6.883: Online Methods in Machine Learning

Alexander Rakhlin

LECTURE 5

This lecture is partly based on chapter 17 in [SSBD14].

1. MULTICLASS AND MULTILABEL PROBLEMS

Today we will talk about an extension of binary classification to multi-class and multi-label cases. In multiclass problems, examples are of the form (x_i, y_i) with $y_i \in \{1, \dots, k\}$. That is, there are k classes, and one of them is "correct" for the given example. In the multi-label setup, each x_i may be thought as belonging simultaneously to several classes. This is summarized by a vector $y_i \in \{0,1\}^k$, with $y_i(j) = 1$ if x_i is labeled as belonging to class j. For instance, in classifying hand-written digits, we might be interested in a multi-class formulation, while classifying a topic of a news article naturally leads to a multi-label setup.

1.1 Multiclass

Recall that for binary classification we studied linear classifiers $\langle w, x\rangle$ and the hinge loss

$$\ell(w, (x, y)) = \max\{0, 1 - y\langle w, x\rangle\}.$$

Should we change the form of the classifier, the loss function, or both?

The main issue is that linear classifiers $\langle w, x\rangle$ are naturally suited to binary problems, not to multi-class. Two standard approaches (one-vs-all and all-pairs) use binary classification as a subroutine. One-vs-all trains a collection of k binary classifiers. The all-pairs approach requires even more computational power, as it tries to discriminate between each pair of classes and then combine this information in some way to produce a single class label. Both of these methods disregard the multiclass nature of the problem and try to reduce it to binary classification.

Another approach is to change the form of the classifier and the loss function. Observe that $\text{sign}(\langle w, x\rangle)$ can be written as

$$\operatorname*{argmax}_{y \in \{\pm 1\}} \langle w, yx\rangle .$$

Let's think of yx as a transformation of x that aligns well with some w^* if y is the "correct" class for x, and does not align well otherwise. This formulation generalizes to multi-class quite naturally. Take $\Psi(x, y)$ to be some mapping, and consider

$$\operatorname*{argmax}_{y \in \{1,\dots,k\}} \langle w, \Psi(x, y)\rangle \tag{1}$$

to be a prediction of class for the given x. It would be nice if we could find Ψ that aligns well with some w^* when x is of class y and does not align well otherwise.

An even more general formulation is to replace $\langle w, \Psi(x, y)\rangle$ with some potentially non-linear *score* function $s(x, y)$. We will only study the linear score function, and, in addition, start with a linear representation for Ψ. The next section is devoted to this scenario.

1.1.1 Linear multiclass formulation with hinge loss

If $x \in \mathbb{R}^d$, think of Ψ as a $k \times d$ vector $\Psi(x, y) = [\dots 0 \dots, x^\intercal, \dots 0 \dots]^\intercal$ composed of zeros on all d-length segments, except the y-th, and let w be a $k \times d$-dimensional vector. Alternatively, we may think of w as a $k \times d$ matrix W, and (1) is written as

$$\underset{y \in \{1,\dots,k\}}{\operatorname{argmax}} \ \langle W_y, x \rangle \,, \tag{2}$$

the largest product of x and a row of the matrix W.

In binary classification, the predictors were associated with hyperplanes; now they are associated with $k \times d$ matrices W. Let

$$\widehat{y} = \underset{i \in \{1,\dots,k\}}{\operatorname{argmax}} \ \langle W_i, x \rangle \tag{3}$$

be the multiclass prediction, given a matrix W. For an example (x, y), the indicator loss is, as before,

$$\mathbf{1}\{\widehat{y} \neq y\} \leq \max_j \{\mathbf{1}\{j \neq y\} + \langle W_j, x \rangle - \langle W_y, x \rangle\} \,. \tag{4}$$

Let's convince ourselves of this fact. If $\widehat{y} = y$, the left-hand side is zero while the right-hand side is at least zero (verify this by taking $j = \widehat{y}$). Same argument holds for the case $\widehat{y} \neq y$. The loss function

$$\boldsymbol{\ell}(W, (x, y)) = \max_j \{\mathbf{1}\{j \neq y\} + \langle W_j, x \rangle - \langle W_y, x \rangle\} \tag{5}$$

will be called *multi-class hinge* loss.

Both the left and the right-hand side of (4) are zero if

$$1 + \langle W_j, x \rangle \leq \langle W_y, x \rangle$$

for any $j \neq y$. That is, no mistake is incurred if the correct class has a margin of 1 with respect to all other classes. For a class j to (erroneously) become the winner in the max, the associated product $\langle W_j, x \rangle$ needs to be at least $\langle W_y, x \rangle - 1$.

Verify that the binary hinge loss is a special case by taking W to be a $2 \times d$ matrix with rows $W_1 = -W_2 = \frac{1}{2} w$. (convince yourself of the rest of the argument)

Given data $(x_1, y_1), \dots, (x_n, y_n)$ with values in $\mathbb{R}^d \times \{1, \dots, k\}$, we aim to minimize the average of multi-class hinge losses

$$f(W) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}(W, (x_i, y_i)), \tag{6}$$

or the multi-class SVM version

$$f(W) = \frac{1}{n} \sum_{i=1}^{n} \boldsymbol{\ell}(W, (x_i, y_i)) + \frac{\lambda}{2} \|W\|^2. \tag{7}$$

The norm in the regularization term is the Frobenius norm (which is the Euclidean norm of the matrix stretched into a vector form).

To define SGD for the multiclass SVM it remains to find subgradients of each component loss. By our earlier argument, to find an element of the subdifferential set $\partial \boldsymbol{\ell}(W, (x_i, y_i))$,

we only need to find a subdifferential of the function for j^* that achieves the maximum in (5). That is, if

$$j^* \in \operatorname*{argmax}_j \{\mathbf{1}\{j \neq y_i\} + \langle W_j, x_i\rangle - \langle W_{y_i}, x_i\rangle\}, \tag{8}$$

then a subgradient (with respect to W) is computed as follows. If $j^* \neq y_i$

$$\nabla_i = [\dots 0 \dots, x_i, \dots, 0 \dots, -x_i, \dots 0]^\mathsf{T}, \tag{9}$$

a matrix whose j^*-th row is x_i and y_i-th row is $-x_i$. If $j^* = y_i$, the subgradient ∇_i is identically zero.

Algorithm 1 SGD for multiclass SVM

Input: $\lambda > 0$ (regularization parameter)
Init: $W_1 = 0$
for t=1,...,T **do**
  Set $\eta_t = \frac{1}{\lambda t}$
  Sample $i \sim \text{Unif}[n]$
  Find $j^* \in \operatorname*{argmax}_j \{\mathbf{1}\{j \neq y_i\} + \langle W_j, x_i\rangle - \langle W_{y_i}, x_i\rangle\}$
  if $j^* \neq y_i$ **then**
    $W_{t+1} = W_t - \eta_t \nabla_i - \eta_t \lambda W_t$
  else
    $W_{t+1} = W_t - \eta_t \lambda W_t$
  end if
end for

1.1.2 General formulation

We have considered a linear form for the function Ψ, and we now come back to the more general definition (1). Let $\Delta(y, y') \in [0, 1]$ be the cost of predicting y when the true label is y', with $\Delta(y, y) = 0$. Suppose the prediction $\widehat{y}$ is given by (1). The analogue of (4) for a general *score* function $s(x, y)$

$$\Delta(\widehat{y}, y) \leq \max_j \{\Delta(j, y) + s(x, j) - s(x, y)\}. \tag{10}$$

Multiclass hinge loss is sometimes written as a further upper bound on the right-hand side:

$$\max_j \{\Delta(j, y) + s(x, j) - s(x, y)\} = \max_{j \neq y} \max\{0, \Delta(j, y) + s(x, j) - s(x, y)\} \tag{11}$$

$$\leq \sum_{j \neq y} \max\{0, \Delta(j, y) + s(x, j) - s(x, y)\} \tag{12}$$

For the linear case with indicator loss, this version of multiclass hinge becomes

$$\sum_{j \neq y} \max\{0, 1 + \langle W_j, x\rangle - \langle W_y, x\rangle\} \tag{13}$$

Homework: derive the SGD update for this form of the loss.

References

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014.