EECS 553: Machine Learning (ECE) **University of Michigan**

Multiclass Classification

Winter 2023 **Clayton Scott**

1 Introduction

In this write-up,

- $\mathcal{X}$ denotes the sample space,
- $k \geq 2$ denotes the number of classes,
- $[k] := \{1, 2, \ldots, k\}$,
- Score function $\mathbf{f} = (f_1, \ldots, f_k) : \mathcal{X} \to \mathbb{R}^k$,
- Given a vector $\mathbf{v} = (v_1, \ldots, v_k) \in \mathbb{R}^k$, define

$$\max \mathbf{v} := \max_{j \in [k]} v_j \quad \text{and} \quad \arg\max \mathbf{v} := \min\{i \in [k] : v_i = \max \mathbf{v}\}.$$

- Define $\arg\max \mathbf{f} : \mathcal{X} \to [k]$ by

$$\mathcal{X} \ni x \mapsto \arg\max \mathbf{f}(x) \in [k].$$

Below, suppose that the sample space $\mathcal{X} = \mathbb{R}^d$.

Definition 1. *Given a matrix $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{d \times k}$, we define the score function $\mathbf{f}_\mathbf{W}$ by $\mathbf{f}_\mathbf{W}(x) = (\mathbf{w}_1' x, \ldots, \mathbf{w}_k' x)' = \mathbf{W}' x$.*

Definition 2. *Let $(x_i, y_i) \in \mathbb{R}^d \times [k]$ for $i = 1, \ldots, n$. The dataset $\{(x_i, y_i)\}_{i \in [n]}$ is* linearly separable *if there exists $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{d \times k}$ such that for all $i \in [n]$, we have $\mathbf{w}_{y_i}' x_i > \mathbf{w}_j' x_i$ for all $j \neq y_i$.*

Exercise 1. A dataset $\{(x_i, y_i)\}_{i \in [n]}$ is *linearly separable* if and only if there exists $\mathbf{W} = (\mathbf{w}_1, \ldots, \mathbf{w}_k) \in \mathbb{R}^{d \times k}$ such that for all $i \in [n]$, $(\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \geq 1$ for all $j \in [k]$, $j \neq y_i$.

2 Multiclass support vector machine

Multiclass support vector machine is a subject of ongoing research with many different variants proposed. See [1]. We will cover two variants by [2] and by [3].

2.1 Hard-margin

In this section, we assume the following.

$$\textbf{Assumption}\text{: the dataset } \{(x_i, y_i)\}_{i \in [n]} \text{ is linearly separable.} \tag{1}$$

The *multiclass hard-margin support vector machine* solves the following optimization:

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times k}} \quad \frac{1}{2} \|\mathbf{W}\|_F^2 \tag{2}$$

$$s.t. \quad (\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \geq 1, \tag{3}$$

$$\forall i \in [n], \forall j \in [k] : j \neq y_i. \tag{4}$$

Note that this is an optimization problem with convex (in fact quadratic) objective and linear constraints. Furthermore, by Exercise 1, the problem is feasible.

Exercise 2. When $k = 2$, i.e., $y_i \in \{1, 2\}$ for all i, show that a solution of Eqn. (2) can be transformed into a solution of the optimization

$$\min_{\mathbf{u} \in \mathbb{R}^d} \quad \|\mathbf{u}\|_2^2 \tag{5}$$

$$s.t. \quad (-1)^{y_i} \mathbf{u}' x_i \geq 1, \, \forall i \in [n]. \tag{6}$$

Note that Eqn. (5) is the *binary* hard-margin support vector machine.

2.2 Soft-margin

By Exercise 1, the optimization (2) is feasible if and only if the dataset is separable. However, we cannot assume that the dataset is separable in practice, i.e., the constraint (3) may be violated for all choices of $\mathbf{W}$.

To remedy this, we introduce *slack variables* ζ_i, for $i = 1, \ldots, n$.

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times k}, \zeta_i \in \mathbb{R}, \, i=1,\ldots,n} \quad \frac{\lambda}{2} \|\mathbf{W}\|_F^2 + \frac{1}{n} \sum_{i=1}^{n} \zeta_i \tag{7}$$

$$s.t. \quad (\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \geq 1 - \zeta_i, \tag{8}$$

$$\zeta_i \geq 0, \tag{9}$$

$$\forall i \in [n], \, \forall j \in [k] : j \neq y_i. \tag{10}$$

The slack variable is constrained to be nonnegative. When (3) is violated for some $i \in [n]$, the slack variable ζ_i can be increased until (8) holds. The quantity λ is a hyperparameter. The optimization (7) is known as the *Crammer-Singer* multiclass support vector machine [2].

2.3 Empiricial risk minimization formulation

Rearranging, we can write (8) as $\zeta_i \geq 1 - (\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i$. For any $t \in \mathbb{R}$, define $[t]_+ := \max\{0, t\}$. The constraint (8) and the nonnegative constraint (10) can be combined into a single constraint:

$$\zeta_i \geq \max \left\{ 0, \max_{j \in [k]: j \neq y_i} 1 - (\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \right\} \tag{11}$$

$$= \left[\max_{j \in [k]: j \neq y_i} 1 - (\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \right]_+ \tag{12}$$

$$= \max_{j \in [k]: j \neq y_i} \left[1 - (\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \right]_+ . \tag{13}$$

Exercise 3. Prove the equality (13).

Recall the binary *hinge loss* $h(t) = [1 - t]_+$ (see Figure 1)
Continuing where we left off at (13), we have

$$\zeta_i \geq \max_{j \in [k]: j \neq y_i} h\left((\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \right).$$

Thus, the soft-margin multiclass SVM (7) can be reformulated as

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times k}} \quad \frac{\lambda}{2} \|\mathbf{W}\|_F^2 + \frac{1}{n} \sum_{i=1}^{n} \max_{j \in [k]: j \neq y_i} h\left((\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \right). \tag{14}$$

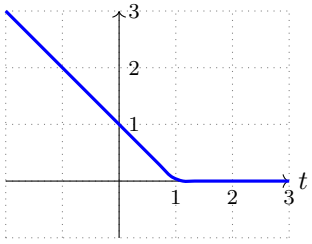

Figure 1: Hinge loss function

Exercise 4. For any $(x, y) \in \mathbb{R}^d \times [k]$, prove that the function

$$\mathbf{W} \mapsto \max_{j \in [k]: j \neq y} h\left((\mathbf{w}_y - \mathbf{w}_j)' x\right)$$

is convex. This function is often referred to as the *Crammer-Singer* hinge loss.

Thus, (14) expresses the multiclass support vector machine as a regularized empirical risk minimization with respect to a convex loss.

Exercise 5. There is another way to introduce slack variables to the hard-margin multiclass SVM. Define slack variables ζ_{ij}, for $i = 1, \ldots, n$, and $j \in [k]$ where $j \neq y_i$.

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times k}, \zeta_{ij} \in \mathbb{R},\, i=1,\ldots,n,\, j \in [k],\, j \neq y_i} \quad \frac{\lambda}{2}\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n} \sum_{j \in [k]: j \neq y_i} \zeta_{ij} \tag{15}$$

$$s.t. \quad (\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i \geq 1 - \zeta_{ij}, \tag{16}$$

$$\zeta_{ij} \geq 0, \tag{17}$$

$$\forall i \in [n],\, \forall j \in [k] : j \neq y_i. \tag{18}$$

The above is known as the Weston-Watkins multiclass support vector machine:
Prove the following:

1. The optimization (15) is equivalent to

$$\min_{\mathbf{W} \in \mathbb{R}^{d \times k}} \quad \frac{\lambda}{2}\|\mathbf{W}\|_F^2 + \frac{1}{n}\sum_{i=1}^{n} \sum_{j \in [k]: j \neq y_i} h\left((\mathbf{w}_{y_i} - \mathbf{w}_j)' x_i\right). \tag{19}$$

2. For any $(x, y) \in \mathbb{R}^d \times [k]$, the function $\mathbf{W} \mapsto \sum_{j \in [k]: j \neq y} h\left((\mathbf{w}_y - \mathbf{w}_j)' x\right)$ is convex. This function is often referred to as the *Weston-Watkins* hinge loss.

3 Multinomial logistic regression

We start off with some notations:

- P denotes a joint distribution over $\mathcal{X} \times [k]$,
- $(X, Y) \sim P$ are random variables,
- (x_i, y_i) are i.i.d realizations of (X, Y), where $i = 1, \ldots, n$,
- $\eta_y(x) := P(Y = y | X = x)$ is the conditional distribution over $[k]$ given sample x,
- Define $\Delta_k = \{\boldsymbol{p} = (p_1, \ldots, p_k) \in [0, 1]^k : p_1 + \cdots + p_k = 1\}$.

3.1 Softmax

At the heart of multinomial logistic regression is the *softmax* function $\boldsymbol{\psi} = (\psi_1, \ldots, \psi_k) : \mathbb{R}^k \to \Delta_k$, which converts a real vector-valued score $\boldsymbol{v} = (v_1, \ldots, v_k)$ to a probability vector $\boldsymbol{\psi}(\boldsymbol{v}) \in \Delta_k$. The softmax is defined as

$$\psi_y(v) := \frac{\exp(v_y)}{\sum_{j\in[k]} \exp(v_j)} = \frac{1}{1 + \sum_{j\in[k]:j\neq y} \exp(-(v_y - v_i))}. \tag{20}$$

Exercise 6. Prove the following:

1. For all $\boldsymbol{v} \in \mathbb{R}^k$ we have $\psi_y(\boldsymbol{v}) > 0$ for all $y \in [k]$. Furthermore,

$$\psi_y(\boldsymbol{v}) = \left(1 + \sum_{j\in[k]:j\neq y} \exp(-(v_y - v_j))\right)^{-1}. \tag{21}$$

   Let $\text{sigmoid}(t) = (1 + \exp(-t))^{-1}$. Conclude that when $k = 2$, $\psi_1((v_1, v_2)) = \text{sigmoid}(v_1 - v_2)$ and $\psi_2((v_1, v_2)) = \text{sigmoid}(v_2 - v_1)$.

2. Let $\boldsymbol{p} \in \Delta_k$ be such that $p_y > 0$ for all $y \in [k]$. Define $\boldsymbol{v} = (v_1, \ldots, v_k)$ by $v_y = \ln(p_y)$ for each $y \in [k]$. Then $\boldsymbol{\psi}(\boldsymbol{v}) = \boldsymbol{p}$.

3. If $\boldsymbol{p} \in \Delta_k$ and $\boldsymbol{v} \in \mathbb{R}^k$ are such that $\boldsymbol{\psi}(\boldsymbol{v}) = \boldsymbol{p}$. Then there exists $c \in \mathbb{R}$ such that $v_y = \ln(p_y) + c$ for each $y \in [k]$. Hint: let $c = \ln\left(\sum_{j\in[k]} \exp(v_j)\right)$.

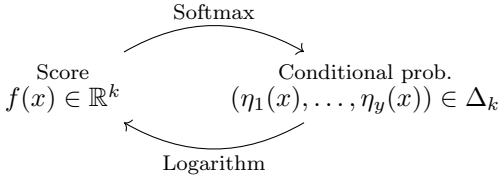

Figure 2: Softmax and logarithm are inverses of each other. See Exercise 6.

3.2 Generative assumption

Recall that $\eta_y(x) := P(Y = y|X = x)$.

$$\begin{aligned} &\textbf{Assumption: } \text{there exists a } \mathbf{W}^* \in \mathbb{R}^{d\times k} \text{ such that} \\ &\qquad P(Y = y|X = x) = \eta_y(x) = \psi_y(\mathbf{f}_{\mathbf{W}^*}(x)). \end{aligned} \tag{22}$$

In contrast with Assumption 1, the assumption here is on P, the data generating distribution.

We would like to find $\mathbf{W}^*$ based on data generated from the model. A natural approach is using *maximum likelihood*. The negative log-likelihood of observing the dataset $(x_1, y_1), \ldots, (x_n, y_n)$ given $\mathbf{W}$ is

$$\begin{aligned} &-\ln\left(\prod_{i=1}^n P(Y = y_i|X = x_i)\right) \\ &= \sum_{i=1}^n -\ln\left(\psi_{y_i}(\mathbf{f}_{\mathbf{W}}(x_i))\right) \\ &= \sum_{i=1}^n \ln\left(1 + \sum_{j\in[k]:j\neq y_i} \exp(-(\boldsymbol{w}_{y_i} - \boldsymbol{w}_i)'x)\right) \quad \because \text{equation (21)}. \end{aligned}$$

Adding a regularizer $\rho(\mathbf{W})$ (e.g., $= \frac{\lambda}{2}\|\boldsymbol{W}\|_F^2$ for Tikhonov regularization), we have

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times k}} \quad \rho(\boldsymbol{W}) + \frac{1}{n}\sum_{i=1}^{n} \ln\left(1 + \sum_{j\in[k]:j\neq y_i} \exp(-(\boldsymbol{w}_{y_i} - \boldsymbol{w}_j)' x_i)\right). \tag{23}$$

Note the similarity of (23) to the multiclass SVMs (19) and (14).

Exercise 7. 1. For any $y \in [k]$, show that the function

$$\mathbb{R}^k \ni \boldsymbol{v} \mapsto -\ln(\psi_y(\boldsymbol{v}))$$

is convex. Hint: use Exercise 6.

2. Conclude that, for any $(x, y) \in \mathbb{R}^d \times [k]$, the function

$$\mathbf{W} \mapsto -\ln(\psi_y(\boldsymbol{f}_{\boldsymbol{W}}(x)))$$

is convex.

The function

$$L(y, \boldsymbol{v}) := -\ln(\psi_y(\boldsymbol{v})) \tag{24}$$

is known as the *multinomial logistic loss*. Using this notation, we can write (23) more succinctly as

$$\min_{\mathbf{W}\in\mathbb{R}^{d\times k}} \quad \rho(\boldsymbol{W}) + \frac{1}{n}\sum_{i=1}^{n} L(y_i, \boldsymbol{W}' x_i). \tag{25}$$

3.3 Connection to cross-entropy

The *cross-entropy* is a "loss" that is popular in training neural networks. In this section, we show the relationship between cross-entropy and the multinomial logistic loss (24).

Definition 3. *The* cross-entropy *between two probability vectors $\boldsymbol{p}, \boldsymbol{q} \in \Delta_k$ is defined as the quantity*

$$\mathrm{CE}(\boldsymbol{q}, \boldsymbol{p}) = \sum_{j=1}^{k} -q_j \ln p_j.$$

Definition 4. *For each $y \in [k]$, define $\boldsymbol{e}_y$ to be the y-th elementary basis vector, i.e.,*

$$\boldsymbol{e}_y = (0, \ldots, 0, \underbrace{1}_{y\text{-}th\ index}, 0, \ldots, 0).$$

Let $\boldsymbol{f}_\theta : \mathcal{X} \to \mathbb{R}^k$ be a score function parametrized by $\theta \in \Theta$, where Θ is the *parameter space*. For example, $\Theta = \mathbb{R}^{d\times k}$ in the case of linear classifiers or Θ is the space of the parameters of a neural network. At a high level, neural networks are trained by minimizing the (regularized) training error

$$\min_{\theta\in\Theta} \quad \rho(\theta) + \frac{1}{n}\sum_{i=1}^{n} \mathrm{CE}(\boldsymbol{e}_{y_i}, \boldsymbol{\psi}(\boldsymbol{f}_\theta(x_i))) \tag{26}$$

where ρ is some (optional) regularizer.

Now, given a labelled data $x, y \in \mathcal{X} \times [k]$, note that

$$\begin{aligned}
\mathrm{CE}(\boldsymbol{e}_y, \boldsymbol{\psi}(\boldsymbol{f}_\theta(x))) &= -\sum_{j=1}^{k} -(\boldsymbol{e}_y)_j \ln \psi_j(\boldsymbol{f}_\theta(x)) \\
&= -\ln \psi_y(\boldsymbol{f}_\theta(x)) \quad \because (\boldsymbol{e}_y)_j = \begin{cases} 0 & : y \neq j \\ 1 & : y = j \end{cases} \\
&= L(y, \boldsymbol{f}_\theta(x))
\end{aligned}$$

where L is the multinomial logistic loss as defined in (24). Thus, (26) can be rewritten as

$$\min_{\theta \in \Theta} \quad \rho(\theta) + \frac{1}{n} \sum_{i=1}^{n} L(y_i, \boldsymbol{f}_\theta(x_i)). \tag{27}$$

Note the similarity to (25).

References

[1] Urun Dogan, Tobias Glasmachers, and Christian Igel, "A unified view on multi-class support vector classification" *Journal of Machine Learning Research*, vol. 17,pp. 1550–1831, 2006.

[2] Koby Crammer and Yoram Singer, "On the algorithmic implementation of multiclass kernel-based vector machines," *Journal of Machine Learning Research*, vol. 2, pp. 265–292, 2001.

[3] J Weston and C Watkins, "Support Vector Machines for Multi-Class Pattern Recognition," *Proc. 7th European Symposium on Artificial Neural Networks*, 1999.