Log-linear MRFs: Ising, Boltzmann, Deep Belief, Metric

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

- Log-linear MRF Applications
  - Ising Model
  - Boltzmann Distribution
  - Energy Based Model
  - Boltzmann Machine
    - Restricted Boltzmann Machine
    - Deep Belief Networks
  - Metric MRF

General Log-linear model with features

- A distribution P is a log-linear model over $\mathcal{H}$ if

$$P(X_1,..,X_n)=\frac{1}{Z}\exp\left[-\sum_{i=1}^{k} w_i f_i(D_i)\right]$$

Note that k is the no of features Not no of subgraphs

  - Can have several functions over same scope
  - Each term is an energy function
- Equivalent to Gibbs distribution

$$P_\Phi(\mathrm{X}_1,..\mathrm{X}_\mathrm{n})=\frac{1}{Z}\tilde{P}(\mathrm{X}_1,..\mathrm{X}_\mathrm{n}) \text{ where } \tilde{P}(\mathrm{X}_1,..\mathrm{X}_\mathrm{n})=\prod_{i=1}^{m}\phi_i(D_i)$$

is an unnomalized measure and $Z=\sum_{\mathrm{X}_1,..\mathrm{X}_\mathrm{n}}\tilde{P}(\mathrm{X}_1,..\mathrm{X}_\mathrm{n})$

- Rewrite factor $\phi(D)$ as $\phi(D)=\exp(-\varepsilon(D))$

  where $\varepsilon(\mathrm{D}) = -\ln\,\phi(D)$ is the *energy* function

Example of Markov Network: Ising Model

- Pairwise and single potentials
  - Edge potentials $\varepsilon_{ij}(x_i,x_j)=-w_{ij}x_ix_j$
    - Contributes w_{ij} when $X_i=X_j$, same, and $-w_{ij}$ otherwise
  - Node potentials are u_i
- Probability distribution (energy function)

$$P(\xi)=\frac{1}{Z}\exp\left(-\sum_{i<j} w_{ij}x_ix_j-\sum_i u_ix_i\right)$$

ξ ε $Val(\boldsymbol{X})$ is a full assignment of the variables

- Edge/node potentials also arise in continuous
  - Gaussian quadratic form: $p(\mathbf{x})=\frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left[(\mathbf{x}-\boldsymbol{\mu})^t\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right]$
    - Using precision matrix $J=\Sigma^{-1}$

$$p(\mathrm{x}) \quad \alpha \quad \exp\left(-\frac{1}{2}\mathrm{x}^tJ\mathrm{x}+(J\boldsymbol{\mu})^t\,\mathrm{x}\right)$$

With $\mathrm{h}=J\mu$, terms involving x_i $-\frac{1}{2}J_{i,i}x_i^2+h_ix_i$ terms involving pairs $-\frac{1}{2}\left[J_{i,j}x_ix_j+J_{j,i}x_jx_i\right]=4J_{i,j}x_ix_j$

Ising Model in Statistical Physics

- Energy of interacting atoms
  - Determined from their spin
    - Atom's spin is sum of its electron spins
    - Each atom associated with binary random variable
      - $X_i \in \{+1,-1\}$ whose value is direction of atom's spin

When $w_{ij}>0$ model prefers aligned spins: ferromagnetism
$w_{ij}<0$: antiferromagnetic
$w_{ij}=0$: non-interacting

- Energy function parametric form

$$\varepsilon_{i,j}(x_i,x_j) = -w_{ij}x_ix_j$$

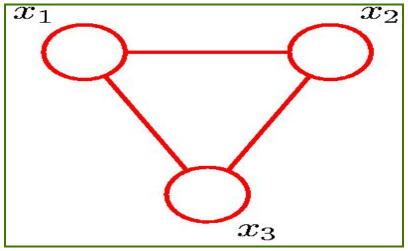
  - Symmetric in X_i, X_j: note scope is pairwise
  - Makes contribution w_{ij} to energy when $X_i=X_j$ (same spin)
  - $-w_{ij}$ otherwise

- Probability distribution over atoms (energy function)

$$\boxed{P(\xi) = \frac{1}{Z}\exp\left(-\sum_{i<j} w_{ij}x_ix_j - \sum_i u_ix_i\right)}$$

$\xi \varepsilon \; Val(\boldsymbol{X})$

Ising Model studies



- To answer a variety of questions
  - Usually as the no. of atoms (variables) goes to infinity

$$P(\xi) = \frac{1}{Z}\exp\left(-\sum_{i<j} w_{ij}x_i x_j - \sum_i u_i x_i\right)$$

- Inference problems, e.g.,
  - Determine probability of a configuration where majority of spins are $+1$ (or -1) versus more mixed ones
    - Answer depends on strength of interactions w_{ij}
    - e.g., Multiply all weights by temperature parameter
  - Many other problems investigated extensively
    - Answers known--some even analytically

Boltzmann Distribution

- Variant of Ising Model
- Variables X_i have value $\{0,1\}$ instead of $\{+1,-1\}$
  - Energy function has same parametric form

    $\varepsilon_{ij}(x_i,x_j) = -w_{ij}x_ix_j$
  - Nonzero contribution $-w_{ij}$ from edge X_i-X_j only when $X_i=X_j=1$
    - Ising model has contribution w_{ij} when variables are same and $-w_{ij}$ when they are different
- Has the same energy function as Ising model

$$\boxed{P(\xi) = \frac{1}{Z}\exp\left(-\sum_{i<j} w_{ij}x_ix_j - \sum_i u_ix_i\right)}$$

Mapping 0 to -1

Boltzmann Distrib. & Statistical Mechanics

- Boltzmann Probability distribution

$$P(\text{state}) \propto \exp[-E\,]/kT$$

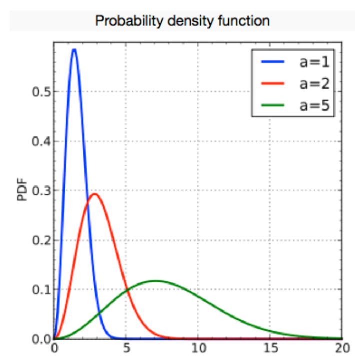
Where E is state energy (varies from state to state)

- kT is a constant of the distribution
  - k = Boltzmann's constant, T = absolute temperature
- Ratio over two states depends on energy difference

$$P(\text{state}_1)|P(\text{state}_2)=\exp[E_2\text{-}E_1]/kT$$

- Later investigated by Josiah Gibbs
  - Boltzmann distribution also known as Gibbs measure
- Maxwell-Boltzmann distribution
  - Is χ^2 with 3 degrees of freedom

$$f(v) = \sqrt{\left(\frac{m}{2\pi kT}\right)^3} 4\pi v^2 e^{-\frac{mv^2}{2kT}}$$



Boltzmann Distribution resembles neuron

- Neuron output is a stochastic function of its connected neighbors
  - Probability distribution of each variable X_i given assignment of neighbors X_j is $\sigma(z)$ where

$$z = -\left(\sum_j w_{ij} x_j\right) - w_i$$

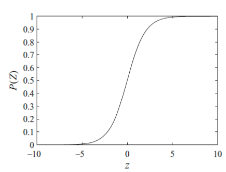
where $\sigma(z) = [1/1+\exp(-z)]$ is a value in $[0,1]$

Conditional Unnormalized

$$\tilde{P}(y=1 \mid \mathrm{x}) = \exp\left\{w_0 + \sum_{i=1}^{d} w_i x_i\right\} \qquad \tilde{P}(y=0 \mid \mathrm{x}) = \exp\{0\} = 1$$

Normalized

$$\boxed{P(y=1 \mid \mathrm{x}) = sigmoid\left\{w_0 + \sum_{i=1}^{d} w_i x_i\right\} \quad \text{where } sigmoid(z) = \frac{e^z}{1+e^z} = \frac{1}{1+e^{-z}}}$$



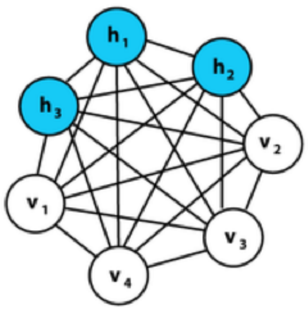
Logistic Regression

Z has term 1 because $P^{\sim}(y=0 \mid \mathrm{x}) = 1$

Boltzmann distribution, Sigmoidal neuron and Logistic Regression have the same form

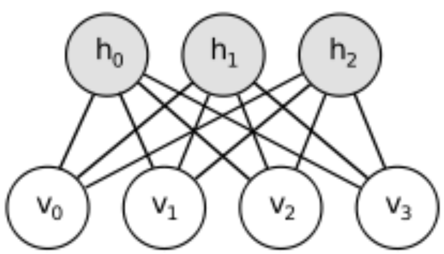
Boltzmann Machine

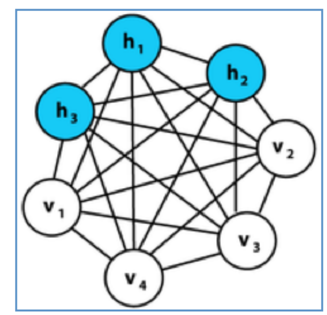
- A form of Energy Based Model
- Structure of a recurrent neural network (RNN):
  - one where there are directed cycles
  - Unlike feed-forward neural networks RNN can use internal memory to process arbitrary sequences
    - Can process time-varying real-valued inputs
  - Have nodes which are inputs, hidden and outputs
- Boltzmann machines are a type of RNN



Restricted Boltzmann Machine

- RBM is a special case of Boltzmann machines and Markov networks
- No visible-visible and hidden-hidden connections– Bipartite graph





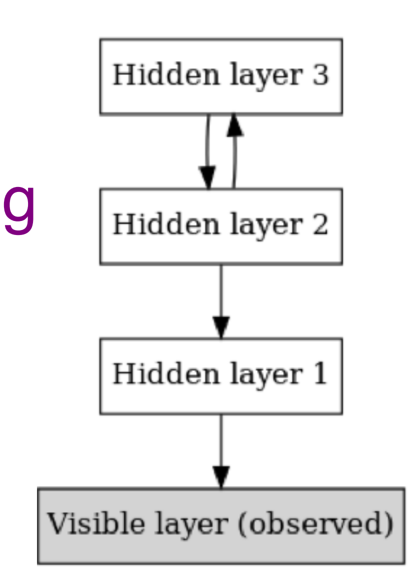
Not an RBM

- Used to learn features for input to neural networks in Deep Learning

Deep Belief Networks (DBNs)

- Consist of several layers of RBMs
  - Stacking RBMs
    - Fine tuning resulting deep network using gradient descent and back-propagation
- DBNs are Generative Models
  - Provide estimates of both

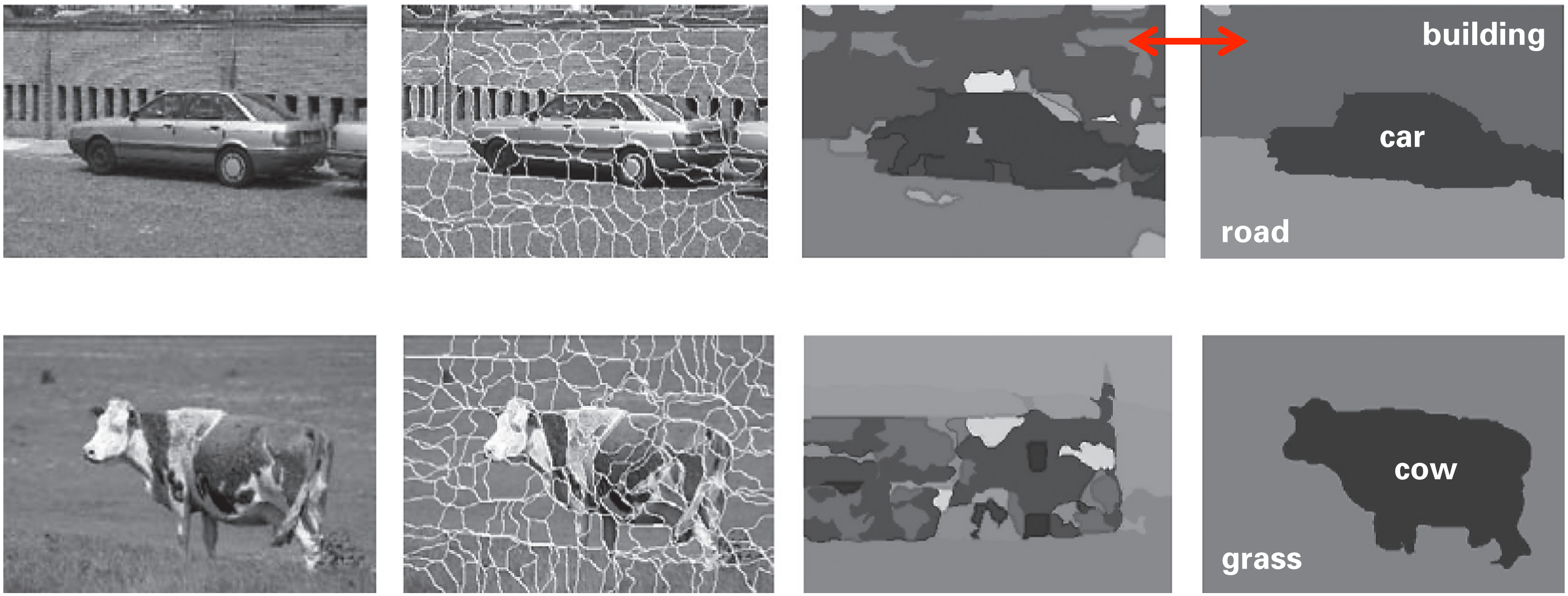
    $p(x|C_k)$ and $p(C_k|x)$
  - Conventional neural networks are discriminative
    - Directly estimate $p(C_k|x)$



Metric MRF for Labeling

- Task:
  - Graph with nodes $X_1,..X_n$, edges $\boldsymbol{E}$
  - Assign to each X_i a label in $V=\{v_1,..v_k\}$
    - E.g., labeling super-pixels in image
  - Each node, in isolation, has a preferred label
    - E.g., color specifies a label
  - However, we want *smoothness* constraint over neighbors
    - Neighboring nodes should have “similar” values

Importance of Modeling Correlations between superpixels



Original image

Oversegmented image-superpixels Each superpixel is a random variable

Classification using node potentials alone-each superpixel classified independently

Segmentation using pairwise Markov Network encoding interactions between adjacent superpixels

Solution for Labeling

- Solution:
  - Encode node preferences as edge potentials
  - Smoothness preferences as edge potentials
  - Encode model in negative log-space, using energy functions
- Energy function

$$E(x_1,..x_n)=\sum_i \varepsilon_i(x_i)+\sum_{i,j\in E}\varepsilon_{ij}(x_i,x_j)$$

  - For MAP objective, ignore partition function
  - Goal: Minimize the energy (MAP objective)

$$\arg\min_{x_1,..x_n} E(x_1,..x_n)$$

  - How to define smoothness? Next.

Smoothness for Metric MRF

- Many variants
- Simplest one is a variant of Ising model

$$\varepsilon_{i,j} = \begin{cases} 0 & x_i = x_j \\ 1 & x_i \neq x_j \end{cases}$$

  - for $\lambda_{ij} \geq 0$
- In this model:
  - Lowest pairwise energy (0) when neighbors have same value
  - Higher energy otherwise λ_{ij}

Generalizations of Smoothness for Metric MRF

1. Potts model (when there are more than two labels)
2. Distance Function on labels
   - Prefer neighboring nodes to have labels smaller distance apart
   - Metric MRF
     - Need a metric $\mu(v_k, v_l)$ on labels

Metric Requirement

- Function μ: $V \times V \rightarrow [0,\infty)$
  - Reflexivity, symmetry and triangle inequality
    - Semi-metric if triangle inequality is violated
- Metric MRF
    - Define $\varepsilon_{i,j}(v_k,v_l) = \mu(v_k,v_l)$
    - Where μ is a metric (or semi-metric)
      - Assume same for all variables
        - » Simplifies no. of parameters needed
        - » Usually holds in practice
    - Example metric p-norm: $\varepsilon(x_i,x_j) = \min(c||x_i\text{-}x_j||_p, \text{dist}_{\max})$
- Metric interactions arise frequently
    - Plays important role in computer vision