Maximum Margin Structured Labeling:
Algorithms, Generalization Bounds, and Consistency

David McAllester
Toyota Technological Institute at Chicago (TTI-Chicago)

Overview

- Review of maximum margin structured labeling.

- Theoretical justifications (or lack thereof).

- Consistency.

Decoding

We wish to "decode" or "interpret" a code x as a signal y.

Decoding:

$$f_w(x) = \operatorname*{argmax}_y \; w^T \Phi(x, y)$$

Examples:

- x might be a pair of images and y a stereo depth map.
- x might be a sentence and y a parse tree (or logical form).
- x might be a database and y a set of inferred statements.

Graph Example

Take x to be a directed graph with labeled edges and take y to be a labeling of the nodes of x.

For example x might be a web site and y a labeling of each page as "professor", "student", "project" or "other".

Define a "feature" to be an arc $L_1 \stackrel{L_2}{\rightarrow} L_3$ where L_1 and L_3 are node labels, L_2 is an arc label and $\stackrel{L_2}{\rightarrow}$ represents an unspecified arc in the graph with label L_2. Note that if there are L_e possible edge labels and L_n node labels then there are $L_e L_n^2$ features.

$\Phi(x, y)$ assigns a count to each feature.

w assigns a weight to each feature.

$f_w(x)$ can be computed with the junction tree algorithm or approximated with loopy BP.

Training

We can consider different top level objectives in selecting w.

$$w^* = \operatorname*{argmin}_w \mathrm{E}_{\langle x, y\rangle \sim D} \left[d(y, f_w(x)) \right] \tag{1}$$

$$w^* = \operatorname*{argmin}_w \mathrm{E}_{\langle x, y\rangle \sim D} \left[\log \frac{1}{P(y|x, w)} \right] \tag{2}$$

$$P(y|x, w) = \frac{1}{Z(x, w)} \exp(w^T \Phi(x, y))$$

(1) seems to reflect what actually matters in many applications.

(2) is convex and therefore possibly easier to optimize.

We focus on (1).

Maximum Margin Training

Taskar Guestrin and Koller proposed the following M^3 hinge loss:

$$w^* = \operatorname*{argmin}_w \sum_i \max_{\hat{y}} \left(H(y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w)\right)_+ + \lambda ||w||^2$$

$H(y, y')$ is the Hamming distance between y and y'.

$m(x, y, \hat{y}, w)$ is the margin $w^T \Phi(x, y) - w^T \Phi(x, \hat{y})$

Computing w^* Efficiently

$$\begin{aligned} \text{minimize} \quad & \alpha + \lambda ||w||^2 \\ \text{subject to} \quad & \alpha \geq \sum_i \max_{\hat{y}} \left(H(y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w) \right)_+ \end{aligned}$$

The constraint

$$\alpha \geq \max_{\hat{y}} \left(H(y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w) \right)_+$$

can be represented by a linear program encoding the junction tree algorithm.

The Linear Program

We want to represent:

$$\alpha \geq \max_{\hat{y}} \left(H(y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w) \right)_+$$

or equivalently:

$$\alpha \geq 0, \quad \alpha \geq \max_{\hat{y}} \; H(y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w)$$

For a tree the second inequality can be represented as follows.

$$\begin{aligned}
\alpha_{n,L} &= I[L \neq y_i(n)] + \sum_{n \xrightarrow{\gamma} m} \alpha_{n,L,\gamma,m} \\
\alpha_{n,L,\gamma,m} &\geq \alpha_{m,L'} - \left(w_{y_i(n) \xrightarrow{\gamma} y_i(m)} - w_{L \xrightarrow{\gamma} L'} \right) \\
\alpha &\geq \alpha_{r=L}
\end{aligned}$$

Hinge Loss Alternatives

Multiclass Hinge Loss (Collins):

$$w^* = \underset{w}{\operatorname{argmin}} \sum_i \max_{\hat{y} \neq y_i} \left(1 - m(x_i, y_i, \hat{y}, w)\right)_+ + \lambda ||w||^2 \qquad (3)$$

Altun and Hoffman Hinge Loss:

$$w^* = \underset{w}{\operatorname{argmin}} \sum_i \max_{\hat{y}} \; d(y_i, \hat{y}) \left(1 - m(x_i, y_i, \hat{y}, w)\right)_+ + \lambda ||w||^2 \quad (4)$$

M^3 Hinge Loss (Taskar Guestrin and Koller):

$$w^* = \underset{w}{\operatorname{argmin}} \sum_i \max_{\hat{y}} \left(H(x_i, y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w)\right)_+ + \lambda ||w||^2 \quad (5)$$

These all generalize binary hinge loss.

Theorem 1

Parameters ℓ, r, and s are defined by $\Phi(x, y) \in \Re^{\ell}$, $|\mathcal{Y}(x)| \leq r$, and $||\Phi(x, y)||_1 \leq s$.

In the structured case r is exponentially larger than s.

Theorem 1:

$$L(Q(w),\ D) \leq \frac{\mathcal{L}_1(w,\ S)}{m} + \frac{||w||^2}{m} + \sqrt{\frac{2s^2||w||^2 \ln\left(\frac{rm}{||w||^2}\right) + \ln\left(\frac{m}{\delta}\right)}{(m-1)}}$$

$$\mathcal{L}_1(w, S) = \sum_{i=1}^{m} \max_{\hat{y} \in \mathcal{Y}(x)} d(y_i,\ \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq 1]$$

PAC-Bayesian Proof

Theorem:

$$L(Q(w),\ D) \leq \frac{\mathcal{L}_1(w,\ S)}{m} + \frac{||w||^2}{m} + \sqrt{\frac{2s^2||w||^2 \ln\left(\frac{rm}{||w||^2}\right) + \ln\left(\frac{m}{\delta}\right)}{(m-1)}}$$

This is derived from the PAC-Bayesian Theorem: For any data distribution D and loss function L with values in $[0, 1]$ with probability $1 - \delta$ over the choice of an IID sample S of m pairs we have

$$L(Q, D) \leq L(Q, S) + \sqrt{\frac{KL(Q, P) + \ln\frac{m}{\delta}}{2(m-1)}}$$

Theorem 2

$$L(Q(w),\ D) \ \leq\ \frac{\mathcal{L}_H(w,\ S)}{m} + \frac{||w||^2}{m} + \sqrt{\frac{||w||^2 \ln\left(\frac{2\ell m}{||w||^2}\right) + \ln\left(\frac{m}{\delta}\right)}{2(m-1)}}$$

$$\mathcal{L}_H(w, S) \ =\ \sum_{i=1}^{m} \quad \max_{\hat{y} \in \mathcal{Y}(x)} \ d(y_i,\ \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq H(x_i, f_w(x_i), \hat{y})]$$

$$H(x, y, \hat{y}) \ =\ ||\Phi(x, y) - \Phi(x, \hat{y})||_1$$

Camparing the Two Theorems

$$mL(Q(w), D) \leq \sum_{i=1}^{m} \max_{\hat{y}} \; d(y_i, \; \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq 1] + f(||w||^2)$$

$$mL(Q(w), D) \leq \sum_{i=1}^{m} \max_{\hat{y}} \; d(y_i, \; \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq H(x_i, f_w(x_i), \hat{y})] + g(||w||^2)$$

$$mL(Q(2sw), D) \leq \sum_{i=1}^{m} \max_{\hat{y}} \; d(y_i, \; \hat{y}) I \left[m(x_i, f_w(x_i), \hat{y}, w) \leq \frac{H(x_i, f_w(x_i), \hat{y})}{2s} \right] + \; g(||2sw||^2)$$

$$g(||2sw||^2) < f(||w||^2)$$

Simplifying the Regularization

$$w^* = \operatorname*{argmin}_w \sum_{i=1}^{m} \max_{\hat{y}} d(y_i, \hat{y}) I[m(x_i, f_w(x_i), \hat{y}, w) \leq H(x_i, f_w(x_i), \hat{y})] + \lambda ||w||^2$$

$$= \operatorname*{argmin}_w \sum_{i=1}^{m} \max_{\hat{y}} d(y_i, \hat{y}) \left(H(x_i, f_w(x_i), \hat{y}) - m(x_i, f_w(x_i), \hat{y}, w)\right)^+ + \lambda ||w||^2$$

Compare to M^3 hinge loss:

$$w^* = \operatorname*{argmin}_w \sum_{i=1}^{m} \max_{\hat{y}} \left(H(x, y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w)\right)_+ + \lambda ||w||^2$$

$$(x)^+ = I[x \geq 0]$$
$$(x)_+ = \max(0, x)$$

Convexity

The generalization bound suggests the following convex hinge loss.

$$w^* = \underset{w}{\text{argmin}} \sum_{i=1}^{m} \max_{\hat{y}} \; d(y_i, \; \hat{y}) \left(H(x_i, y_i, \hat{y}) - m(x_i, y_i, \hat{y}, w)\right)_+ \; + \; \lambda ||w||^2$$

Consistency

Consider the top level goal:

$$d^* = \inf_w cal E(w)$$

$$\mathcal{E}(w) = \mathrm{E}_{\langle x,\, y\rangle \sim D}\left[d(y, f_w(x))\right]$$

Let A be an algorithm taking as input a sample S and producing a weight vector as output.

Let S be an infinite sample and let S_m be the first m elements of S.

I will call algorithm A is consistent if $\mathcal{E}(A(S_m)) \rightarrow d^*$ (the sequence approaches d^* in probability).

Consistency of the Generalization Bound

There exists a regularization schedule λ_m under which the following generalization bound algorithm is consistent:

$$w^* \;=\; \underset{w}{\operatorname{argmin}} \;\; \sum_{i=1}^{m} \max_{\hat{y}} \; d(y_i,\; \hat{y}) \left(H(x_i, f_w(x_i), \hat{y}) - m(x_i, f_w(x_i), \hat{y}, w)\right)^+ \;\; + \;\; \lambda_m ||w||^2$$

Inconsistency of Multiclass Hinge Loss

Consider the case where all x values are the same and there is an independent weight for each $\hat{y}$.

The multiclass margin m_i can now be written as follows.

$$m_i = w_i - \max_{j \neq i} w_j$$

In the limit of infinite training data we have the following.

$$w^* = \operatorname*{argmin}_w \sum_i p_i \left(1 - m_i\right)_+ \tag{6}$$

Assume $p_i < 1/2$ forall i.

In this case one can show that w^* is uniform — all weights are the same.

For example, if we increase the weight of the most likely label then we increase the margin for that label but decrease the margin of every other label by the same amount.

Kernels and Semi-Supervised Learning

In some applications x may be contain vectorial data. Consider speech recognition or sequential character recognition.

In this case we take $\Phi_i(x, y)$ to be a vector.

$$
\begin{aligned}
w^T \Phi(x, y) &= F_f(x, y) = \sum_i f(\Phi_i(x, y)) \\
f^* &= \operatorname*{argmin}_f \sum_t \max_{\hat{y}} \left(H(\hat{y}, y_t) - m(x, y_t, \hat{y}, f)\right)_+ + \lambda ||f||^2 \\
m(x, y_t, \hat{y}) &= F_f(x, y_t) - F_f(x, \hat{y})
\end{aligned}
$$

Here f in an RKHS defined by a kernel on the vectors $\Phi_i(x, y)$.

In semi-supervised learning we can use unlabeled data to construct a graph kernel (Altun and McAllester 05).

Summary

- The most effective choice of hinge loss remains unclear.

- There is a fundamental conflict between consistency and convexity.