Graphical Models

Instructor: Pranjal Awasthi
So Far

• \( y = f^*(\hat{X}) \)
  - Use train set \((\hat{X}_1, y_1), (\hat{X}_2, y_2), \ldots, (\hat{X}_m, y_m)\) to build a predictor

• What if want to predict multiple \( y'_i \)'s?
  - \( y_1 = f_1^*(\hat{X}), y_2 = f_2^*(\hat{X}), \ldots, y_k = f_k^*(\hat{X}) \)
  - \( y'_i \)'s are correlated
Structured Prediction

\[ X = \text{pixel intensities/RGB values} \]
\[ Y_i = \text{label of pixel } i \]
\[ y_1 = f_1(X), y_2 = f_2(X) \]

Input Image

Segmentation Maps
Structured Prediction

\[ X = \text{sentence} \]
\[ Y = \text{part of speech tag} \]
Structured Prediction
Structured Prediction

- \( y = f^* (\vec{X}) \)
  - Use train set \((\vec{X}_1, y_1), (\vec{X}_2, y_2), ..., (\vec{X}_m, y_m)\) to build a predictor

- What if want to predict multiple \( y_i' \)'s?
  - \( y_1 = f_1^* (\vec{X}), y_2 = f_2^* (\vec{X}), ..., y_k = f_k^* (\vec{X}) \)
  - \( y_i' \)'s are correlated

- Trivial solutions
  - Solve each problem independently (ignores useful info)
  - Solve a giant multiclass problem (not very efficient, if \( y_i' \in \Sigma_1^{k-1} \) need \( 2^k \) labels)
Graphical Models

• An elegant approach for structured prediction
• High level idea:
  – Model the joint density \( P(\vec{X}, y_1, y_2, \ldots y_k) = P(\vec{X}, \vec{Y}) \)
  – Model captures dependencies and correlations
  – Learn the model from data
  – Compute \( P(\vec{Y}|\vec{X}) \) to make predictions
• Can do more
  – Might want to just model \( P(\vec{X}) \)
  – For instance, build a model for generating images, speech etc.
Graphical Models

• Model \( P(\vec{X}, \vec{Y}) \) or \( P(\vec{X}) \)

• Notation:
  – From now on will just use \( P(\vec{X}) = P(X_1, X_2, ..., X_n) \)

• How to capture dependencies among \( X_i's \)?
  – Given by a graph \( G \) over \( X_1, X_2, ... X_n \)
  – Require \( P(\vec{X}) \) to be “consistent” with \( G \).
Graphical Models

\[ P(X_1, X_2, X_3, X_4, X_5) \]

- \( G \) can be directed or undirected
- \( P() \) is consistent with \( G \)
  - Satisfies Markov constraints

\( \text{will study undirected for now} \)
Graphical Models

- Markov Constraints
  - For each $X_i$, let $N(X_i) = \{X_j: X_i \sim X_j \text{ in } G\}$
  - Then $X_i$ is independent of the rest of the variables given $N(X_i)$
  - $P(X_i \mid X_1, X_2, ... X_{i-1}, X_{i+1}, X_n) = P(X_i \mid N(X_i))$
Graphical Models

\[ P(X_1, X_2, X_3, X_4, X_5) \]

Given \( X_5 \), \( X_1 \cdots X_4 \) are independent.
Graphical Models

$P(X_1, X_2, X_3, X_4, Y)$

This is the Naive Bayes classifier. The $X_i$'s are independent given $Y$. 
Graphical Models

Case i: Model $P(X_1, \ldots, X_n)$ when $X_i$'s are independent.

Associated graph:

Case ii: Model $P(X_1, \ldots, X_n)$ when no assumption on $X_i$'s.

Associated graph:

(clique or complete graph)
Markov Random Fields (MRFs)

Definition:
If $P()$ is consistent with a graph $G$ then $P + G$ is a Markov random field.
Markov Random Fields (MRFs)

Theorem [Hammersley, Clifford '72]:
If $P + G$ is an MRF then $P(\vec{X}) \propto e^{\sum C \theta_c \phi_c(\vec{X})}$, where the sum is over cliques of $G$, and $\phi_c(\vec{X})$ is a function defined over variables in $C$. 

\[ P(X_1, X_2, X_3, X_4, X_5) \]

\[ \Theta_c = \text{Parameters of model} \]
\[ \phi_c(X) = \text{local functions} \]
Markov Random Fields (MRFs)

\[ P(X_1, X_2, X_3, X_4, X_5) \]

\[ P(X_1, X_2, \ldots, X_5) \propto e^{\sum_{i=1}^{5} \Theta_i \phi(x_i) + \sum_{j=1}^{4} \Theta_{ij} \phi(x_j x_5)} \]

\[ = \prod_{i=1}^{5} e^{\Theta_i \phi(x_i)} \prod_{j=1}^{4} e^{\Theta_{ij} \phi(x_j x_5)} \]

\[ \propto P(X_5) \prod_{j=1}^{4} P(x_j | x_5) \]
Markov Random Fields (MRFs)

Theorem [Hammersley, Clifford’72]:
If $P + G$ is an MRF then $P(\tilde{X}) \propto e^{\sum_{c} \theta_c \phi_c(\tilde{X})}$, where the sum is over cliques of $G$, $\theta_c \in \mathbb{R}$ and $\phi_c(\tilde{X})$ is a function defined over variables in $C$.

$$P(\tilde{X}) = e^{\sum_{c} \theta_c \phi_c(\tilde{X})} - A(\theta)$$

Gibbs Distribution

$$e = \frac{A(\theta)}{\sum_{\tilde{X}} e^{\sum_{c} \theta_c \phi_c(\tilde{X})}}$$

Normalization factor
Examples (Ising Model)

\[ X_i \in \{+1, -1\} \]

Local functions:
\[ \phi_i(x_i) = x_i \]
\[ \phi_{i,j}(x_i x_j) = x_i x_j \]

Cliques are nodes and edges:
\[ \mathcal{C} = \{i, j\} \]

\[ P(\mathcal{C}) \propto e^{\sum_{i \in \mathcal{C}} x_i + \sum_{i,j \in \mathcal{C}} \theta_{i,j} x_i x_j} \]
Examples (Hidden Markov Model)

Predict \( x_i \)'s given \( y_i \)'s

\[
P(x, y) \propto p(x_0) p(y_0|x_0) p(x_1|x_0) p(y_1|x_1) \cdots
\]

\[
= p(x_0) p(y_0|x_0) \prod p(x_{i-1}|x_i) p(y_i|x_i)
\]

Product of factors over nodes & edges.
$\alpha$ and $\beta$ are parameters.

$\Theta \in \mathbb{R}^k$

$\theta_i$: Proportion in topic $i$.

Document generation process:

1. Pick $\Theta$ from Dirichlet($\alpha$)
2. Pick $z \in \{1, \ldots, k\}$ according to $\Theta$
3. Given $z = j$, pick word $w$ from the distribution $\beta_j$.

$p(\Theta, Z, w) \propto p(\Theta) p(z|\Theta) p(w|z)$

Product of local functions
Examples (Gaussian Graphical Model)

\[ p(x_1, \ldots, x_n) \propto e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)} \]

- \( \mu \) \text{ mean}
- \( \Sigma^{-1} \) \text{ inverse covariance}

\( \Sigma_{ij} \neq 0 \) only if \( i \neq j \) in \( G \).
Graphical Models

• Learning problems
  – Given $\vec{X}_1, \vec{X}_2, ... \vec{X}_m$ learn the graph structure $G$
  – Given $\vec{X}_1, \vec{X}_2, ... \vec{X}_m$ and graph $G$, learn the parameters

• Inference
  – Given $G + P(\vec{X})$
    • Compute $P(X_i)$
    • Compute $P(X_i \mid X_j)$
    • Generate sample from $P(\vec{X})$
    • Compute $\text{argmax}_{\vec{X}} P(\vec{X})$
Inference in Graphical Models

• Inference
  – Given $G + P(\vec{X})$
    • Compute $P(X_i)$
    • Compute $P(X_i | X_j)$

\[
p(x_1) = \sum_{x_2 \ldots x_n} p(x_1, x_2, \ldots, x_n)
\]

\[
p(x_i | x_j) = \frac{p(x_i, x_j)}{p(x_j)}
\]
Inference in Graphical Models

• Inference
  – Given $G + P(\bar{X})$
    • Generate sample from $P(\bar{X})$
      \[ P(X) = P(X_1)P(X_2 \ldots X_n | X_1) \]
      - Generate a sample from $P(X_1)$
      - Fix $X_1$ & generate a sample from $P(X_2 \ldots X_n | X_1)$
Inference in Graphical Models

• Inference
  – Given $G + P(\tilde{X})$
    • Compute $\arg\max_{\tilde{X}} P(\tilde{X})$

$$\arg\max_{\tilde{X}} f(\tilde{X}) = \arg\max_{\tilde{X}} \sum_{\theta \in \Theta} \phi_{\tilde{X}}(\tilde{X})$$

$P(\tilde{X}) \propto \sum_{\theta \in \Theta} \phi_{\tilde{X}}(\tilde{X})$

$might \ not \ be \ a \ concave \ function \ in \ general.$
Inference in Graphical Models

• Techniques
  – Junction Tree Algorithm
  – Markov Chain Monte Carlo Methods
  – Variational Methods

Gives exact solution but not always polynomial time.