} \left[\delta_{u=v}(1 - P_{iu}) - \delta_{u \neq v} P_{iv} \right]. \tag{13}$$

$$\frac{\partial \log P(\vec{y}|X, W)}{\partial W_{uj}} = \sum_{i=1}^{n} x_{ij} P_{iu} - \sum_{i|y_i=u} x_{ij} \tag{14}$$

$$\frac{\partial^2 \log P(\vec{y}|X, W)}{\partial W_{uj} \partial W_{vk}} = \sum_{i=1}^{n} x_{ij} x_{ik} P_{iu} \left[\delta_{u=v}(1 - P_{iu}) - \delta_{u \neq v} P_{iv} \right] \tag{15}$$

By the Diagonal Dominance Theorem (see the Appendix), the Hessian (the matrix of second derivatives) is positive semi-definite (PSD). Theorem 2.6.1 of Cover and Thomas (1991) gives us that an objective with a PSD Hessian is convex. If we add an L2 regularizer, $C \sum_u W_u W_u^T$, to the objective, then the Hessian is positive definite and hence the objective is strictly convex.

Appendix

Theorem 1 (Diagonal Dominance Theorem) *Suppose that M is symmetric and that for each $i = 1, \ldots, n$, we have*

$$M_{ii} \geq \sum_{j \neq i} |M_{ij}|. \tag{16}$$

Then M is positive semi-definite (PSD). Furthermore, if the inequalities above are all strict, then M is positive definite.

Proof: Recall that an eigenvector is a vector $\vec{x}$ such that $M\vec{x} = \gamma\vec{x}$. γ is called the eigenvalue for $\vec{x}$. Let $M \in \mathbb{R}^{n \times n}$ be a symmetric matrix. Then M has n eigenvectors with real eigenvalues. Consider an eigenvector, $\vec{x}$, of M with eigenvalue γ. Then, $M\vec{x} = \gamma\vec{x}$. In particular, $M_{ii}x_i + \sum_{j \neq i} M_{ij}x_j = \gamma x_i$. Let i be such that $|x_i| \geq |x_j|\ \forall j$. Now, assume $M_{ii} \geq \sum_{j \neq i} |M_{ij}|\ \forall i$. Then we see that $\gamma \geq 0$. Hence, all eigenvalues of M are non-negative and M is PSD. If the inequalities in our assumption are strict, then eigenvalues of M are positive and M is positive definite. $\square$

References

Cover, T., & Thomas, J. (1991). *Elements of information theory*. John Wiley & Sons, Inc.