The Structured Weighted Violations Perceptron Algorithm - Supplementary Material

Rotem Dror and **Roi Reichart**
Faculty of Industrial Engineering and Management, Technion, IIT
{rtmdrr@campus|roiri@ie}.technion.ac.il

We have bounded the number of mistakes SWVP is making in an on-line setup. We next provide guarantees as to how well the algorithm generalizes to a new example.

Generalization Bound Let us consider the training set D as an ordered sequence: $D = \{(x^1, y^1), \ldots, (x^n, y^n)\}$, and let us run the SWVP online algorithm on this sequence. At each round $t = 1, \ldots, n$, the algorithm may update the weight vector $\mathbf{w}$, so we get a sequence of weight vectors $\mathbf{w}^1, \ldots, \mathbf{w}^n$, from which we can create an hypotheses sequence of the form $h^t(x) = \arg\max_{y' \in \mathcal{Y}(x)} \mathbf{w}^t \cdot \phi(x, y')$.

To check the algorithm success in generalizing to a new test example (x^{n+1}, y^{n+1}), we need to decide which hypothesis to use from the above sequence, under the assumption that both the training examples and the new test example are drawn i.i.d from an (unknown) distribution $P(x, y)$.

Freund and Schapire (1999) presented the voted perceptron, a batch variant of the perceptron algorithm, and (Collins, 2002) presented an approximation for this variant called the averaged parameters perceptron that holds the same generalization guarantees. We adapt the averaged parameters setting to our algorithm. The resulting adaptation of (Freund and Schapire, 1999) then states:

Theorem 1 (Freund & Schapire 99)**.** *Assume all examples are generated i.i.d. at random. Let $(x^1, y^1), \ldots, (x^n, y^n)$ be a sequence of training examples and let (x^{n+1}, y^{n+1}) be a test example. For a pair $\boldsymbol{u}$, δ such that $\|\boldsymbol{u}\| = 1$ and $\delta > 0$ define $D_{\boldsymbol{u},\delta}$ as before. Then the probability (over the choice of $n+1$ examples) that the voted SWVP algorithm does not predict y^{n+1} on test instance x^{n+1} is at most*

$$\frac{2}{n+1}\mathbb{E}_{n+1}\left(\inf_{\boldsymbol{u}:\|\boldsymbol{u}\|=1,\delta>0} \frac{(R^{JJ} + D_{U,\delta})^2}{(\delta + r^{diff})^2}\right)$$

where $\mathbb{E}_{n+1}$ is an expected value taken over $n+1$ examples.

Note that the adaptation of (Freund and Schapire, 1999) to the original CSP algorithm provided by (Collins, 2002) gives the generalization bound of $\frac{2}{n+1}\mathbb{E}_{n+1}\left(\inf_{\mathbf{u}:\|\mathbf{u}\|=1,\delta>0} \frac{(R+D_{U,\delta})^2}{\delta^2}\right)$. This means that the generalization bound of SWVP is upper bounded by the generalization bound of CSP (convergence property 1 and theorem 2).

References

Michael Collins. 2002. Discriminative training methods for hidden markov models: Theory and experiments with perceptron algorithms. In *Proc. of EMNLP*, pages 1–8.

Yoav Freund and Robert E Schapire. 1999. Large margin classification using the perceptron algorithm. *Machine learning*, 37(3):277–296.