Notes: Multivariate Classification with SGD

1 References

1. The textbook covers multivariate classification in Chapter 17.
2. `https://chrisyeh96.github.io/2018/06/11/logistic-regression.html`

2 Multivariate classification background

The softmax function is defined to be

$$\text{softmax} : \mathbb{R}^d \to \mathbb{R}^d \tag{1}$$

$$\text{softmax}(\mathbf{a})_i = \frac{\exp(\mathbf{a}_i)}{\sum_{j=1}^{d} \exp(\mathbf{a}_j)} \tag{2}$$

the softmax function has the important property that it "squashes" its input vector into a probability distribution.

The loss function used in multi-class classification is called the "softmax cross entropy" loss, and is the result of composing the softmax function above with the cross entropy (which is a way to measure the difference of probability distributions).

Assume we are solving a multi-class classification problem with k classes and d input dimensions. The softmax cross entropy loss is given by

$$\ell(W; (\mathbf{x}, y)) = -\log \frac{\exp(-\mathbf{w}_y^T \mathbf{x})}{\sum_{j=1}^{k} \exp(-\mathbf{w}_j^T \mathbf{x})} \tag{3}$$

where for each class $i \in [k]$, $\mathbf{w}_i : \mathbb{R}^d$ is the parameter vector associated with class i; the variable $W : \mathbb{R}^{k \times d} = (\mathbf{w}_1; \mathbf{w}_2; ...; \mathbf{w}_k)$ is the full parameter matrix; $\mathbf{x} : \mathbb{R}^d$ is the feature vector; and $y \in [k]$ is the class label.

3 Loss function properties

Fact 1. The softmax cross entropy function ℓ is convex with respect to W.

Fact 2. Assume that $\|\mathbf{x}\|_2 \leq \rho$. Then softmax cross entropy function ℓ is ρ-Lipschitz with respect to W.

Fact 3. For each class $i \in [k]$, assume that $\|\mathbf{w}_i\|_2 \leq B$. Then $\|W\|_F \leq \sqrt{k}B$.

Theorem 14.8 of Shalev-Shwartz and Ben-David then states that if SGD is run for T iterations to compute parameter estimate $\bar{W}$, then

$$\mathbb{E}\, L_S(\bar{W}) - L_S(W^*) \leq \frac{\sqrt{k}B\rho}{\sqrt{T}} \tag{4}$$

where $W^* = \arg\min L_S(W)$.

Theorem 14.12 of Shalev-Shwartz and Ben-David then states that if SGD is run for T iterations to compute parameter estimate $\bar{W}$, then

$$\mathbb{E}\, L_D(\bar{W}) - L_D(W^*) \leq \frac{\sqrt{k}B\rho}{\sqrt{T}} \tag{5}$$

where $W^* = \arg\min L_D(W)$.

Page intentionally left blank for handwritten notes.

4 L2 regularization

Recall that when we apply L2 regularization, we are minimizing the objective function

$$f(W) = \tfrac{1}{m}\sum_{i=1}^{m} \ell(W; (\mathbf{x}_i, y_i)) + \tfrac{\lambda}{2}\|W\|_F^2. \tag{6}$$

This function is λ-strongly convex in W. (Why?)

Then Theorem 14.11 tells us that after performing T iterations of SGD, we have that

$$\mathbb{E}\, f(\bar{W}) - f(W^*) \leq \frac{4\rho^2}{\lambda T}(1 + \log T) \tag{7}$$

where $W^* = \arg\min f(W)$. Notice that there is no dependence on k in this inequality!

If we substitute $f(W) = L_D(W) + \|W\|_F^2$ into Equation 8 above, then we get

$$\mathbb{E}\, L_D(\bar{W}) - L_D(W^*) \leq \frac{4\rho^2}{\lambda T}(1 + \log T) + \tfrac{\lambda}{2}\|W^*\|_F^2 - \tfrac{\lambda}{2}\|\bar{W}\|_F^2 \leq \frac{4\rho^2}{\lambda T}(1 + \log T) + \frac{\lambda k B^2}{2}. \tag{8}$$

5 What to do when k is big?

A common way to reduce sample complexity when k is large is to factor the parameter matrix as $W = VU$, where $V : \mathbb{R}^{k\times e}$, $U : \mathbb{R}^{e\times d}$, and $e << k$. Then, each $\mathbf{w}_i = \mathbf{v}_i U$, and the cross entropy softmax loss is

$$\ell(VU; (\mathbf{x}, y)) = -\log \frac{\exp(-\mathbf{v}_y^T U\mathbf{x})}{\sum_{j=1}^k \exp(-\mathbf{v}_j^T U\mathbf{x})}. \tag{9}$$

Fact 4. The softmax cross entropy function ℓ is convex with respect to U and V.

Fact 5. Assume that $\|\mathbf{x}\|_2 \leq \rho$. Then softmax cross entropy function ℓ is ρ-Lipschitz with respect to U and V.

Fact 6. (i) For each class $i \in [k]$, assume that $\|\mathbf{v}_i\|_2 \leq 1$. Then, $\|V\|_F \leq \sqrt{k}$. (ii) Let $\mathbf{u}_i$ denote the ith column of U. For each column $i \in [e]$, assume that $\|\mathbf{u}_i\|_2 \leq B$. Then $\|U\|_F \leq \sqrt{e}B$.

Assume that V is fixed and known in advance. Theorem 14.12 of Shalev-Shwartz and Ben-David then states that if SGD is run for T iterations to compute parameter estimate $\bar{U}$, then

$$\mathbb{E}\, L_D(V\bar{U}) - L_D(VU^*) \leq \frac{\sqrt{e}B\rho}{\sqrt{T}} \tag{10}$$

where $U^* = \arg\min_U L_D(VU)$. (A similar result holds for L_S based on Theorem 14.8.)

Assume that U is fixed and known in advance. Theorem 14.12 of Shalev-Shwartz and Ben-David then states that if SGD is run for T iterations to compute parameter estimate $\bar{U}$, then

$$\mathbb{E}\, L_D(\bar{V}U) - L_D(V^*U) \leq \frac{\sqrt{k}\rho}{\sqrt{T}} \tag{11}$$

where $V^* = \arg\min_V L_D(VU)$. (A similar result holds for L_S based on Theorem 14.8.)

6 Open Research Problem (very informal discussion)

The limitation of the previous technique is that the class labels must have a linear structure. In this section, we review how to take advantage of arbitrary metric structure.

Let $\mathcal{L}$ be a metric space of labels, and $d_{i,j}$ be the distance between class labels i and j.

Build a "cover tree" from the class labels. Let p_i denote the parent label for class i. Then we can rewrite that parameter vector for each class i as

$$\mathbf{w}_i = \mathbf{v}_i + \mathbf{w}_{p_i}. \tag{12}$$

Then the goal is to learn the $\mathbf{v}_i$ vectors instead of the $\mathbf{w}_i$ vectors. The matrix $V = (\mathbf{v}_1, ..., \mathbf{v}_k)$ can be bounded to have size $\sqrt{\dim(\mathcal{L})}B$.

It then follows from Theorem 14.12 of Shalev-Shwartz and Ben-David that if SGD is run for T iterations to compute parameter estimate $\bar{W}$, then

$$\mathbb{E}\, L_D(\bar{W}) - L_D(W^*) \le \frac{\sqrt{\dim \mathcal{L}} B\rho}{\sqrt{T}} \tag{13}$$

where $W^* = \arg\min L_D(W)$.