Hybrid model of Conditional Random Field and Support Vector Machine

Qinfeng Shi
Australia National University and NICTA
Canberra, Australia
qinfeng.shi@anu.edu.au

Mark Reid
Australia National University
Canberra, Australia
mark.reid.anu.edu.au

Tiberio Caetano
Australia National University and NICTA
Canberra, Australia
Tiberio.Caetano@nicta.com.au

1 Introduction

Conditional Random Fields (CRFs) [4, 13, 3, 17] are semi-generative (despite often being classified as discriminative models) in the sense that it estimates the conditional probability $D(y|x)$ (given any observation x) of any label y, which is **generated** from $D(y|x)$. Estimating $D(y|x)$ is usually more efficient than estimating $D(x|y)$ when there aren't sufficient observation x per class or there are too many labels (e.g. there are exponential many y for a chain-like x). To avoid causing terminology confusion, we call the models that estimate underlying distribution (either $D(y|x)$ or $D(x|y)$) probabilistic models. Unlike CRFs, Support Vector Machine (SVM) is a **pure** discriminative model in the sense that it seeks for a predicting function regardless of modeling the underlying distribution. We are interested in revealing the nature of probabilistic models and pure discriminative models, in order to obtain a model having the advantages of both.

It is known that probabilistic models often converge to the true distribution asymptotically (i.e. fisher consistent). However, the consistency is often useless in practice, since in real world it is impossible to fit the models with infinite many data in a finite time. SVM [1, 14, 16] is fisher inconsistent in multiclass [9] and structured label case, however, it does provide a PAC bound on the true error (known as generalization bound). Particularly, its PAC-Bayes margin bound [7] is rather tight, which states that, knowing training sample size m, hypothesis space $\mathcal{H}$ and margin threshold γ, with overwhelming probability at least $1 - \delta$, the true error is upper bounded by the empirical error $+O(\sqrt{\frac{\gamma^{-2}\log|\mathcal{H}|\log m + \log \delta^{-1}}{m}})$. Clearly when the m and γ are large, the true error is closely bounded by the empirical error. It gives a quantitative evaluation of the goodness of the algorithm in practice. Several studies ([12, 6, 5, 8]) show the bounds are useful in real applications (e.g. model selections).

Is there a model that is fisher consistent for classification and has a generalization bound? We use a naive combination of two models by simply weighted summing up the losses of two. It turns out a surprising theoretical result — the hybrid loss could be fisher consistent in some circumstance and it has a PAC-bayes bound on its true error.

2 Hybrid Loss and Risk Minimization

Given the observation domain $\mathcal{X}$, the label domain $\mathcal{Y}$, a feature map $\phi(x, y) : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, $\alpha \in [0, 1]$, and define $f(x, y) = \langle w, \phi(x, y)\rangle$, the hybrid loss ℓ is

$$\ell(f; x, y) := \alpha(-\log p(y|x; w)) + (1 - \alpha)[1 - f(x, y) + f(x, y^*)]_+, \tag{1}$$

where the estimated distribution $p(y|x;w) = \frac{\exp(f(x,y))}{\sum_{y' \in \mathcal{Y}} \exp(f(x,y'))}$, $[z]_+ = z$ when $z > 0$, and equals 0 elsewhere. And

$$y_i^* \in \underset{y' \neq y_i, y' \in \mathcal{Y}}{\text{argmax}} \{1 + f(x_i, y') - f(x_i, y_i)\} \quad (2)$$

$$= \underset{y' \neq y_i, y' \in \mathcal{Y}}{\text{argmin}} \{f(x_i, y_i) - f(x_i, y')\}. \quad (3)$$

Here $\min_{y' \neq y_i, y' \in \mathcal{Y}} \{f(x_i, y_i) - f(x_i, y')\}$ is the margin denoted by $M(x_i, y_i; f)$. When f is expressed by w explicitly, the margin is also written as $M(x_i, y_i; w)$.

Given training data $\mathbb{X} = \{x_1, \ldots, x_m\}$ and $\mathbb{Y} = \{y_1, \ldots, y_m\}$, the empirical risk is:

$$\frac{\lambda ||w||^2}{2} + \sum_{i=1}^{m} \ell(f; x_i, y_i). \quad (4)$$

Here w is estimated by minimizing (4). The convexity of (4) gives a unique global optimum.

3 Fisher Consistency For Classification

Estimator converging to the true distribution asymptotically is a desirable property. An estimator or algorithm is **fisher consistent for classification**[1] also known as "classification calibration" (see [9] and [15]), iff given entire data population the estimated model f predicts as good as predicting via the true data distribution $D(y|x)$ for all D and all x, that is,

$$\underset{y}{\text{argmax}}\, f(x, y) \subset \underset{y}{\text{argmax}}\, D(y|x), \quad \forall x. \quad (5)$$

Note that the above definition applies to binary, multiclass and structured output y. The original classification calibration definition in [15] on binary classification ignores the case when there are ties on the choice of y according to $D(y|x)$. Thus when ties happen, the estimated f can be arbitrary thus it can perform very poorly. The subset relation here ensures that f still performs reasonably well with ties. Liu [9] shows that SVM isn't fisher consistent for multiclass, because when there is no dominant class, i.e. $\max_y D(y|x) < 1/2$, the minimizer f of the expected SVM loss (known as hinge loss) is constant, i.e. $f(x, y) = f(x, y')$ for all y, y'. If there is no restriction on f, a straightforward way to check fisher consistency, is to get

$$f^* \in \underset{f}{\text{argmin}}\, \mathbb{E}_{y \sim D(y|x)}[\ell(f; x, y)], \quad (6)$$

and then check whether (5) holds for all f^*. The non-parametric [2] CRF loss is known to be fisher consistent, because the derivative of the expectation is zero, iff $p = D$.

Theorem 1 (Margin Condition) *The hybrid loss is fisher consistent, iff* $M(x, y = j_0; f^*) > 0$, *where* $j_0 = \text{argmax}_j D(y = j|x)$, *for all* x.

Proof By the definition of margin, clearly $f(x_i, j_0) > f(x_i, y')$ for $y' \neq j_0 \Leftrightarrow \text{argmax}_y f(x, y) = \text{argmax}_y D(y|x)$. So the theorem 1 follows. ■

Theorem 2 (Necessary Condition) *For* k *classes problem, if the hybrid loss is fisher consistent, there exists a* f^*, *such that*

$$\Big(\alpha \sum_{j=1}^{k} D(j|x) \log p(j|x; f^*) + (1 - \alpha)(2D(j_0|x) - 1) M(x, j_0; f^*)\Big) > \alpha \log(\frac{1}{k}) \quad (7)$$

[1] Note that the fisher consistency for classification is weaker than fisher consistency for density estimation. The former requires the same prediction only, while the latter requires the estimated density is the same as the true data distribution. In this paper, we focus on the former only.

[2] Note that the fisher consistency analysis is usually for non-parametric models instead of parametric ones like (4). One can argue that when the w doesn't enforce any restriction on the f (e.g. w is in a RKHS with infinite dimensionality), the fisher consistency analysis holds. However, when w has low dimensionality, the classical fisher consistency analysis doesn't work any more. And it is interesting to exploit the gap, which won't be covered in this paper.

Proof The lemma 4 in [9] holds here as well. By adding the negation of the CRF loss, the minimizer f^* in (6) equals to

$$\operatorname*{argmax}_f \Big(\alpha \sum_{j=1}^{k} D(j|x) \log p(j|x; f) + (1-\alpha)(2D(j_0|x) - 1) M(x, j_0; f)\Big). \tag{8}$$

If the hybrid loss is fisher consistent, the estimated distribution p won't be uniform. So the theorem holds. ■

Distribution dependent consistency Varying α changes the fisher consistency of the hybrid loss. We conjecture that there exists a threshold $\tau \neq 1$, such that for all $\alpha \geq \tau$, (5) holds for the hybrid loss. Such τ depends on the location D on the probability simplex. Let B_D be the maximum ball centered at D, enclosed by hyperplanes $q(j_0|x) = q(j|x), j \neq j_0$, where $j_0 = \operatorname{argmax}_j D(j|x)$. Clearly when the estimated p falls into B_D, the (5) holds. When D is close to the center of the simplex or any of the above hyperplanes, the ball becomes very small, i.e. the τ is close to 1.

4 Generalization Bound

McAllester introduced PAC-Bayes analysis [10, 11] which was further refined [12, 6, 5, 8]. Germain et al. [2] recently gave a simplified PAC-Bayesian bound proof on Gibbs classifier. Here we provide a PAC-Bayes generalization bound for the proposed model.

Theorem 3 (Generalization Bound) *For any data distribution D, for any prior P over w, for any $\delta \in (0, 1]$ and $\alpha \in [0, 1)$ and for any $\gamma \geq 0$, for any w, with probability at least $1 - \delta$ over random samples S from D with m instances, we have*

$$\mathbb{E}_D\Big[\Big(\gamma - M(x, y; w)\Big)_+\Big] \leq \frac{1}{m}\sum_{i=1}^{m}\Big(\gamma - M(x_i, y_i; w)\Big)_+ + \frac{1}{(1-\alpha)}\left(\alpha\sqrt{\frac{1}{m}} + \sqrt{\frac{\ln \frac{1}{P(w)} + \ln A(\alpha, w) + \ln \frac{1}{\delta(1-e^{-2})}}{2m}}\right),$$

where

$$R(\alpha, w) = \alpha\,\mathbb{E}_D\Big[-\ln p(y|x; w)\Big] + (1-\alpha)\,\mathbb{E}_D\Big[\Big(\gamma - M(x, y; w)\Big)_+\Big], \tag{9}$$

$$R_S(\alpha, w) = \Big[\alpha\frac{\sum_{i=1}^m - \ln p(y_i|x_i; w)}{m} + (1-\alpha)\frac{\sum_{i=1}^m \Big(\gamma - M(x_i, y_i; w)\Big)_+}{m}\Big], \tag{10}$$

$$A(\alpha, w) = \mathbb{E}_{s\sim D^m}\, e^{2m(R(\alpha,w) - R_S(\alpha,w))^2}. \tag{11}$$

Here A is upper bounded independently of D. For example, for a zero-one loss, it is upper bounded by $m + 1$ (see [2]). The theorem gives a bound on the true margin error of the hybrid model. The theorem follows theorem 5 in the appendix immediately.

5 Conclusion and Discussion

We show that a naive hybrid loss has a surprising theoretical result — the hybrid loss could be fisher consistent and has a PAC-bayes bound. Current fisher consistency analysis focus on non-parametric models, whereas in real world parametric models are very popular such as non-kernelized CRFs and linear SVM. How to apply fisher consistency analysis to parametric models with finite dimensionality is still an open question.

6 Acknowledgement

This work has been supported by the Australian National University and National ICT Australia. We thank to the mystery second reviewer for pointing out the distinction between parametric and non-parametric models for fisher consistency.

References

[1] K. Crammer and Y. Singer. On the algorithmic implementation of multiclass kernel-based vector machines. *J. Mach. Learn. Res.*, 2:265–292, 2001.

[2] Pascal Germain, Alexandre Lacasse, Francois Laviolette, and Mario Marchand. Pac-bayesian learning of linear classifiers. In *ICML*, 2008.

[3] J. Lafferty, X. Zhu, and Y. Liu. Kernel conditional random fields: Representation and clique selection. In *Proc. Intl. Conf. Machine Learning*, volume 21, page 64, San Francisco, CA, 2004. Morgan Kaufmann.

[4] J. D. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic modeling for segmenting and labeling sequence data. In *Proc. Intl. Conf. Machine Learning*, volume 18, pages 282–289, San Francisco, CA, 2001. Morgan Kaufmann.

[5] John Langford. Tutorial on practical prediction theory for classification. *JMLR*, 6:273–306, 2005.

[6] John Langford, Matthias Seeger, and Nimrod Megiddo. An improved predictive accuracy bound for averaging classifiers. In *ICML*, 2001.

[7] John Langford and Mattias Seeger. Bounds for averaging classifiers. Technical Report 102, Computer Sciense, CMU, 1 2001.

[8] John Langford and John Shawe-Taylor. Pac-bayes and margin. In *Neural Information Processing Systems*. MIT Press, 2003.

[9] Yufeng Liu. Fisher consistency of multicategory support vector machines. In *Proc. Intl. Conf. Machine Learning*, 2007.

[10] David A. McAllester. Some pac-bayesian theorems. In *COLT*, pages 230–234, 1998.

[11] David A. McAllester. Pac-bayesian model averaging. In *COLT*, 1999.

[12] David A. McAllester. Pac-bayesian stochastic model selection. *ML*, 2001.

[13] F. Sha and F. Pereira. Shallow parsing with conditional random fields. In *Proceedings of HLT-NAACL*, pages 213–220, Edmonton, Canada, 2003. Association for Computational Linguistics.

[14] B. Taskar, C. Guestrin, and D. Koller. Max-margin Markov networks. In S. Thrun, L. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 25–32, Cambridge, MA, 2004. MIT Press.

[15] A. Tewari and P.L. Bartlett. On the consistency of multiclass classification methods. *Journal of Machine Learning Research*, 8:1007–1025, 2007.

[16] I. Tsochantaridis, T. Hofmann, T. Joachims, and Y. Altun. Support vector machine learning for interdependent and structured output spaces. In *Proc. Intl. Conf. Machine Learning*, New York, NY, USA, 2004. ACM Press.

[17] S. V. N. Vishwanathan, Nicol N. Schraudolph, Mark Schmidt, and Kevin Murphy. Accelerated training conditional random fields with stochastic gradient methods. In *Proc. Intl. Conf. Machine Learning*, pages 969–976, New York, NY, USA, 2006. ACM Press.

Lemma 4 (PAC-Bayesian bound[12, 2]) *For any data distribution D, for any prior P and posterior Q over w, for any $\delta \in (0,1]$, for any loss ℓ. With probability at least $1-\delta$ over random sample S from D with m instances, we have*

$$R(Q,\ell) \leq R_S(Q,\ell) + \sqrt{\frac{KL(Q||P) + \ln(\frac{1}{\delta}\mathbb{E}_{s\sim D^m}\mathbb{E}_{w\sim P} e^{2m(R(Q,\ell)-R_S(Q,\ell))^2})}{2m}},$$

where $KL(Q||P) := \mathbb{E}_{w\sim Q}\ln(\frac{Q(w)}{P(w)})$ is the Kullback-Leibler divergence between Q and P, and $R(Q,\ell) = \mathbb{E}_{Q,D}[\ell(x,y;w)], R_S(Q,\ell) = \mathbb{E}_Q \frac{\sum_{i=1}^m \ell(x_i,y_i;w)}{m}$.

Theorem 5 (Bound on Averaging classifier) *For any data distribution D, for any prior P and posterior Q over w, for any $\delta \in (0,1]$ and $\alpha \in [0,1)$ and for any $\gamma \geq 0$. With probability at least $1-\delta$ over random sample S from D with m instances, we have*

$$\mathbb{E}_{Q,D}\Big[[\gamma - M(x,y;w)]_+\Big] \leq \frac{1}{m}\mathbb{E}_Q\Big[\sum_{i=1}^m[\gamma - M(x_i,y_i;w)]_+\Big] + \frac{\alpha}{1-\alpha}\sqrt{\frac{1}{m}} + \frac{1}{1-\alpha}\sqrt{\frac{KL(Q||P) + \ln A(\alpha) + \ln\frac{1}{\delta(1-e^{-2})}}{2m}}, \qquad (12)$$

where $KL(Q||P) := \mathbb{E}_{w\sim Q}\ln(\frac{Q(w)}{P(w)})$ is the Kullback-Leibler divergence between Q and P, and

$$R(\alpha) = \alpha\,\mathbb{E}_{Q,D}\Big[-\ln p(y|x;w)\Big] + (1-\alpha)\,\mathbb{E}_{Q,D}\Big[\Big(\gamma - M(x,y;w)\Big)_+\Big], \qquad (13)$$

$$R_S(\alpha) = \mathbb{E}_Q\Big[\alpha\frac{\sum_{i=1}^m -\ln p(y_i|x_i;w)}{m} + (1-\alpha)\frac{\sum_{i=1}^m\Big(\gamma - M(x_i,y_i;w)\Big)_+}{m}\Big], \qquad (14)$$

$$A(\alpha) = \mathbb{E}_{s\sim D^m}\mathbb{E}_{w\sim P} e^{2m(R(\alpha)-R_S(\alpha))^2}. \qquad (15)$$

Proof Since $\mathbb{E}_D\Big(\mathbb{E}_Q\Big[\frac{\sum_{i=1}^m -\ln p(y_i|x_i;w)}{m}\Big]\Big) = \mathbb{E}_{Q,D}\Big[-\ln p(y|x;w)\Big]$, by Chernoff bound we have

$$\Pr_{S\sim D^m}\Big(\mathbb{E}_Q\Big[\frac{\sum_{i=1}^m -\ln p(y_i|x_i;w)}{m}\Big] - \mathbb{E}_{Q,D}\Big[-\ln p(y|x;w)\Big] < \epsilon\Big) > 1 - e^{-2m\epsilon^2}.$$

Define $B(S) := \mathbb{E}_Q\Big[\frac{\sum_{i=1}^m -\ln p(y_i|x_i;w)}{m}\Big] - \mathbb{E}_{Q,D}\Big[-\ln p(y|x;w)\Big]$.

Applying Lemma 4 for $R(\alpha)$ and $R_S(\alpha)$, we have for any P, Q

$$\begin{aligned}
\delta &> \Pr_{S\sim D^m}\left(R(\alpha) \geq R_S(\alpha) + \sqrt{\frac{KL(Q||P) + \ln\frac{1}{\delta} + \ln A(\alpha)}{2m}}\right) \\
&\geq \Pr_{S\sim D^m}\left(R(\alpha) \geq R_S(\alpha) + \sqrt{\frac{KL(Q||P) + \ln\frac{1}{\delta} + \ln A(\alpha)}{2m}}, B(S) < \epsilon\right) \\
&\geq \Pr_{S\sim D^m}\Bigg((1-\alpha)\,\mathbb{E}_{Q,D}\left[\Big(\gamma - M(x,y;w)\Big)_+\right] \geq (1-\alpha)\frac{\sum_{i=1}^m \Big(\gamma - M(x_i,y_i;w)\Big)_+}{m} \\
&\quad +\alpha\epsilon + \sqrt{\frac{KL(Q||P) + \ln\frac{1}{\delta} + \ln A(\alpha)}{2m}}, B(S) < \epsilon\Bigg) \\
&= \Pr_{S\sim D^m}\Bigg((1-\alpha)\,\mathbb{E}_{Q,D}\left[\Big(\gamma - M(x,y;w)\Big)_+\right] \geq (1-\alpha)\frac{\sum_{i=1}^m \Big(\gamma - M(x_i,y_i;w)\Big)_+}{m} \\
&\quad +\alpha\epsilon + \sqrt{\frac{KL(Q||P) + \ln\frac{1}{\delta} + \ln A(\alpha)}{2m}}\Bigg| B(S) < \epsilon\Bigg) \Pr_{S\sim D^m}\Big(B(S) < \epsilon\Big) \\
&\geq \Pr_{S\sim D^m}\Bigg((1-\alpha)\,\mathbb{E}_{Q,D}\left[\Big(\gamma - M(x,y;w)\Big)_+\right] \geq (1-\alpha)\frac{\sum_{i=1}^m \Big(\gamma - M(x_i,y_i;w)\Big)_+}{m} \\
&\quad +\alpha\epsilon + \sqrt{\frac{KL(Q||P) + \ln\frac{1}{\delta} + \ln A(\alpha)}{2m}}\Bigg) \Pr_{S\sim D^m}\Big(B(S) < \epsilon\Big)
\end{aligned}$$

Divide two sides by $\Pr_{S\sim D^m}\Big(B(S) < \epsilon\Big)$, we get

$$\begin{aligned}
&\Pr_{S\sim D^m}\Bigg((1-\alpha)\,\mathbb{E}_{Q,D}\left[\Big(\gamma - M(x,y;w)\Big)_+\right] \geq (1-\alpha)\frac{\sum_{i=1}^m \Big(\gamma - M(x_i,y_i;w)\Big)_+}{m} \\
&\quad +\alpha\epsilon + \sqrt{\frac{KL(Q||P) + \ln\frac{1}{\delta} + \ln A(\alpha)}{2m}}\Bigg) \leq \frac{\delta}{\Pr_{S\sim D^m}\Big(B(S) < \epsilon\Big)} \leq \frac{\delta}{1 - e^{-2m\epsilon^2}}.
\end{aligned}$$

Let $\epsilon = \sqrt{\frac{1}{m}}$, and then let $\delta' = \frac{\delta}{1-e^{-2m(\epsilon^2)}} = \frac{\delta}{1-e^{-2}}$, we get $\delta = \frac{1}{\delta'(1-e^{-2})}$. The theorem follows by substituting δ with δ' and dividing by $(1-\alpha)$ on both sides of the inequality inside of the probability. ■