Probabilistic Graphical Models (3): Learning

Qinfeng (Javen) Shi

The Australian Centre for Visual Technologies,
The University of Adelaide, Australia

27 May 2011

Course Outline

Probabilistic Graphical Models:

1. Representation
2. Inference
3. Learning (Today)
4. Sampling-based approximate inference
5. Temporal models
6. . . .

Learning

- Learning graph structure
- Learning parameters in Bayes Net
- Learning parameters in MRFs
- Conditional Random Fields
- Structured Support Vector Machines
- Max Margin Markov Network
- Maximum Entropy Discrimination Markov Networks.
- ...

Learning Graph Structure

- Manually construct graphs (as Bayes nets or MRFs) using relation between independencies and graph (covered in tutorial 1).
- Automatic methods to build the graphs.

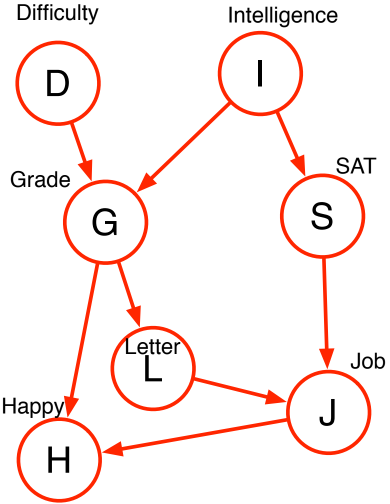
Learning Graph Structure Automatically

- Constraint-based: have a distribution that satisfies a set of independencies, and the goal is to find a graphical model that represents these independencies.
  disadvantage: sensitive to failure of individual independency tests.
- Score-based: design a scoring function, and compute the score for all possible models. Pick a model with highest score.
  disadvantage: enumerating scores for all models is often NP-hard. Resort to heuristic search.
- Bayesian model averaging: ensemble of possible models.
  disadvantage: some has no close-form resorting to approximations.

Learning parameters in Bayes Net

- with discrete variables
  An example will be given.
- with continuos variables (such as kalman filter)
  We will defer this to advance topic dynamic bayes net ($\subset$ temporal models).

An Example


Difficulty
Intelligence
D
I
Grade
SAT
G
S
Letter
L
Job
J
Happy
H

An Example

Y = Yes. N = No.

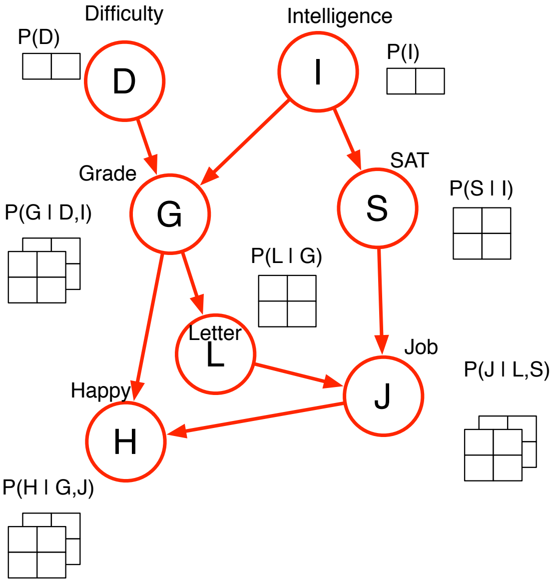
Case	D	I	G	S	L	H	J
1	Y	Y	Y	Y	Y	N	Y
2	N	N	Y	N	N	Y	N
3	Y	N	Y	N	N	Y	N
⋮							

$$P(D = d) = \frac{N_{D=d}}{N_{total}}$$

$$P(G = g|D = d, I = i) = \frac{N_{G=g,D=d,I=i}}{N_{D=d,I=i}}$$

⋮

An Example


Difficulty
P(D)
D
Intelligence
I
P(I)
Grade
G
SAT
S
P(S I I)
P(G I D,I)
P(L I G)
Letter
L
Job
J
P(J I L,S)
Happy
H
P(H I G,J)

An Example

Problems?

The Problems

- not minimise classification error.
- not much flexibility on the features nor the parameters.

Learning parameters in MRFs - EF

Exponential Family (EF) (vector parameter form)

$$P(x|w) = \frac{1}{Z(w)} h(x) \exp\Big(\langle \eta(w), T(x) \rangle \Big), \qquad (1)$$

with
natural parameter $w \in \mathbb{R}^m$,
natural parameter function $\eta(w) : \mathbb{R}^m \to \mathbb{R}^d$,
sufficient statistics $T(x) : \mathcal{X} \to \mathbb{R}^d$,
auxiliary measure $h(x) : \mathcal{X} \to \mathbb{R}^+$,
partition function $Z(w) = \sum_x h(x) \exp\Big(\langle \eta(w), T(x) \rangle \Big)$.
When $\eta(w) = w, m = d$, the EF is said in canonical form.
Special case: normal distribution, binomial distribution ...

Learning parameters in MRFs - ERM

Regularised Empirical Risk Minimisation

$$\min_{\mathbf{w}} J(\mathbf{w}) := \lambda\Omega(\mathbf{w}) + R_{emp}(\mathbf{w}),$$

$$\text{where} \quad R_{emp}(\mathbf{w}) := \frac{1}{m}\sum_{i=1}^{m} \ell(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})$$

is the empirical risk and $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m) \in \mathcal{X} \times \mathcal{Y}$ is the training sample of input-output pairs and $\mathbf{w}$ is a parameter vector. The model complexity is controlled by regulariser $\lambda\Omega(\mathbf{w})$ (with $\lambda > 0$), which usually is (piecewise) differentiable and cheap to compute. For instance, let the regulariser $\Omega(\mathbf{w}) = \frac{1}{2}||\,\mathbf{w}\,||^2$, and the loss $\ell(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})$ be the binary hinge loss, $[1 - \mathbf{y}_i \langle \mathbf{w}, \mathbf{x}_i \rangle]_+$, we recover the soft margin linear SVM.

Probabilistic Approaches - MAP, ML

A likelihood function $\mathcal{L}(\mathbf{w})$ is the modelled probability or density for the occurrence of a sample configuration $(\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)$ given the probability density $\mathbf{P_w}$ parameterised by $\mathbf{w}$. That is,

$$\mathcal{L}(\mathbf{w}) = \mathbf{P_w}\left((\mathbf{x}_1, \mathbf{y}_1), \ldots, (\mathbf{x}_m, \mathbf{y}_m)\right).$$

Maximum a Posteriori (MAP) estimates $\mathbf{w}$ by maximising $\mathcal{L}(\mathbf{w})$ times a prior $P(\mathbf{w})$. That is

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}} \mathcal{L}(\mathbf{w})P(\mathbf{w}). \tag{2}$$

Assuming $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{1 \le i \le m}$ are I.I.D. samples from $\mathbf{P_w}(\mathbf{x}, \mathbf{y})$, (2) becomes

$$\begin{aligned}\mathbf{w}^* &= \operatorname*{argmax}_{\mathbf{w}} \prod_{1 \le i \le m} \mathbf{P_w}(\mathbf{x}_i, \mathbf{y}_i)P(\mathbf{w}) \\ &= \operatorname*{argmin}_{\mathbf{w}} \sum_{1 \le i \le m} -\ln \mathbf{P_w}(\mathbf{x}_i, \mathbf{y}_i) - \ln P(\mathbf{w}).\end{aligned}$$

Maximum Likelihood (ML) is a special case of MAP when $P(\mathbf{w})$ is uniform. Alternatively, one can replace the joint distribution $\mathbf{P_w}(\mathbf{x}, \mathbf{y})$ by the conditional distribution $\mathbf{P_w}(\mathbf{y} \mid \mathbf{x})$ that gives a discriminative model called Conditional Random Fields (CRFs)

Probabilistic Approaches - ME

Maximum Entropy (ME) estimates **w** by maximising the entropy. That is,

$$\mathbf{w}^* = \operatorname*{argmax}_{\mathbf{w}} \sum_{\mathbf{x}\in\mathcal{X},\mathbf{y}\in\mathcal{Y}} -\mathbf{P}_{\mathbf{w}}(\mathbf{x},\mathbf{y}) \ln \mathbf{P}_{\mathbf{w}}(\mathbf{x},\mathbf{y}).$$

Duality between maximum likelihood, and maximum entropy, subject to moment matching constraints on the expectations of features.

Probabilistic Approaches - CRFs - 1

Assume the conditional distribution over $\mathcal{Y} \,|\, \mathcal{X}$ has a form of exponential families, *i.e.*,

$$\mathbf{P}(\mathbf{y} \,|\, \mathbf{x}; \mathbf{w}) = \frac{\exp(\langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}) \rangle)}{Z(\mathbf{w} \,|\, \mathbf{x})}, \tag{3}$$

where

$$Z(\mathbf{w} \,|\, \mathbf{x}) = \sum_{\mathbf{y}' \in \mathcal{Y}} \exp(\langle \mathbf{w}, \Phi(\mathbf{x}, \mathbf{y}') \rangle), \tag{4}$$

and

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{i \in \mathcal{V}} \Phi_1(\mathbf{x}, \mathbf{y}^{(i)}) + \sum_{(ij) \in \mathcal{E}} \Phi_2(\mathbf{x}, \mathbf{y}^{(ij)}). \tag{5}$$

via the Hammersley – Clifford theorem if only node and edge features are considered. More generally speaking, the global feature can be decomposed into local features on cliques (fully connected subgraphs).

Probabilistic Approaches - CRFs - 2

Denote $(\mathbf{x}_1, \ldots, \mathbf{x}_m)$ as $\mathbf{X}$, $(\mathbf{y}_1, \ldots, \mathbf{y}_m)$ as $\mathbf{Y}$. The classical approach is to maximise the conditional likelihood of $\mathbf{Y}$ on $\mathbf{X}$, incorporating a prior on the parameters. This is a Maximum a Posteriori (MAP) estimator, which consists of maximising

$$\mathbf{P}(\mathbf{w} \mid \mathbf{X}, \mathbf{Y}) \propto P(\mathbf{w})\, \mathbf{P}(\mathbf{Y} \mid \mathbf{X}; \mathbf{w}).$$

From the i.i.d. assumption we have

$$\mathbf{P}(\mathbf{Y} \mid \mathbf{X}; \mathbf{w}) = \prod_{i=1}^{m} \mathbf{P}(\mathbf{y}_i \mid \mathbf{x}_i; \mathbf{w}),$$

and we impose a Gaussian prior on $\mathbf{w}$

$$P(\mathbf{w}) \propto \exp\left(\frac{-\|\,\mathbf{w}\,\|^2}{2\sigma^2}\right).$$

Probabilistic Approaches - CRFs - 3

Maximising the posterior distribution can also be seen as minimising the negative log-posterior, which becomes our risk function $R(\mathbf{w} \,|\, \mathbf{X}, \mathbf{Y})$

$$R(\mathbf{w} \,|\, \mathbf{X}, \mathbf{Y}) = -\ln(P(\mathbf{w})\,\mathbf{P}(\mathbf{Y} \,|\, \mathbf{X}; \mathbf{w})) + c$$

$$= \frac{||\,\mathbf{w}\,||^2}{2\sigma^2} - \sum_{i=1}^{m} \underbrace{(\langle \Phi(\mathbf{x}_i, \mathbf{y}_i), \mathbf{w}\rangle) - \ln(Z(\mathbf{w} \,|\, \mathbf{x}_i))}_{:=\ell_L(\mathbf{x}_i, \mathbf{y}_i, \mathbf{w})} + c,$$

where c is a constant and ℓ_L denotes the log loss *i.e.* negative log-likelihood. Now learning is equivalent to

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} R(\mathbf{w} \,|\, \mathbf{X}, \mathbf{Y}).$$

Probabilistic Approaches - CRFs- 4

Above is a convex optimisation problem on $\mathbf{w}$ since $\ln Z(\mathbf{w} \mid \mathbf{x})$ is a convex function of $\mathbf{w}$. The solution can be obtained by gradient descent since $\ln Z(\mathbf{w} \mid \mathbf{x})$ is also differentiable. We have

$$\nabla_{\mathbf{w}} R(\mathbf{w} \mid \mathbf{X}, \mathbf{Y}) = -\sum_{i=1}^{m} (\Phi(\mathbf{x}_i, \mathbf{y}_i) - \nabla_{\mathbf{w}} \ln(Z(\mathbf{w} \mid \mathbf{x}_i)).$$

It follows from direct computation that

$$\nabla_{\mathbf{w}} \ln Z(\mathbf{w} \mid \mathbf{x}) = \mathbb{E}_{\mathbf{y} \sim \mathbf{P}(\mathbf{y} \mid \mathbf{x}; \mathbf{w})}[\Phi(\mathbf{x}, \mathbf{y})].$$

Probabilistic Approaches - CRFs - 5

Since our sufficient statistics $\Phi(\mathbf{x}, \mathbf{y})$ are decomposed over nodes and edges (eq. 5), it is straightforward to show that the expectation also decomposes into expectations on nodes $\mathcal{V}$ and edges $\mathcal{E}$

$$\mathbb{E}_{\mathbf{y} \sim \mathbf{P}(\mathbf{y} \,|\, \mathbf{x};\mathbf{w})}[\Phi(\mathbf{x}, \mathbf{y})] =$$
$$\sum_{i \in \mathcal{V}} \mathbb{E}_{\mathbf{y}^{(i)} \sim \mathbf{P}(\mathbf{y}^{(i)} \,|\, \mathbf{x};\mathbf{w})}[\Phi_1(\mathbf{x}, \mathbf{y}^{(i)})] + \sum_{(ij) \in \mathcal{E}} \mathbb{E}_{\mathbf{y}^{(ij)} \sim \mathbf{P}(\mathbf{y}^{(ij)} \,|\, \mathbf{x};\mathbf{w})}[\Phi_2(\mathbf{x}, \mathbf{y}^{(ij)})],$$

where the node and edge expectations can be computed given $\mathbf{P}(\mathbf{y}^{(i)} \,|\, \mathbf{x}; \mathbf{w})$ and $\mathbf{P}(\mathbf{y}^{(ij)} \,|\, \mathbf{x}; \mathbf{w})$, which can be computed exactly by variable elimination or junction tree or approximately using *e.g.* (loopy) belief propagation. This is the main computational problem with MAP estimation, which can be circumvented through sampling.

Max Margin Approaches

In learning, we look for a F that predicts labels well via

$$\mathbf{y}^* = \max_{\mathbf{y} \in \mathcal{Y}} F(\mathbf{x}_i, \mathbf{y}).$$

Margin: a scoring gap between $F(\mathbf{x}_i, \mathbf{y}_i)$ and best $F(\mathbf{x}_i, \mathbf{y})$ for $\mathbf{y} \neq \mathbf{y}_i$. That is

$$M(\mathbf{x}_i, \mathbf{y}_i) = F(\mathbf{x}_i, \mathbf{y}_i) - \max_{\mathbf{y} \in (\mathcal{Y} - \mathbf{y}_i)} F(\mathbf{x}_i, \mathbf{y})$$

Max Margin Approaches- Structured SVM - 1

$$\min_{\mathbf{w},\xi} \frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{m}\xi_i \quad \text{s.t.} \tag{6a}$$

$$\forall i, \mathbf{y}, \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y})\rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}) - \xi_i. \tag{6b}$$

or its dual problem in kernels $k(,) := \langle \Phi, \Phi \rangle$:

$$\max_{\alpha} \frac{1}{2} \sum_{i,j,\mathbf{y},\mathbf{y}'} \alpha_{i\mathbf{y}}\alpha_{j\mathbf{y}'} \langle \Phi(\mathbf{x}_i, \mathbf{y}), \Phi(\mathbf{x}_j, \mathbf{y}')\rangle - \sum_{i,\mathbf{y}} \Delta(\mathbf{y}_i, \mathbf{y})\alpha_{i\mathbf{y}}$$

$$\forall i, \mathbf{y}, \sum_{\mathbf{y}} \alpha_{i\mathbf{y}} \leq C, \ \alpha_{i\mathbf{y}} \geq 0.$$

Max Margin Approaches- Structured SVM - 2

Cutting plane method needs to find the label for the most violated constraint in (6b)

$$\mathbf{y}_i^\dagger = \operatorname*{argmax}_{\mathbf{y} \in \mathcal{Y}} \Delta(\mathbf{y}_i, \mathbf{y}) + \langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}) \rangle . \tag{7}$$

With $\mathbf{y}_i^\dagger$, one can solve following relaxed problem (with much fewer constraints)

$$\min_{\mathbf{w},\xi} \frac{1}{2} \|\mathbf{w}\|^2 + C \sum_{i=1}^{m} \xi_i \quad \text{s.t.} \tag{8a}$$

$$\forall i, \left\langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y}_i^\dagger) \right\rangle \geq \Delta(\mathbf{y}_i, \mathbf{y}_i^\dagger) - \xi_i . \tag{8b}$$

Max Margin Approaches- Structured SVM - 3

Input: data $\mathbf{x}_i$, labels $\mathbf{y}_i$, sample size m
Initialise $S_i = \emptyset$ for all i, and $\mathbf{w}_0 = 0$ or a random vector.
repeat
  for $i = 1$ **to** m **do**
    $\mathbf{w}_t = \sum_i \sum_{y \in S_i} \alpha_{i\mathbf{y}} \Phi(\mathbf{x}_i, \mathbf{y})$
    $\mathbf{y}_i^\dagger = \operatorname{argmax}_{\mathbf{y} \in \mathcal{Y} - \mathbf{y}_i} \langle \mathbf{w}_t, \Phi(\mathbf{x}_i, \mathbf{y}) \rangle + \Delta(\mathbf{y}_i, \mathbf{y})$,
    $\xi_i = \left[\Delta(\mathbf{y}_i, \mathbf{y}) + \left\langle \mathbf{w}_t, \Phi(\mathbf{x}_i, \mathbf{y}_i^\dagger) - \Phi(\mathbf{x}_i, \mathbf{y}_i) \right\rangle \right]_+$,
    if $\xi_i > 0$ **then**
      Increase constraint set $S_t \leftarrow S_t \cup \mathbf{y}_t^\dagger$
    end if
  end for
  $\alpha \leftarrow$ optimise dual QP with constraint set S_t.
until S has not changed in this iteration

Max Margin Approaches- Max Margin Markov Net - 1

Max Margin Markov Network (M3N) transform the structured SVM dual into

$$\max_{\alpha} \; -\frac{1}{2}\|\sum_{i,\mathbf{y}} \alpha_{i\,\mathbf{y}}[\Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y})]\|^2 + \sum_{i,\mathbf{y}} \Delta(\mathbf{y}_i, \mathbf{y})\alpha_{i\,\mathbf{y}}$$

$$\forall i, \mathbf{y} \; \sum_{\mathbf{y}} \alpha_{i\,\mathbf{y}} = C, \; \alpha_{i\,\mathbf{y}} \geq 0.$$

Now the dual variable $\frac{\alpha_{i\,\mathbf{y}}}{C}$ can be viewed as a distribution over **y** given **x**. Thus the dual object becomes

$$\max_{\alpha} \; -\frac{1}{2}\|\sum_{i} \mathbb{E}_{\mathbf{y}\sim\alpha_{i\,\mathbf{y}}}[\Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y})]\|^2 + \sum_{i} \mathbb{E}_{\mathbf{y}\sim\alpha_{i\,\mathbf{y}}} \Delta(\mathbf{y}_i, \mathbf{y}) \tag{9}$$

$$\forall i, \mathbf{y} \; \sum_{\mathbf{y}} \frac{\alpha_{i\,\mathbf{y}}}{C} = 1, \; \alpha_{i\,\mathbf{y}} \geq 0.$$

Max Margin Approaches- Max Margin Markov Net - 2

Denote $\mathbf{y} \sim \mathbf{y}^{(a)}$ as the value of the component $\mathbf{y}^{(a)}$ is consistent with that in $\mathbf{y}$. Decomposing global features into local node and edge features as (5), we get

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}\sim\alpha_{i\,\mathbf{y}}} \Phi(\mathbf{x}_i, \mathbf{y}) &= \sum_{\mathbf{y}} \alpha_{i\,\mathbf{y}} \Phi(\mathbf{x}_i, \mathbf{y}) \\
&= \sum_{\mathbf{y}} \alpha_{i\,\mathbf{y}} \sum_{a \in \mathcal{V}} \Phi_1(\mathbf{x}_i, \mathbf{y}^{(a)}) + \sum_{(ab)\in\mathcal{E}} \Phi_2(\mathbf{x}_i, \mathbf{y}^{(ab)}) \\
&= \sum_{a\in\mathcal{V}} \sum_{\mathbf{y}:\mathbf{y}\sim\mathbf{y}^{(a)}} \alpha_{i\,\mathbf{y}}(\mathbf{y}) \Phi_1(\mathbf{x}_i, \mathbf{y}^{(a)}) + \sum_{(ab)\in\mathcal{E}} \sum_{\mathbf{y}:\mathbf{y}\sim\mathbf{y}^{(ab)}} \alpha_{i\,\mathbf{y}}(\mathbf{y}) \Phi_2(\mathbf{x}_i, \mathbf{y}^{(ab)}) \\
&= \sum_{a\in\mathcal{V}} \sum_{\mathbf{y}^{(a)}} \mu_{\mathbf{x}_i}(\mathbf{y}^{(a)}) \Phi_1(\mathbf{x}_i, \mathbf{y}^{(a)}) + \sum_{(ab)\in\mathcal{E}} \sum_{\mathbf{y}^{(ab)}} \mu_{\mathbf{x}_i}(\mathbf{y}^{(ab)}) \Phi_2(\mathbf{x}_i, \mathbf{y}^{(ab)}),
\end{aligned}$$

where marginals

$$\mu_{\mathbf{x}_i}(\mathbf{y}^{(a)}) = \sum_{\mathbf{y}:\mathbf{y}\sim\mathbf{y}^{(a)}} \alpha_{i\,\mathbf{y}}(\mathbf{y}), \quad \mu_{\mathbf{x}_i}(\mathbf{y}^{(ab)}) = \sum_{\mathbf{y}:\mathbf{y}\sim\mathbf{y}^{(ab)}} \alpha_{i\,\mathbf{y}}(\mathbf{y}).$$

Similarly if $\Delta(\mathbf{y}_i, \mathbf{y}) = \sum_{a\in\mathcal{V}} \Delta(\mathbf{y}_i, \mathbf{y}^{(a)})$, then

$$\mathbb{E}_{\mathbf{y}\sim\alpha_{i\,\mathbf{y}}} \Delta(\mathbf{y}_i, \mathbf{y}) = \sum_{a\in\mathcal{V}} \mu_{\mathbf{x}_i}(\mathbf{y}^{(a)}) \Delta(\mathbf{y}_i, \mathbf{y}^{(a)}).$$

To ensure the marginals resulting from a valid distribution $\alpha_{i\,\mathbf{y}}(\mathbf{y})$, one must ensure following consistency constraint

$$\sum_{\mathbf{y}^{(b)}} \mu_{\mathbf{x}_i}(\mathbf{y}^{(ab)}) = \mu_{\mathbf{x}_i}(\mathbf{y}^{(a)}), \forall (a, b) \sim \mathcal{E}, \forall i.$$

Max Margin Approaches- MED

Maximum Entropy Discrimination (MED) that maximises the entropy — or minimises the KL divergence $\mathrm{KL}(Q(\mathbf{w})||P(\mathbf{w})) = \int \ln \frac{Q(\mathbf{w})}{P(\mathbf{w})} dQ(\mathbf{w})$ between the posterior Q and the prior P — with a constraint that the expected margin with respect to the posterior $Q(\mathbf{w})$ over model parameter $\mathbf{w}$ is not less than certain threshold (that is a weighted max margin constraint or weighted hinge loss via the posterior) for binary classification.

Max Margin Approaches- MEDN

Maximum Entropy Discrimination Markov Networks (MEDN)

$$\min_{\mathbf{w},\xi} \text{KL}(Q(\mathbf{w})||P(\mathbf{w})) + C\sum_{i=1}^{m}\xi_i \quad \text{s.t.}$$

$$\forall i, \mathbf{y}, \int \Big[\langle \mathbf{w}, \Phi(\mathbf{x}_i, \mathbf{y}_i) - \Phi(\mathbf{x}_i, \mathbf{y})\rangle - \Delta(\mathbf{y}_i, \mathbf{y}) \Big] dQ(\mathbf{w}) \geq -\xi_i.$$

Again $\mathbf{y}$ can be replaced by the most-violated $\overline{\mathbf{y}}_i$. Apparently letting $\mathbf{y}$ be scalar y, MEDN recovers MED. Letting $P(\mathbf{w})$ be a zero mean, identity variance gaussian over $\mathbf{w}$, MEDN recovers M3N.

Recap

Introduction to Probabilistic Graphical Models

1. representation (tutorial 1)
2. inference (tutorial 2)
3. learning (tutorial 3, today)

Next tutorial:
Particle (or sampling)-based approximate inference
(importance sampling, markov chain monte carlo)