10-708: Probabilistic Graphical Models 10-708, Spring 2016

25 : Posterior Regularization: an integrative paradigm for learning GMs

Lecturer: Matt Gormley *Scribes: Tzu-Ming Kuo*

1 Continued: Maximum Entropy Discrimination Markov Network

1.1 Motivation

Likelihood-based estimation and max-margin learning method are two widely used learning approaches in machine learning community. Likelihood-based estimation is usually applied to train probabilistic model. It basically tries to find a model that maximizes the likelihood on some given training samples. On the other hand, max-margin methods tries to find a model that generates correct output while keeping the decision value of the correct output away from that of incorrect output. This two different approaches has different advantages:

- Likelihood Estimation
  - Easy to perform Bayesian learning,
  - Incorporate prior knowledge, latent structures, missing data
  - Bayesian or direct regularization
- Max-Margin Learning
  - Dual sparsity
  - Sound theoretical guarantee
  - Easy to incorporate kernel trick

Since these approaches enjoy different advantages, it is natural to come up a model that incorporates these two approaches.

1.2 Maximum Entropy Discrimination

This work is a attempt to integrate likelihood-based method an max-margin learning. It takes the idea of model averaging from Bayesian learning and max-margin for max-margin learning. In model averaging, instead of using only one model to generate output, every models are used to generate outputs and the final outputs is the weighted average of outputs from all models. During training procedure, the goal is to find the weight that has the best performance on training samples. Take binary classification as an example, the model output $\hat{y}$ of an instance $\boldsymbol{x}$ is

$$\hat{y} = \text{sign}\left(\int p(\boldsymbol{w})F(\boldsymbol{x};\boldsymbol{w})d\boldsymbol{w}\right).$$

Substitute the model output into max-margin constraints yields the expected margin constraint

$$\int p(\boldsymbol{w}) \left[yF(\boldsymbol{x}; \boldsymbol{w}) - \xi \right] d\boldsymbol{w} \geq 0.$$

Sometimes we may have prior knowledge about the distribution $p(\boldsymbol{w})$. For example, if we think $\boldsymbol{w}$ should be close to the origin, we may want to use a zero-mean Gaussian distribution as the prior for $\boldsymbol{w}$. Given a prior distribution $p_0(\boldsymbol{w})$ for $\boldsymbol{w}$, we wants $p(\boldsymbol{w})$ to be close $p_0(\boldsymbol{w})$ as much as possible. An approach is to minimize the KL-divergence between $p(\boldsymbol{w})$ and $p_0(\boldsymbol{w})$.

Putting objectives above together yields the optimization problem of Maximum Entropy Discrimination for binary classification.

$$\begin{aligned} &\min_{p(\boldsymbol{w})} \mathrm{KL}\left(p(\boldsymbol{w}) || p_0(\boldsymbol{w})\right) \\ \text{subjects to } &\int p(\boldsymbol{w}) \left[y_i F(\boldsymbol{x}_i; \boldsymbol{w}) - \xi_i \right] d\boldsymbol{w} \geq 0, \forall i \\ &\xi_i \geq 0. \end{aligned}$$

1.3 Maximum Entropy Discrimination Markov Networks

Maximum Entropy Discrimination Markov Networks extends the idea of Maximum Entropy Discrimination on Markov Networks and consider structured prediction problems. The optimization problem is

$$\begin{aligned} &\min_{p(\boldsymbol{w}), \boldsymbol{\xi}} \mathrm{KL}\left(p(\boldsymbol{w}) || p_0(\boldsymbol{w})\right) + U(\boldsymbol{\xi}) \\ &\text{subjects to } p(\boldsymbol{w}) \in \mathcal{F}_1, \xi_i \geq 0, \end{aligned}$$

where U is the loss penalizing the violation of expected margin constraints, $\mathcal{F}_1$ is the expected margin constraints

$$\mathcal{F}_1 = \{p(\boldsymbol{w}) : \int p(\boldsymbol{w}) [\Delta F_i(y; \boldsymbol{w}) - \Delta l_i(y)] d\boldsymbol{w} - \xi_i \geq 0, \forall y \neq y_i\},$$

$\Delta F_i(y; \boldsymbol{w}) \equiv F(\boldsymbol{x}_i, y_i; \boldsymbol{w}) - F(\boldsymbol{x}_i, y; \boldsymbol{w})$, and $\Delta_i(y)$ is the margin required between predicting y_i rather than y.

Given $p(\boldsymbol{w})$ the prediction rule is

$$h_1(\boldsymbol{x}, p(\boldsymbol{w})) = \arg \max_{y \in \mathcal{Y}(x)} \int p(\boldsymbol{w}) F(\boldsymbol{x}, y, \boldsymbol{w}) d\boldsymbol{w}.$$

1.3.1 Optimization Problem

It can be shown that the optimal posterior distribution $p^*(\boldsymbol{w})$ can be expressed as

$$\frac{1}{Z(\alpha)} p_0(\boldsymbol{w}) \exp\left(\sum_{i,y} \alpha_i(y) [\Delta F_i(y; \boldsymbol{w}) - \Delta l_i(y)] \right),$$

where α are the solution of the dual problem

$$\begin{aligned} &\max_{\alpha} -\log Z(\alpha) - U^*(\alpha) \\ &\text{subjects to } \alpha_i(y) \geq 0, \forall y, i, \end{aligned}$$

and U^* is conjugate of U.

1.3.2 Gaussian Maximum Entropy Discrimination Markov Networks

An interesting observation is that if we specify model F, margin penalty U, and prior $p_0(\boldsymbol{w})$ as following

$$\begin{aligned} F(\boldsymbol{x}, y, \boldsymbol{w}) &\equiv \boldsymbol{w}^T f(\boldsymbol{x}, y), \\ U(\xi) &\equiv \sum_i \xi_i, \\ p_0(\boldsymbol{w}) &\equiv \mathcal{N}(0, I), \end{aligned}$$

the dual problem is

$$\begin{aligned} \max_{\alpha} \sum_{i,y} \alpha_i(y)\Delta_i(y) - \frac{1}{2}\| \sum_{i,y} \alpha_i(y)\Delta f_i(y)\|^2 \\ \text{subjects to } \sum_y \alpha_i(y) = C, \forall i \\ \alpha_i(y) \geq 0, \forall i, y. \end{aligned}$$

The corresponding primal optimal is

$$p^*(\boldsymbol{w}) = \mathcal{N}(\mu_{\boldsymbol{w}}, I),$$

where

$$\mu_{\boldsymbol{w}} = \sum_{i,y} \alpha_i(y)\Delta f_i(y).$$

The prediction rule is

$$\hat{y} = \arg \max_{y \in \mathcal{Y}(x)} \int p(\boldsymbol{w}) F(\boldsymbol{x}, y; \boldsymbol{w}) d\boldsymbol{w} = \mu_{\boldsymbol{w}}^T f(\boldsymbol{x}, y).$$

The dual problem and the prediction rule are identical to that of M^3N. Therefore, in essence, Gaussian Maximum Entropy Discrimination Markov Networks is probabilistic version of M^3N. It uses the posterior mean of $p^*(\boldsymbol{w})$ instead of $\boldsymbol{w}^*$. This means that Maximum Entropy Discrimination Markov Networks is a more general framework, subsumes M^3N and offers some advantages for being probabilistic. The advantages including PAC-Bayesian prediction error guarantee, introducing useful biases via entropy regularization, and integrating Generative and Discriminative principles.

1.3.3 Laplace Maximum Entropy Discrimination Markov Networks

The Gaussian prior can be replaced with a Laplace prior

$$p_o(\boldsymbol{w}) = \prod_{k=1}^{K} \frac{\sqrt{\lambda}}{2} e^{-\sqrt{\lambda}|w_l|} = \left(\frac{\sqrt{\lambda}}{2}\right)^K e^{-\sqrt{\lambda}\|\boldsymbol{w}\|}.$$

This is called Laplace Maximum Entropy Discrimination Markov Network. The optimization problem becomes

$$\begin{aligned} \min_{\mu,\xi} \sqrt{\lambda} \sum_{k=1}^{K} \left(\sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}} \log \frac{\sqrt{\lambda\mu_k^2 + 1} + 1}{2} \right) + C \sum_{i=1}^{N} \xi_i \\ \text{subjects to } \boldsymbol{\mu}^T \Delta f_i(y) - \Delta l_i(y) + \xi_i \geq 0, \xi_i \geq 0, \forall i, \forall y \neq y_i. \end{aligned}$$

The nature of Laplace prior has a l_1 regularization effect on the components of $\boldsymbol{w}$, push the weights towards zero. The hyper-parameter λ controls the regularization effect. As λ increases, the model becomes more regularized. This can be seen by the fact that

$$\langle w_k \rangle_p = \frac{2\eta_k}{\lambda - \eta_k^2}, \forall k,$$

where $\boldsymbol{\eta}$ is is a linear combination of the support vectors:

$$\boldsymbol{\eta} = \sum_{i,y} \alpha_i(y) \Delta f_i(y).$$

The shrinkage effect makes the primal solution of Laplace Maximum Entropy Discrimination Markov Network to be sparse. This means that a large number of components of $\boldsymbol{w}$ will be 0 and informative components will be more likely be non-zero. In addition, since we only have non-zero value on the dual variables associated with active constraints in primal problem and solution of the dual problem will be sparse. Therefore, Laplace Maximum Entropy Discrimination Markov Network have not only primal sparsity but also dual sparsity. Define KL-norm as

$$\|\mu\|_{\mathrm{KL}} \equiv \sum_{k=1}^{K} \left(\sqrt{\mu_k^2 + \frac{1}{\lambda}} - \frac{1}{\sqrt{\lambda}} \log \frac{\sqrt{\lambda \mu_k^2 + 1} + 1}{2} \right),$$

and compare with l_1 and l_2 norm by plotting a level curve.

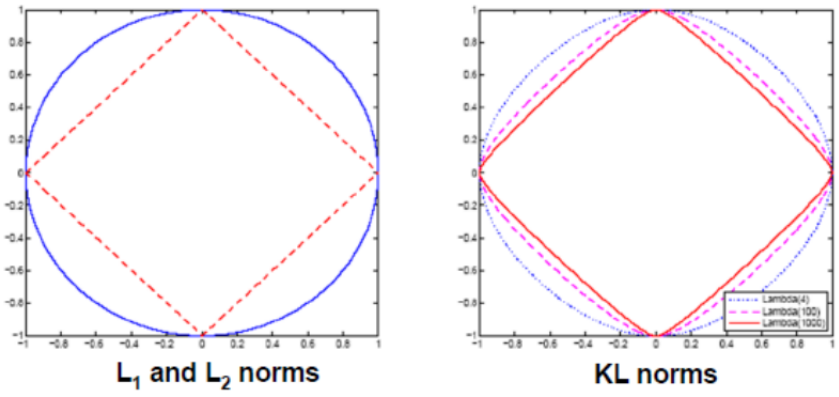

Figure 1: The KL-norm can be considered an interpolation between l_1 and l_2 norm controlled by parameter λ.

1.3.4 Variational Learning of Laplace Maximum Entropy Discrimination Markov Networks

Though Laplace Maximum Entropy Discrimination Markov Network has nice property, the both primal problem and the dual problem is infeasible to be exactly solved. A alternative way to solve this problem is to utilize the hierarchical representation of Laplace prior. Using this property we get an upper bound of

KL-norm

$$\begin{aligned} KL(p\|p_0) &= -H(p) - \langle \log \int p(\boldsymbol{w}|\tau)p(\tau|\lambda)d\tau \rangle_p \\ &\leq -H(p) - \langle \log q(\tau) \int \frac{p(\boldsymbol{w}|\tau)p(\tau|\lambda)}{q(\tau)} d\tau \rangle_p \\ &\equiv \mathcal{L}(p(\boldsymbol{w}), q(\tau)). \end{aligned}$$

And we minimize the following problem, which is an upper bound of the primal problem

$$\min_{p(\boldsymbol{w})\in\mathcal{F}_1, q(\tau), \xi} \mathcal{L}(p(\boldsymbol{w}), q(\tau)) + U(\xi).$$

Though this problem is still difficult to solve if we optimize all variables at the same time. However, a nice property of this problem is that when $q(\tau)$ is fixed, the problem is reduced to a M^3N problem. And if $p(\boldsymbol{w})$ and ξ_i is fixed instead, $q(\tau)$ has a closed-form solution and its expectation is

$$\langle \frac{1}{\tau_k} \rangle_q = \sqrt{\frac{\lambda}{\langle w_k \rangle_p}}.$$

This suggests that we can solve this problem by alternatively fixing $q(\tau)$ and $p(\boldsymbol{w}), \xi$ and solves for the optimal for other variables.

2 Posterior Regularization: an integrative paradigm for learning GMs

2.1 Introduction

This lecture is not just about regularization, it is about to integrate what we learnt in this class so far. In the first part of class, parameter estimation for graphical models are based on maximum likelihood principle because that is the most common objective function we use on graphs. In the latter part of class, we have seen some different types of loss functions. For example, we introduce a prior distribution for parameters and we end up optimizing the posterior probability of the parameters of the model given samples. In the last lecture, we also learnt to integrate max-margin learning principle with Markov Network and results in a probability model with dual/primal sparsity and generalization guarantee. All this method enjoys different advantages between each-other. It is natural to ask if we can find a way to integrate those principles.

2.2 Bayesian Inference

Bayesian inference relies on a prior distribution over models and a likelihood function of data given a model. Bayesian inference allow us not only to utilize prior knowledge by designing the prior distribution over models but also have trade-off between the evidence and the prior knowledge.

Bayesian inference is a coherent framework of dealing with uncertainties:

$$p(\mathcal{M}|x) = \frac{p(x|\mathcal{M})\pi(\mathcal{M})}{\int p(x|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}},$$

where $p(x|\mathcal{M})$ is a likelihood function of data x given model $\mathcal{M}$, and $\pi(\mathcal{M})$ is a prior distribution of model $\mathcal{M}$.Bayes rule offers a mathematically rigorous computational mechanism for combining prior knowledge with evidence.

2.2.1 Parametric Bayesian Inference

In parametric Bayesian inference, a model $\mathcal{M}$ can represented by a finite set of parameters θ. The posterior distribution is

$$p(\theta|x) = \frac{p(x|\theta)\pi(\theta)}{\int p(x|\theta)\pi(\theta)d\theta} \propto p(x|\theta)\pi(\theta).$$

There are some examples of conjugacy in picking the appropriate prior:

- Gaussian distribution prior + 2D Gaussian likelihood → Gaussian posterior distribution
- Dirichlet distribution prior + 2D Multinomial likelihood → Dirichlet posterior distribution
- Sparsity-inducing priors + some likelihood models → Sparse Bayesian inference

2.2.2 Nonparametric Bayesian Inference

If the parameter can grow as the number of data increase (i.e., model cannot be represented by finite number of parameters). In this case the posterior distribution is

$$p(\mathcal{M}|x) = \frac{p(x|\mathcal{M})\pi(\mathcal{M})}{\int p(x|\mathcal{M})\pi(\mathcal{M})d\mathcal{M}} \propto p(x|\mathcal{M})\pi(\mathcal{M}).$$

Note that the formula is s only symbolically true. We cannot write a closed-form for an infinite term. we need process definitions such as Dirichlet process, Indian buffet process, and Gaussian process, which allow us to construct conditional distributions of one instance given all the other instances.