Probabilistic Graphical Models

Lecture 5: Conditional Random Fields

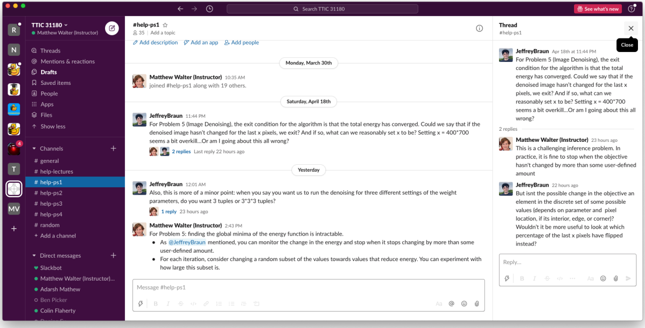
Matthew Walter

TTI-Chicago

April 21, 2020

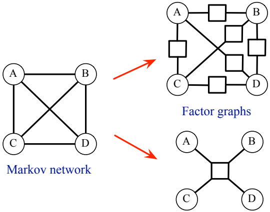
Problem Set 1 and Slack

- See Canvas for updates on Problem Set 1
- Don't forget to join the Slack channel #help-ps1
- See Slack channel for discussion and hints



Factor Graphs (Revisited)

- The Markov network H does not make explicit the structure of the distribution, i.e., maximum cliques vs. complete graph subsets
- A **factor graph** is a bipartite undirected graph with variable nodes (oval) and factor nodes (square). Edges exist only between variable nodes and factor nodes
- Each factor node is associated with a single potential, the scope of which is the variables that are the factor's neighbors



- Distribution is the same as an MRF, just a different data structure

Boltzmann Distribution (Revisited)

- We can rewrite a factor $\phi(\boldsymbol{D}) : \text{Val}(\boldsymbol{D}) \rightarrow \mathbb{R}^+$ as

$$\phi(\boldsymbol{D}) = \exp(-\psi(\boldsymbol{D}))$$

where $\psi(\boldsymbol{D}) = -\log \phi(\boldsymbol{D})$ is the **energy function** (not surprisingly, derived from statistical physics)

Boltzmann Distribution (Revisited)

- We can rewrite a factor $\phi(\boldsymbol{D}) : \text{Val}(\boldsymbol{D}) \rightarrow \mathbb{R}^+$ as

$$\phi(\boldsymbol{D}) = \exp(-\psi(\boldsymbol{D}))$$

  where $\psi(\boldsymbol{D}) = -\log \phi(\boldsymbol{D})$ is the **energy function** (not surprisingly, derived from statistical physics)
- The factorized distribution then becomes (Boltzmann distribution)

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{k=1}^{K} \exp\left(-\psi_k(\boldsymbol{D}_k)\right) = \frac{1}{Z} \exp\left(-\sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)\right)$$

- $\sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)$ is referred to as the "free energy"

Boltzmann Distribution (Revisited)

- We can rewrite a factor $\phi(\boldsymbol{D}) : \text{Val}(\boldsymbol{D}) \rightarrow \mathbb{R}^+$ as

$$\phi(\boldsymbol{D}) = \exp(-\psi(\boldsymbol{D}))$$

  where $\psi(\boldsymbol{D}) = -\log \phi(\boldsymbol{D})$ is the **energy function** (not surprisingly, derived from statistical physics)
- The factorized distribution then becomes (Boltzmann distribution)

$$P(X_1, \ldots, X_n) = \frac{1}{Z} \prod_{k=1}^{K} \exp\left(-\psi_k(\boldsymbol{D}_k)\right) = \frac{1}{Z} \exp\left(-\sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)\right)$$

- $\sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)$ is referred to as the "free energy"
- Gives rise to interpretation as energy minimization

$$\arg\max\ P(X_1, \ldots, X_n) = \arg\min\ \sum_{k=1}^{K} \psi_k(\boldsymbol{D}_k)$$

Log-Linear Markov Networks with Features (Revisited)

- A **feature** is a function $f : \text{Val}(\boldsymbol{D}_i) \rightarrow \mathbb{R}$
- A distribution P is a **log-linear model** over a Markov network H if it is associated with
  - A set of features $\boldsymbol{F} = \{f_1(\boldsymbol{D}_1), \ldots, f_K(\boldsymbol{D}_M)\}$ where $\boldsymbol{D}_i$ is a complete subgraph in H
  - A set of weights $\{w_1, \ldots, w_M\}$

  such that

$$P(X_1, \ldots, X_n) \propto \exp\left(-\sum_{i=1}^{M} w_i f_i(\boldsymbol{D}_i)\right)$$

Log-Linear Markov Networks with Features (Revisited)

- A **feature** is a function $f : \text{Val}(\boldsymbol{D}_i) \rightarrow \mathbb{R}$
- A distribution P is a **log-linear model** over a Markov network H if it is associated with
  - A set of features $\boldsymbol{F} = \{f_1(\boldsymbol{D}_1), \ldots, f_K(\boldsymbol{D}_M)\}$ where $\boldsymbol{D}_i$ is a complete subgraph in H
  - A set of weights $\{w_1, \ldots, w_M\}$

  such that

$$P(X_1, \ldots, X_n) \propto \exp\left(-\sum_{i=1}^{M} w_i f_i(\boldsymbol{D}_i)\right)$$

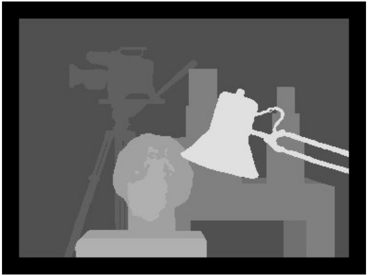
- Multiple features can be defined over the same variables
- Log-linear model can represent tabular potentials, but is more general
- Features and weights can be reused for different factors
- Historically, features designed by hand and weights learned from data

Example: Stereo Vision

Left Image

Right Image

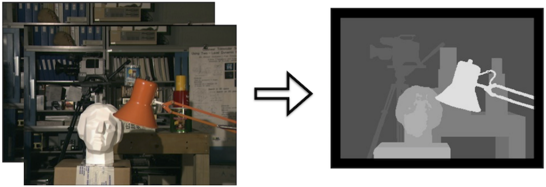
Example: Stereo Vision



Disparity Image

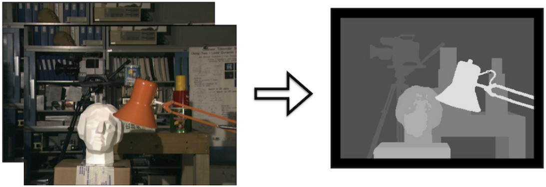
- Dense stereo reconstruction: For every pixel (i_l, j_l) in left image, we are interested in the image-space distance (**disparity**) y_{i_l,j_l} to the corresponding pixel (i_r, j_r) in the right image

Example: Stereo Vision



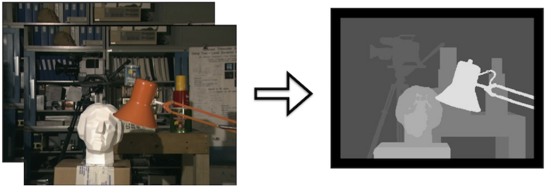
- We could model this as a Markov random field

Example: Stereo Vision



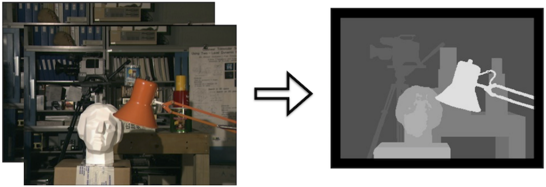
- We could model this as a Markov random field
- Joint distribution $P(\boldsymbol{Y}, \boldsymbol{X})$ over disparity $\boldsymbol{Y}$ and pixel intensities $\boldsymbol{X}$:

Example: Stereo Vision



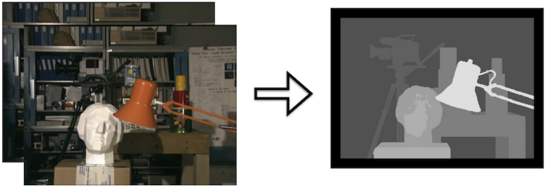
- We could model this as a Markov random field
- Joint distribution $P(\boldsymbol{Y}, \boldsymbol{X})$ over disparity $\boldsymbol{Y}$ and pixel intensities $\boldsymbol{X}$:
  - $P(\boldsymbol{Y} \mid \boldsymbol{X})P(\boldsymbol{X})$ or $P(\boldsymbol{X} \mid \boldsymbol{Y})P(\boldsymbol{Y})$

Example: Stereo Vision



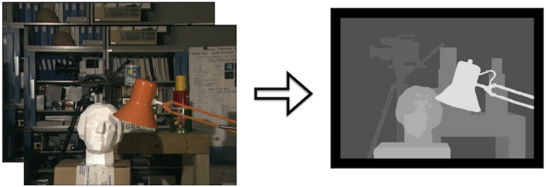
- We could model this as a Markov random field
- Joint distribution $P(\boldsymbol{Y}, \boldsymbol{X})$ over disparity $\boldsymbol{Y}$ and pixel intensities $\boldsymbol{X}$:
  - $P(\boldsymbol{Y} \mid \boldsymbol{X})P(\boldsymbol{X})$ or $P(\boldsymbol{X} \mid \boldsymbol{Y})P(\boldsymbol{Y})$
  - Both involve a (conditional) distribution over natural images!!!

Example: Stereo Vision



- We could model this as a Markov random field
- Joint distribution $P(\boldsymbol{Y}, \boldsymbol{X})$ over disparity $\boldsymbol{Y}$ and pixel intensities $\boldsymbol{X}$:
  - $P(\boldsymbol{Y} \mid \boldsymbol{X})P(\boldsymbol{X})$ or $P(\boldsymbol{X} \mid \boldsymbol{Y})P(\boldsymbol{Y})$
  - Both involve a (conditional) distribution over natural images!!!
- Requires that we choose a parametric form over the pixel values

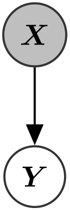
Example: Stereo Vision



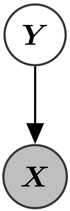
- We could model this as a Markov random field
- Joint distribution $P(\boldsymbol{Y}, \boldsymbol{X})$ over disparity $\boldsymbol{Y}$ and pixel intensities $\boldsymbol{X}$:
  - $P(\boldsymbol{Y} \mid \boldsymbol{X})P(\boldsymbol{X})$ or $P(\boldsymbol{X} \mid \boldsymbol{Y})P(\boldsymbol{Y})$
  - Both involve a (conditional) distribution over natural images!!!
- Requires that we choose a parametric form over the pixel values
- Non-local features (e.g., image gradients) are useful, but difficult to capture with a joint distribution

Generative vs. Discriminative Models

- Let $\boldsymbol{X}$ denote the input/observation (e.g., an image) and $\boldsymbol{Y}$ be the output (e.g., disparity, label, etc.)
- Flexibility in the structure and parametrization of the model



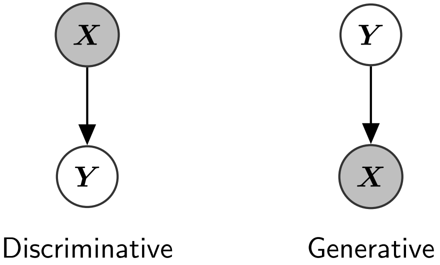
Discriminative



Generative

Generative vs. Discriminative Models

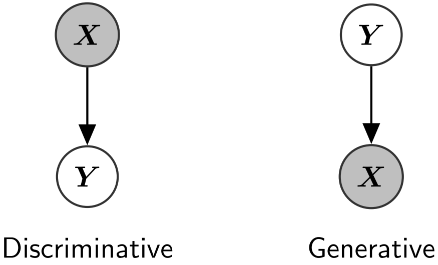
- Let $\boldsymbol{X}$ denote the input/observation (e.g., an image) and $\boldsymbol{Y}$ be the output (e.g., disparity, label, etc.)
- Flexibility in the structure and parametrization of the model



- Discriminative models are concerned with modeling $P(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$, where $\boldsymbol{X} = \boldsymbol{x}$ is treated as a parameter

Generative vs. Discriminative Models

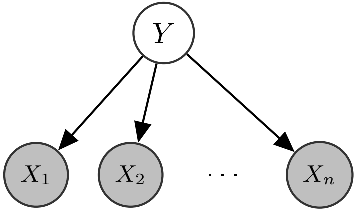
- Let $\boldsymbol{X}$ denote the input/observation (e.g., an image) and $\boldsymbol{Y}$ be the output (e.g., disparity, label, etc.)
- Flexibility in the structure and parametrization of the model



- Discriminative models are concerned with modeling $P(\boldsymbol{Y} \mid \boldsymbol{X} = \boldsymbol{x})$, where $\boldsymbol{X} = \boldsymbol{x}$ is treated as a parameter
- Generative models are interested in the joint distribution $P(\boldsymbol{X}, \boldsymbol{Y}) = P(\boldsymbol{X} \mid \boldsymbol{Y})P(\boldsymbol{Y})$
  - $P(\boldsymbol{X} \mid \boldsymbol{Y})$: Given the target label, generate the input

Generative vs. Discriminative Classifiers: Naive Bayes

- Classify e-mails as being spam ($Y = 1$) or not spam ($Y = 0$)
  - Let $i \in \{1, \ldots, n\}$ index English words
  - $X_i = 1$ if word i appears in the e-mail
  - E-mails are drawn from the joint distribution $P(Y, X_1, \ldots, X_n)$
- Words are conditionally independent (d-separated) given Y

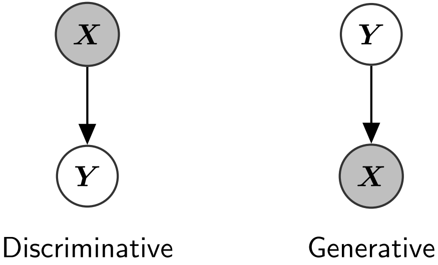


- Prediction follows via Bayes' Rule as

$$P(Y = 1 \mid x_1, \ldots, x_n) = \frac{P(Y = 1) \prod_{i=1}^{n} P(x_i \mid Y = 1)}{\sum_{y=\{0,1\}} P(Y = y) \prod_{i=1}^{n} P(x_i \mid Y = 1)}$$

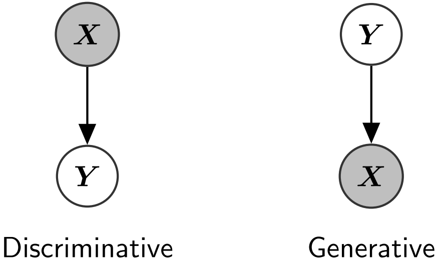
Generative vs. Discriminative Models

- These are **equivalent** models of $P(\boldsymbol{Y}, \boldsymbol{X})$



Generative vs. Discriminative Models

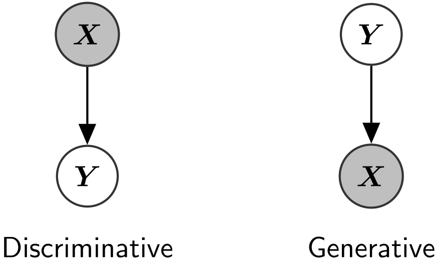
- These are **equivalent** models of $P(\boldsymbol{Y}, \boldsymbol{X})$



- But, suppose that we are only interested in $P(\boldsymbol{Y} \mid \boldsymbol{X})$ for prediction

Generative vs. Discriminative Models

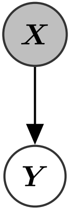
- These are **equivalent** models of $P(\boldsymbol{Y}, \boldsymbol{X})$



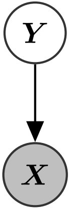
- But, suppose that we are only interested in $P(\boldsymbol{Y} \mid \boldsymbol{X})$ for prediction
- A **generative** model requires representing both $P(\boldsymbol{Y})$ and $P(\boldsymbol{X} \mid \boldsymbol{Y})$
  - Generative, since we can *generate* $\boldsymbol{X}$ given $\boldsymbol{Y}$
  - $P(\boldsymbol{Y} \mid \boldsymbol{X})$ is determined via Bayes' Rule

Generative vs. Discriminative Models

- These are **equivalent** models of $P(\boldsymbol{Y}, \boldsymbol{X})$



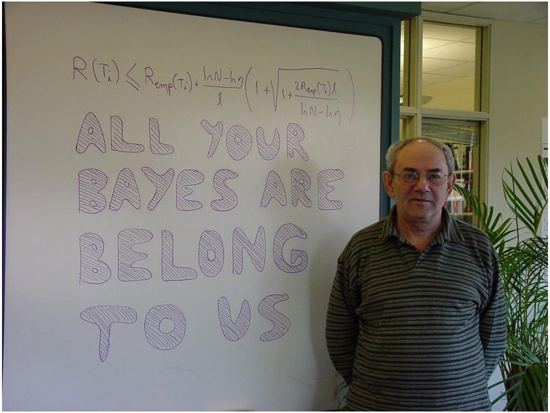
Discriminative



Generative

- But, suppose that we are only interested in $P(\boldsymbol{Y} \mid \boldsymbol{X})$ for prediction
- A **generative** model requires representing both $P(\boldsymbol{Y})$ and $P(\boldsymbol{X} \mid \boldsymbol{Y})$
  - Generative, since we can *generate* $\boldsymbol{X}$ given $\boldsymbol{Y}$
  - $P(\boldsymbol{Y} \mid \boldsymbol{X})$ is determined via Bayes' Rule
- A **discriminative** model only requires a representation of the conditional distribution $P(\boldsymbol{Y} \mid \boldsymbol{X})$
  - Discriminative, since we can *discriminate* between different $\boldsymbol{Y}$
  - We never need to estimate $P(\boldsymbol{X})$ (which can be hard)

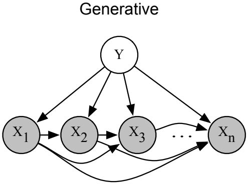
Generative vs. Discriminative Models

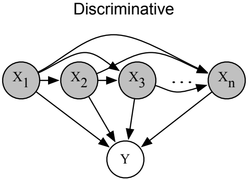


"one should solve the (classification) problem directly and never solve a more general problem as an intermediate step" (Vapnik, 1998)

Generative vs. Discriminative Models

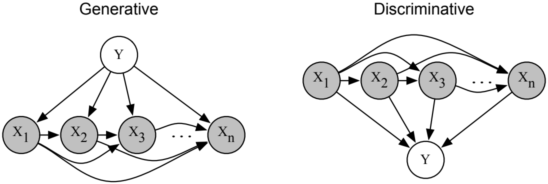
- For the two models to be equivalent, we need:





Generative vs. Discriminative Models

- For the two models to be equivalent, we need:

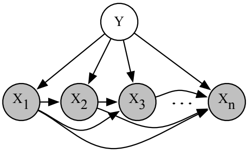


- Modeling requires the following decisions:
  - Generative model: How do we parametrize $P(X_i \mid \mathrm{Pa}^G_{X_i}, \boldsymbol{Y})$?

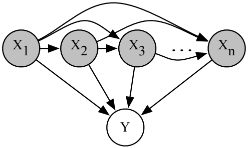
Generative vs. Discriminative Models

- For the two models to be equivalent, we need:

Generative



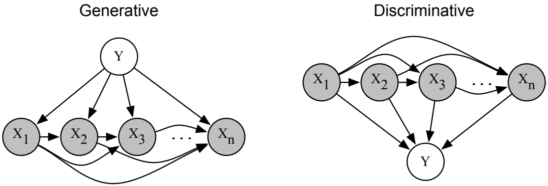
Discriminative



- Modeling requires the following decisions:
  - Generative model: How do we parametrize $P(X_i \mid \mathrm{Pa}_{X_i}^G, \boldsymbol{Y})$?
    Can be hard (e.g., distribution over $(3 \times 255)^{640 \times 480}$ pixel intensities)

Generative vs. Discriminative Models

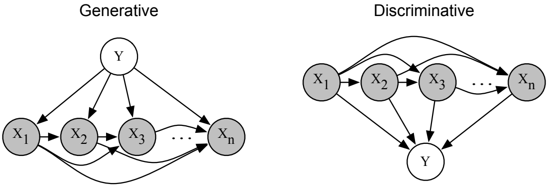
- For the two models to be equivalent, we need:



- Modeling requires the following decisions:
  - Generative model: How do we parametrize $P(X_i \mid \text{Pa}^G_{X_i}, \boldsymbol{Y})$?
    Can be hard (e.g., distribution over $(3 \times 255)^{640 \times 480}$ pixel intensities)
  - Discriminative model: How do we parametrize $P(\boldsymbol{Y} \mid \boldsymbol{X})$?

Generative vs. Discriminative Models

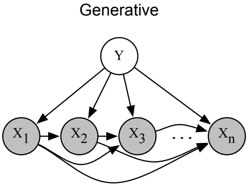
- For the two models to be equivalent, we need:

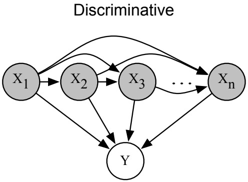


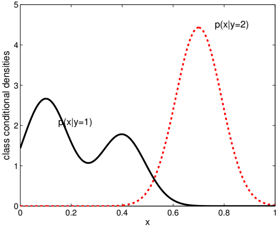
- Modeling requires the following decisions:
  - Generative model: How do we parametrize $P(X_i \mid \mathrm{Pa}^G_{X_i}, \boldsymbol{Y})$?
    Can be hard (e.g., distribution over $(3 \times 255)^{640 \times 480}$ pixel intensities)
  - Discriminative model: How do we parametrize $P(\boldsymbol{Y} \mid \boldsymbol{X})$?
    Allows us to ignore encoding distribution over $\boldsymbol{X}$

Generative vs. Discriminative Models

- Modeling requires the following decisions:
  - Generative model: How do we parametrize $P(X_i \mid \mathsf{Pa}_{X_i}, Y)$?
  - Discriminative model: How do we parametrize $P(\boldsymbol{Y} \mid \boldsymbol{X})$?

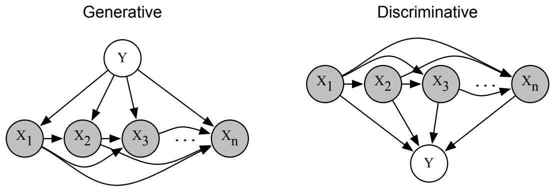


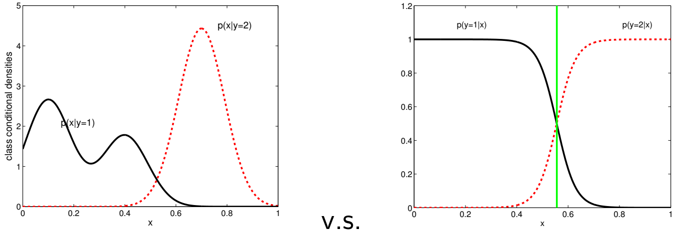




Generative vs. Discriminative Models

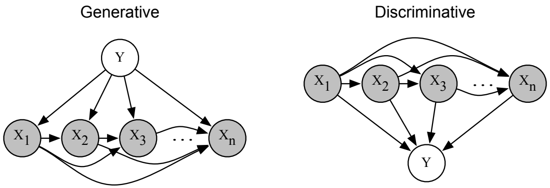
- Modeling requires the following decisions:
  - Generative model: How do we parametrize $P(X_i \mid \mathrm{Pa}_{X_i}, Y)$?
  - Discriminative model: How do we parametrize $P(\boldsymbol{Y} \mid \boldsymbol{X})$?





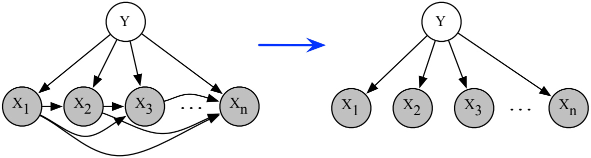
Generative vs. Discriminative Models

- Modeling requires the following decisions:
  - Generative model: How do we parametrize $P(X_i \mid \mathrm{Pa}_{X_i}, \boldsymbol{Y})$?
  - Discriminative model: How do we parametrize $P(\boldsymbol{Y} \mid \boldsymbol{X})$?



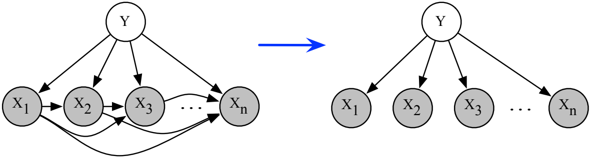
Generative vs. Discriminative Models

- For the generative model, ignore the dependencies: assume that $X_i \perp \boldsymbol{X}_{-i} \mid \boldsymbol{Y}$ (naive Bayes)



Generative vs. Discriminative Models

- For the generative model, ignore the dependencies: assume that $X_i \perp \boldsymbol{X}_{-i} \,|\, \boldsymbol{Y}$ (naive Bayes)



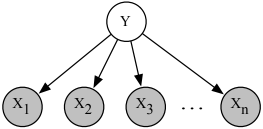
- For the discriminative model, assume that

$$P(Y = 1 \,|\, \boldsymbol{x}; w) = \frac{e^{w_0 + \sum_{i=1}^{n} w_i x_i}}{1 + e^{w_0 + \sum_{i=1}^{n} w_i x_i}} = \frac{1}{1 + e^{-w_0 - \sum_{i=1}^{n} w_i x_i}}$$

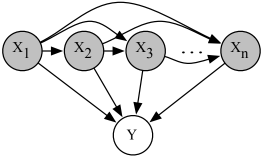
$$= \frac{1}{1 + e^{-z}}$$ **logistic function**

Generative vs. Discriminative Models

Generative (naive Bayes)



Discriminative (logistic regression)

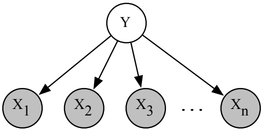


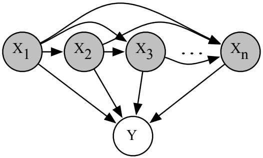
1. For the generative model, ignore the dependencies: assume that $X_i \perp \boldsymbol{X}_{-i} \,|\, \boldsymbol{Y}$ (naive Bayes)
2. For the discriminative model, assume that

$$P(Y = 1 \,|\, \boldsymbol{x}; w) = \frac{e^{w_0 + \sum_{i=1}^{n} w_i x_i}}{1 + e^{w_0 + \sum_{i=1}^{n} w_i x_i}} = \frac{1}{1 + e^{-w_0 - \sum_{i=1}^{n} w_i x_i}}$$

- We can show (problem set) that the first assumption implies the latter
- Every conditional distribution that can be represented via naive Bayes can also be represented using the logistic model

Generative vs. Discriminative Models





- Unlike naive Bayes, logistic models don't assume $X_i \perp \boldsymbol{X}_{-i} \,|\, \boldsymbol{Y}$
- Ignoring dependencies results in double-counting evidence, e.g.,
  - Suppose that $X_i = 1$ ("transaction" in e-mail) and $X_j = 1$ ("account" in e-mail)
  - Irrespective of being spam, these always occur together, i.e., $X_i = X_j$
  - Learning with naive Bayes ($P(X_i \,|\, \boldsymbol{Y}) = P(X_j \,|\, \boldsymbol{Y})$) double-counts evidence
  - Learning with logistic regression sets $w_i = 0$ for one of the words, thereby ignoring it

Generative vs. Discriminative Models

- Discriminative model requires that $\boldsymbol{X}$ be fully observed
  - Generative models allow you to marginalize over unseen variables to compute $P(\boldsymbol{Y} \mid \boldsymbol{X}_o)$
- Maximum likelihood estimation of generative models is more efficient than training discriminative models [Ng and Jordan, 2002][1]
  - Consider number of samples necessary to get close to infinite data case
  - Logistic regression requires $\mathcal{O}(n)$ samples
  - Naive Bayes requires $\mathcal{O}(\log n)$ samples
  - Naive Bayes converges with fewer samples, but not necessarily to better estimates

[1] Ng and Jordan, "On Discriminative vs. Generative Classifiers: A Comparison of Logistic Regression and Naive Bayes," NeurIPS 2002

Generative vs. Discriminative Models

- Ng and Jordan (2002) show that discriminative logistic regression has lower asymptotic error than a generative naive Bayes classifier, but that naive Bayes converges faster

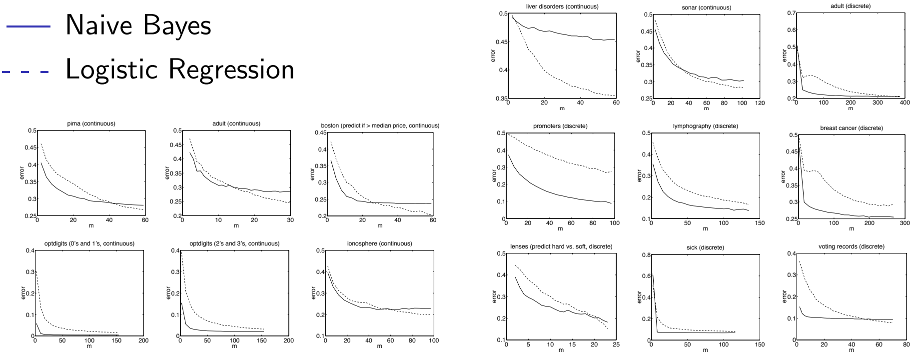
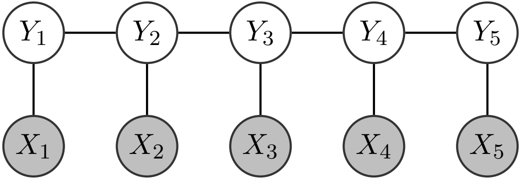


Figure: Error rate vs. number of samples for 15 datasets from the UCI Machine Learning repository. Courtesy: Ng and Jordan, 2002.

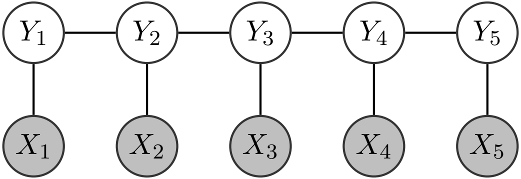
Conditional Random Fields



(linear chain CRF)

- Undirected graph with nodes for $\boldsymbol{Y}$ (target variables) and $\boldsymbol{X}$ (observed variables) (alt., partially directed, with $\boldsymbol{X}$ the parent of $\boldsymbol{Y}$)

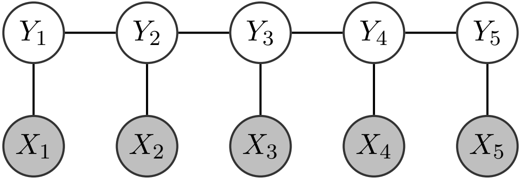
Conditional Random Fields



(linear chain CRF)

- Undirected graph with nodes for $\boldsymbol{Y}$ (target variables) and $\boldsymbol{X}$ (observed variables) (alt., partially directed, with $\boldsymbol{X}$ the parent of $\boldsymbol{Y}$)
- Parametrized by a set of factors $\phi_1(\boldsymbol{D}_1), \phi_2(\boldsymbol{D}_2), \ldots, \phi_m(\boldsymbol{D}_m)$

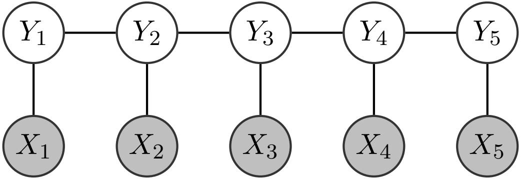
Conditional Random Fields



(linear chain CRF)

- Undirected graph with nodes for $\boldsymbol{Y}$ (target variables) and $\boldsymbol{X}$ (observed variables) (alt., partially directed, with $\boldsymbol{X}$ the parent of $\boldsymbol{Y}$)
- Parametrized by a set of factors $\phi_1(\boldsymbol{D}_1), \phi_2(\boldsymbol{D}_2), \ldots, \phi_m(\boldsymbol{D}_m)$
- Represent conditional distribution $P(\boldsymbol{Y} \mid \boldsymbol{X})$ rather than joint distribution $P(\boldsymbol{Y}, \boldsymbol{X})$

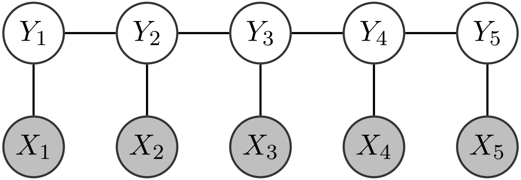
Conditional Random Fields



(linear chain CRF)

- Undirected graph with nodes for $\boldsymbol{Y}$ (target variables) and $\boldsymbol{X}$ (observed variables) (alt., partially directed, with $\boldsymbol{X}$ the parent of $\boldsymbol{Y}$)
- Parametrized by a set of factors $\phi_1(\boldsymbol{D}_1), \phi_2(\boldsymbol{D}_2), \ldots, \phi_m(\boldsymbol{D}_m)$
- Represent conditional distribution $P(\boldsymbol{Y} \mid \boldsymbol{X})$ rather than joint distribution $P(\boldsymbol{Y}, \boldsymbol{X})$
- Avoid representing dist. over $\boldsymbol{X} \Rightarrow$ no potentials involving only $\boldsymbol{X}$

Conditional Random Fields



(linear chain CRF)

- A **conditional random field** (CRF) is an undirected graph H over $\boldsymbol{X}$ and $\boldsymbol{Y}$ defined in terms of a set of factors $\phi_1(\boldsymbol{D}_1), \ldots, \phi_m(\boldsymbol{D}_m)$, where $\boldsymbol{D}_i \not\subseteq \boldsymbol{X}$, that encodes the conditional distribution

$$P(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

$$Z(\boldsymbol{X}) = \sum_{\boldsymbol{Y}} \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

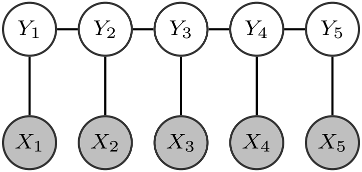
Conditional Random Fields

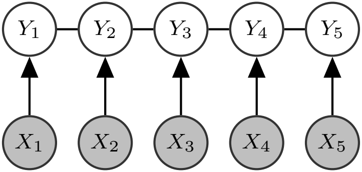
$$P(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

$$Z(\boldsymbol{X}) = \sum_{\boldsymbol{y}} \prod_{i=1}^{m} \phi_i(\boldsymbol{D}_i)$$

- Two variables are connected by an undirected edge if they appear in the scope of the same factor
- Just like a Markov network, except the partition function depends on (i.e., changes with) the observed variables (input) $\boldsymbol{X}$
- Trained to maximize *conditional* (not joint) probability of the output given the input

Linear-Chain CRFs



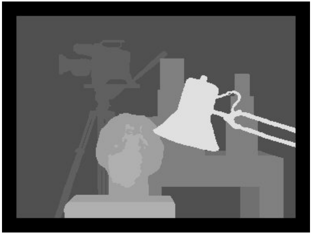


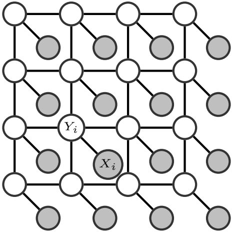
- The linear-chain CRF can be represented as an undirected (left) or partially directed graph (right)
- The conditional distribution factorizes as

$$P(\boldsymbol{Y} \mid \boldsymbol{X}) = \frac{1}{Z(\boldsymbol{X})} \prod_{i=1}^{k-1} \phi(Y_i, Y_{i+1}) \prod_{i=1}^{k} \phi(Y_i, X_i)$$

$$Z(\boldsymbol{X}) = \sum_{\boldsymbol{Y}} \prod_{i=1}^{k-1} \phi(Y_i, Y_{i+1}) \prod_{i=1}^{k} \phi(Y_i, X_i)$$

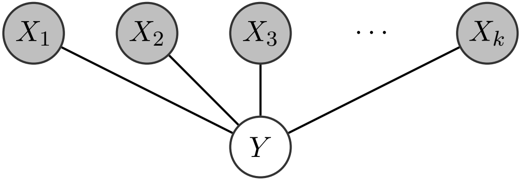
Conditional Random Fields





- By not modeling the distribution over $\boldsymbol{X}$, we can consider representations of the data with complex, non-parametric interactions
- We can employ a rich set of features without concern over their joint distribution (e.g., image gradients)

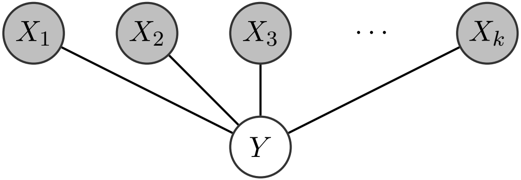
Naive Markov Model



- Let $\boldsymbol{X} = \{X_1.X_2, \ldots, X_k\}$ and Y be binary random variables
- Assume $\boldsymbol{X}$ (observed) and Y are related by the following factors

$$\phi_0(Y) = \exp\{w_0 \mathbb{1}[Y = 1]\}$$
$$\phi_i(X_i, Y) = \exp\{w_i \mathbb{1}[X_i = 1, Y = 1]\}$$

Naive Markov Model



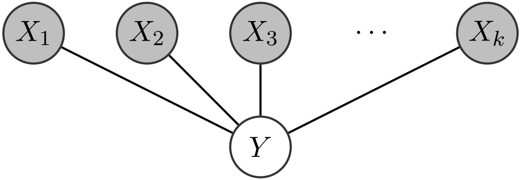
- Let $\boldsymbol{X} = \{X_1.X_2, \ldots, X_k\}$ and Y be binary random variables
- Assume $\boldsymbol{X}$ (observed) and Y are related by the following factors

$$\phi_0(Y) = \exp\{w_0 \mathbb{1}[Y = 1]\}$$
$$\phi_i(X_i, Y) = \exp\{w_i \mathbb{1}[X_i = 1, Y = 1]\}$$

- The conditional distribution becomes

$$\tilde{P}(Y = 1 \,|\, x_1, \ldots, x_k) = \exp\left\{w_0 + \sum_{i=1}^{k} w_i x_i\right\}$$
$$\tilde{P}(Y = 0 \,|\, x_1, \ldots, x_k) = 1$$

Naive Markov Model



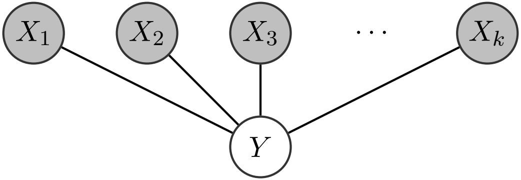
- Let $\boldsymbol{X} = \{X_1.X_2, \ldots, X_k\}$ and Y be binary random variables
- Assume $\boldsymbol{X}$ (observed) and Y are related by the following factors

$$\phi_0(Y) = \exp\{w_0 \mathbb{1}[Y = 1]\}$$
$$\phi_i(X_i, Y) = \exp\{w_i \mathbb{1}[X_i = 1, Y = 1]\}$$

- The conditional distribution becomes

$$P(Y = 1 \mid x_1, \ldots, x_k) = \frac{\exp\left\{w_0 + \sum_{i=1}^k w_i x_i\right\}}{1 + \exp\left\{w_0 + \sum_{i=1}^k w_i x_i\right\}} = \text{sigmoid}\left(w_0 + \sum_{i=1}^k w_i x_i\right)$$

Naive Markov Model



- Let $\boldsymbol{X} = \{X_1.X_2, \ldots, X_k\}$ and Y be binary random variables
- Assume $\boldsymbol{X}$ (observed) and Y are related by the following factors

$$\phi_0(Y) = \exp\{w_0 \mathbb{1}[Y = 1]\}$$
$$\phi_i(X_i, Y) = \exp\{w_i \mathbb{1}[X_i = 1, Y = 1]\}$$

- The conditional distribution becomes

$$P(Y = 1 \mid x_1, \ldots, x_k) = \frac{\exp\left\{w_0 + \sum_{i=1}^k w_i x_i\right\}}{1 + \exp\left\{w_0 + \sum_{i=1}^k w_i x_i\right\}} = \text{sigmoid}\left(w_0 + \sum_{i=1}^k w_i x_i\right)$$

- Recall earlier discriminative (logistic regression) discussion

CRF Parametrization

- Factors may depend on a large number of variables
- Typically, parametrize factors using log-linear representation

$$\phi_c(\boldsymbol{X}_c, \boldsymbol{Y}_c) = \exp\left(\boldsymbol{w}_c^\top \boldsymbol{f}_c(\boldsymbol{X}_c, \boldsymbol{Y}_c)\right)$$

where

  - $\boldsymbol{f}_c(\boldsymbol{X}_c, \boldsymbol{Y}_c)$ is a feature vector (e.g., local image gradients)
  - $\boldsymbol{w}_c$ is a weight vector that is learned from data

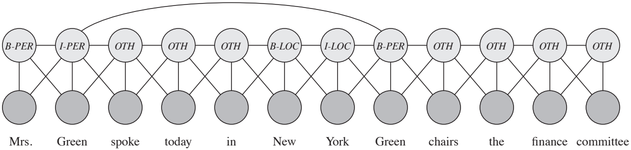
Example: Named Entity Recognition

- Objective: Given a sentence, segment phrases into different locations, people, organizations, etc.

  "Mrs. Green spoke today in New York Green chairs the finance committee"

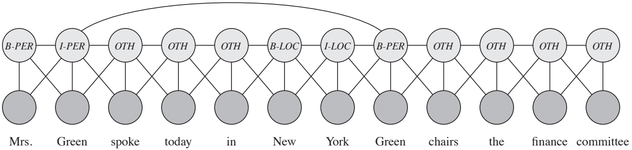
- Entries often span multiple words and label isn't obvious without considering context
- Define a random variable Y_i for each word X_i that expresses its entity type ("BIO notation")
  - B-PER/B-LOC: Beginning of a person/location
  - I-PER/I-LOC: Inside or end of named entity phrase for person/location
  - OTH: Word is not part of an entity

Example: Named Entity Recognition



Interested in $P(\boldsymbol{Y} \mid \boldsymbol{X}) \Rightarrow$ Model as a CRF w/ three types of factors

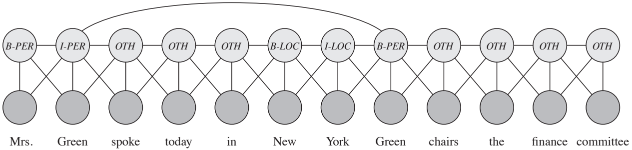
Example: Named Entity Recognition



Interested in $P(\boldsymbol{Y} \mid \boldsymbol{X}) \Rightarrow$ Model as a CRF w/ three types of factors

- $\phi^1(Y_t, Y_{t+1})$: Expresses dependency between neighboring entities (similar to transition distribution in HMMs)

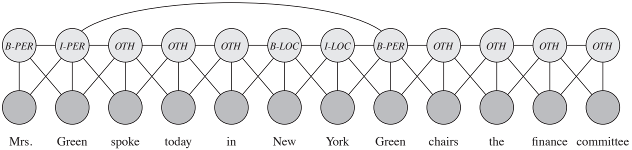
Example: Named Entity Recognition



Interested in $P(\boldsymbol{Y} \mid \boldsymbol{X}) \Rightarrow$ Model as a CRF w/ three types of factors

- $\phi^1(Y_t, Y_{t+1})$: Expresses dependency between neighboring entities (similar to transition distribution in HMMs)
- $\phi^2(Y_t, X_1, \ldots, X_T)$: Expresses dependency between an entity and the entire word sequence (context), encoded via log-linear model
    - O(1000) features over current word X_t (e.g., capitalized), neighboring words, and entire sequence (e.g., number of sports-related words)

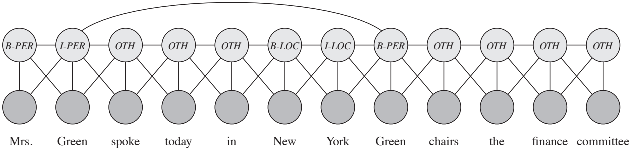
Example: Named Entity Recognition



Interested in $P(\boldsymbol{Y} \mid \boldsymbol{X}) \Rightarrow$ Model as a CRF w/ three types of factors

- $\phi^1(Y_t, Y_{t+1})$: Expresses dependency between neighboring entities (similar to transition distribution in HMMs)
- $\phi^2(Y_t, X_1, \ldots, X_T)$: Expresses dependency between an entity and the entire word sequence (context), encoded via log-linear model
  - O(1000) features over current word X_t (e.g., capitalized), neighboring words, and entire sequence (e.g., number of sports-related words)
- $\phi^3(Y_t, Y_{t'})$: For all pairs t, t' s.t. $X_t = X_{t'}$, since words that appear twice should be the same entity (skip-chain CRF)

Example: Named Entity Recognition

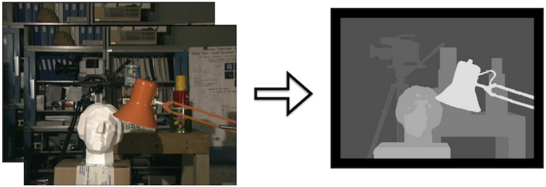


Interested in $P(\boldsymbol{Y} \mid \boldsymbol{X}) \Rightarrow$ Model as a CRF w/ three types of factors

- $\phi^1(Y_t, Y_{t+1})$: Expresses dependency between neighboring entities (similar to transition distribution in HMMs)
- $\phi^2(Y_t, X_1, \ldots, X_T)$: Expresses dependency between an entity and the entire word sequence (context), encoded via log-linear model
  - O(1000) features over current word X_t (e.g., capitalized), neighboring words, and entire sequence (e.g., number of sports-related words)
- $\phi^3(Y_t, Y_{t'})$: For all pairs t, t' s.t. $X_t = X_{t'}$, since words that appear twice should be the same entity (skip-chain CRF)

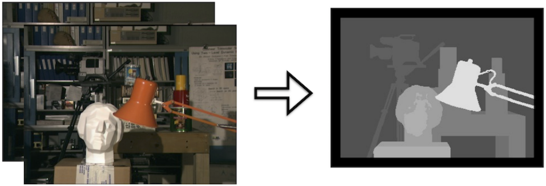
Graph structure changes depending on sentence

Example: Stereo Vision



- Given two images $\boldsymbol{X}_l$ and $\boldsymbol{X}_r$, estimate the disparity $y_{i_l,j_l} \in \boldsymbol{Y}$ between a pixel at (i_l, j_l) in the left image and the corresponding pixel $(i_r, j_r) = (i_l + y_{i_l,j_l}, j_l)$ in the right image (assume 1D)

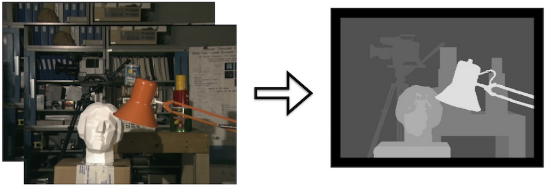
Example: Stereo Vision



- Define local node potential as

$$\phi_{i_l,j_l}(y_{i_l,j_l}, \boldsymbol{X}) \propto \exp\left(-\frac{1}{2\sigma^2}\left(x_l(i_l, j_l) - x_r(i_l + y_{i_l,j_l}, j_l)\right)^2\right)$$

Example: Stereo Vision



- Define local node potential as

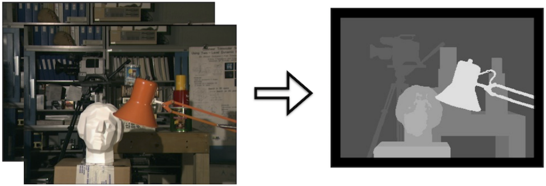
$$\phi_{i_l,j_l}(y_{i_l,j_l}, \boldsymbol{X}) \propto \exp\left(-\frac{1}{2\sigma^2}\left(x_l(i_l, j_l) - x_r(i_l + y_{i_l,j_l}, j_l)\right)^2\right)$$

- Define smoothness potential over neighboring pixels s and t

$$\phi_{s,t}(y_s, y_t) \propto \left(-\frac{1}{2\gamma^2}(y_s - y_t)^2\right)$$

Can be defined only for pairs s and t that don't cross boundaries

Example: Stereo Vision



- Define local node potential as

$$\phi_{i_l,j_l}(y_{i_l,j_l}, \boldsymbol{X}) \propto \exp\left(-\frac{1}{2\sigma^2}\left(x_l(i_l,j_l) - x_r(i_l + y_{i_l,j_l}, j_l)\right)^2\right)$$

- Define smoothness potential over neighboring pixels s and t

$$\phi_{s,t}(y_s, y_t) \propto \left(-\frac{1}{2\gamma^2}(y_s - y_t)^2\right)$$

Can be defined only for pairs s and t that don't cross boundaries

Graph structure changes depending on the image

Conditional Random Fields

There are advantages and disadvantages to MRF and CRF formulations (just as with discriminative and generative classifiers):

Conditional Random Fields

There are advantages and disadvantages to MRF and CRF formulations (just as with discriminative and generative classifiers):

+ CRFs don't "waste resources" modeling things that are observed (data) when we only care about distribution over labels given data

Conditional Random Fields

There are advantages and disadvantages to MRF and CRF formulations (just as with discriminative and generative classifiers):

+ CRFs don't "waste resources" modeling things that are observed (data) when we only care about distribution over labels given data
+ CRFs easily accommodate data-dependent potentials (factors)
    - In image segmentation, we can disable smoothing between labels separated by an edge
    - In named entity recognition, we can tie repeated words together

Conditional Random Fields

There are advantages and disadvantages to MRF and CRF formulations (just as with discriminative and generative classifiers):

- \+ CRFs don't "waste resources" modeling things that are observed (data) when we only care about distribution over labels given data
- \+ CRFs easily accommodate data-dependent potentials (factors)
  - In image segmentation, we can disable smoothing between labels separated by an edge
  - In named entity recognition, we can tie repeated words together
- \+ CRFs easily accommodate labels that depend on global properties of data (difficult to do with MRFs)

Conditional Random Fields

There are advantages and disadvantages to MRF and CRF formulations (just as with discriminative and generative classifiers):

+ CRFs don't "waste resources" modeling things that are observed (data) when we only care about distribution over labels given data
+ CRFs easily accommodate data-dependent potentials (factors)
    - In image segmentation, we can disable smoothing between labels separated by an edge
    - In named entity recognition, we can tie repeated words together
+ CRFs easily accommodate labels that depend on global properties of data (difficult to do with MRFs)

– CRFs require labeled training data and are slower to train