Derive Multi-class SVM from Logistic Regression

Chunpai Wang

February 25, 2018

1 Model From Logistic Regression View

More specifically, we can consider a log-linear model, which has a parameter $w \in \mathbb{R}^n$ that defines a distribution over m labels for a given input $x \in X$ as follows

$$p_w(y|x) := \frac{\exp(w^\top \phi(x,y))}{\sum_{y'=1}^{m} \exp(w^\top \phi(x,y'))} \quad \forall y \in [m] \tag{1}$$

note: using this class probability is equivalent to using a logistic regression model.

Recall the logistic regression in binary classification,

$$\begin{aligned} h_w(x) &= \frac{1}{1+\exp(-w^\top x)} \\ 1 - h_w(x) &= 1 - \frac{1}{1+\exp(-w^\top x)} = \frac{\exp(-w^\top x)}{1+\exp(-w^\top x)} = \frac{1}{1+\exp(w^\top x)} \end{aligned} \tag{2}$$

and the main property of logistic loss function is

$$\begin{aligned} g(z) &= \frac{1}{1+\exp(-z)} \\ g(-z) &= \frac{1}{1+\exp(z)} = \frac{1}{1+\frac{1}{\exp(-z)}} = \frac{1}{\frac{1+\exp(-z)}{\exp(-z)}} = \frac{\exp(-z)}{1+\exp(-z)} \\ g(z) + g(-z) &= 1 \end{aligned} \tag{3}$$

2 Standard SVM

Given the training data $S = \{(x_1, y_1), ..., (x_n, y_n)\}$,and the class label $y_i \in \{1, -1\}$. Now, we consider the probability for the class label that are proportional to the exponential of a linear function of the data

$$\begin{aligned} p(y_i = 1|w, x_i) &\propto \exp(w^\top x_i) \\ p(y_i = -1|w, x_i) &\propto \exp(-w^\top x_i) \end{aligned} \tag{4}$$

where are ignore the bias term for simplicity.

Now, we can find the optimal w by maximizing the log-likelihood, or equivalently minimizing the negative log-likelihood

$$\min_w - \log \prod_i p(y_i|w, x_i) = \min_w - \sum_i \log p(y_i|w, x_i) \tag{5}$$

Since the optimal solution w is not necessary unique, since we can rescale the w and the optimal parameters may be unbounded. To yield a unique solution and avoid over-fitting, we typically add a penalty on the ℓ_2-norm of the parameter vector and compute the penalized maximum likelihood estimate

$$\arg\min_w - \sum_i \log p(y_i|w, x_i) + \lambda \|w\|_2^2 \tag{6}$$

With the estimate w, we can predict the label with

$$\hat{y} = \begin{cases} 1 & if \;\; p(y_i = 1|w, x_i) > p(y_i = -1|w, x_i) \\ -1 & if \;\; p(y_i = 1|w, x_i) < p(y_i = -1|w, x_i) \end{cases} \tag{7}$$

If we want the training set S is correctly trained, we can further generalize above to

$$\forall i \frac{p(y_i|w, x_i)}{p(-y_i|w, x_i)} \geq c \tag{8}$$

where $c > 1$.

The exact choice of c is arbitrary, since if we can satisfy this for some $c > 1$, then we can also satisfy it for any $c' > 1$ by rescaling w. View it as different shape of logistic function.

Taking logarithms on both side, we get

$$\forall i, \log p(y_i|w, x_i) - \log p(-y_i|w, x_i) \geq \log c \tag{9}$$

Now, we can plug in the definition of $p(y_i|w, x_i)$, which is proportion to $\exp(w^\top x_i)$, and since we can rescale c, we can get

$$\forall i, 2y_i w^\top x_i \geq \log c \tag{10}$$

If we pick c such that $\frac{1}{2}\log c = 1$, so that our conditions can be written in a very simple form

$$\forall i, y_i w^\top x_i \geq 1 \tag{11}$$

The above is a linear feasibility problem, and it can be solved using techniques from linear programming. However,

- the solution may not be unique
- there may be no solution

For the first issue, we can restrict ℓ_2-norm regularization on w, and leads to quadratic program

$$\begin{aligned} &\min_w \lambda \|w\|_2^2 \\ &s.t. \quad \forall i y_i w^\top x_i \geq 1 \end{aligned} \tag{12}$$

For the second issue, we can introduce the slack variables ξ, and get

$$\begin{aligned} &\min_{w,\xi} \sum_i \xi_i + \lambda \|w\|_2^2 \\ &s.t. \quad \forall i y_i w^\top x_i \geq 1 - \xi_i \quad \forall i \xi_i \geq 0 \end{aligned} \tag{13}$$

3 Multi-class SVM

In binary SVM, we use one hyperplane to separate 2 classes. Now in multi-class SVM, we will use k hyperplanes to separate k classes. That is, each class is associated with one weight vector w_k, and we consider

$$p(y_i = k|w_k, x_i) \propto \exp(w_k^\top x_i) \tag{14}$$

and

$$\hat{y_i} = \max_k p(y_i = k|w_k, x_i) \tag{15}$$

In order to make all training instances are classified correctly, we would like

$$\forall i \frac{p(y_i|w, x_i)}{\max_{k \neq y_i} p(y_i = k|w_k, x_i)} \geq c \tag{16}$$

And we also introduce the slack variables ξ, and lead to multi-class svm formulation

$$\begin{aligned} &\min_{w,\xi} \sum_i \xi_i + \lambda \|w\|_2^2 \\ &s.t. \forall i, \forall k \neq y_i, \quad w_{y_i}^\top x_i - w_k^\top x_i \geq 1 - \xi_i \\ &\forall i \xi_i \geq 0 \end{aligned} \tag{17}$$

An equivalent unconstrained optimization problem where we eliminate the slack variables is

$$\min_w \sum_i \max_{k \neq y_i} \{0, (1 - w_{y_i}^\top x_i + w_k^\top x_i)\} + \lambda \|w\|_2^2 \tag{18}$$

References

[1] http://karlstratos.com/notes/svms.pdf

[2] https://www.cs.ubc.ca/~schmidtm/Documents/2009_Notes_StructuredSVMs.pdf