

Conditional Random Fields

Sargur N. Srihari

srihari@cedar.buffalo.edu

Machine Learning Course:

<http://www.cedar.buffalo.edu/~srihari/CSE574/index.html>

Outline

1. Generative and Discriminative Models
2. Classifiers
 - Naïve Bayes and Logistic Regression
3. Sequential Models
 - HMM and CRF
 - Markov Random Field
4. CRF vs HMM performance comparison
 - NLP: Table extraction, POS tagging, Shallow parsing, Document analysis
5. CRFs in Computer Vision
6. Summary
7. References

Methods for Classification

- **Generative Models (Two-step)**
 - Infer class-conditional densities $p(x|C_k)$ and priors $p(C_k)$
 - then use Bayes theorem to determine posterior probabilities.
$$p(C_k | x) = \frac{p(x | C_k)p(C_k)}{p(x)}$$
- **Discriminative Models (One-step)**
 - Directly infer posterior probabilities $p(C_k/x)$
- **Decision Theory**
 - In both cases use decision theory to assign each new x_3 to a class

Generative Classifier

- Given M variables, $\mathbf{x} = (x_1, \dots, x_M)$, class variable y and joint distribution $p(\mathbf{x}, y)$ we can
 - Marginalize
$$p(y) = \sum_{\mathbf{x}} p(\mathbf{x}, y)$$
 - Condition
$$p(y | \mathbf{x}) = \frac{p(\mathbf{x}, y)}{p(\mathbf{x})}$$
 - By conditioning the joint pdf we form a classifier
- Huge need for samples
 - If x_i are binary, need 2^M values to specify $p(\mathbf{x}, y)$
 - If $M=10$, there are two classes and 100 samples are needed to estimate a given probability, then we need 2048 samples

Classification of ML Methods

- **Generative Methods**

- “Generative” since sampling can generate synthetic data points
- Popular models
 - Naïve Bayes, Mixtures of multinomials
 - Mixtures of Gaussians, Hidden Markov Models
 - Bayesian networks, Markov random fields

- **Discriminative Methods**

- Focus on given task— better performance
- Popular models
 - Logistic regression, SVMs
 - Traditional neural networks, Nearest neighbor
 - Conditional Random Fields (CRF)

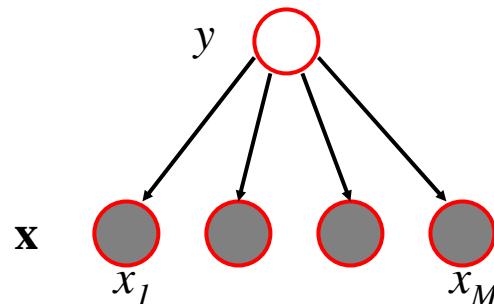
Generative-Discriminative Pairs

- Naïve Bayes and Logistic Regression form a *generative-discriminative* pair
- Their relationship mirrors that between HMMs and linear-chain CRFs

Graphical Model Relationship

GENERATIVE

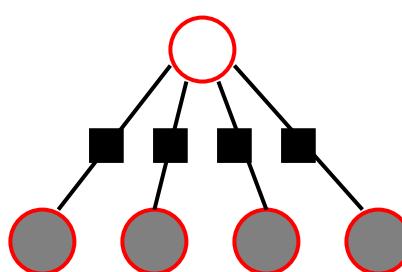
Naïve Bayes Classifier



SEQUENCE

$p(y, \mathbf{x})$

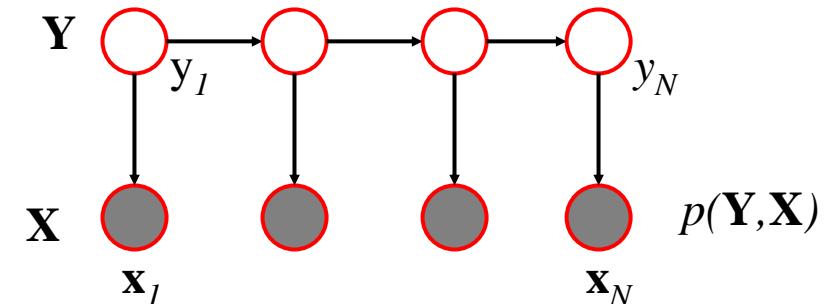
CONDITION



DISCRIMINATIVE

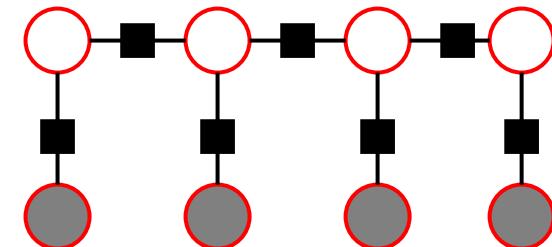
Logistic Regression

Hidden Markov Model



CONDITION

$p(\mathbf{Y}/\mathbf{X})$



Conditional Random Field

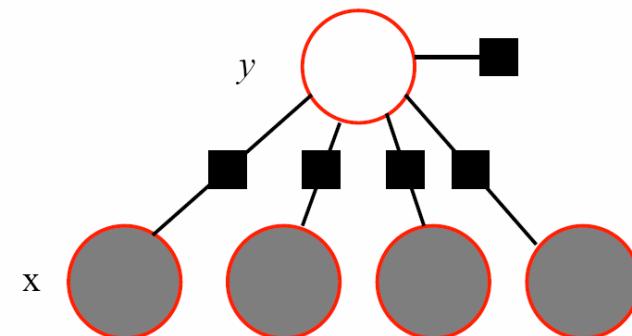
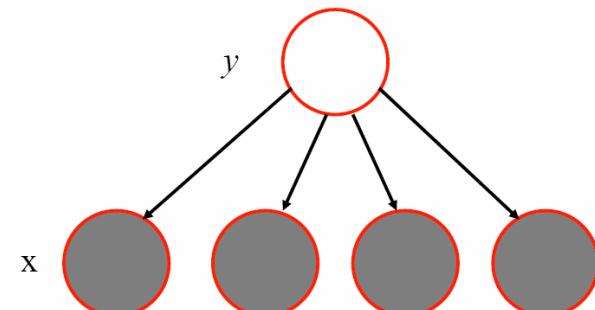
Naïve Bayes Classifier

- Goal is to predict single class variable y given a vector of features $\mathbf{x}=(x_1, \dots, x_M)$
- Assume that once class labels are known the features are independent
- Joint probability model has the form

$$p(y, \mathbf{x}) = p(y) \prod_{m=1}^M p(x_m | y)$$

- Need to estimate only M probabilities
- Factor graph obtained by defining factors

$$\psi(y) = p(y), \quad \psi_m(y, x_m) = p(x_m | y)$$



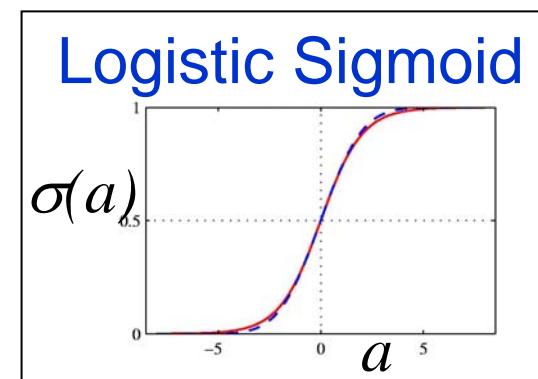
Logistic Regression Classifier

- Feature vector \mathbf{x}
- Two-class classification: class variable y has values C_1 and C_2
- *A posteriori* probability $p(C_1/x)$ written as

$$p(C_1/x) = f(\mathbf{x}) = \sigma(\mathbf{w}^T \mathbf{x}) \text{ where}$$

$$\sigma(a) = \frac{1}{1 + \exp(-a)}$$

- Known as logistic regression in statistics
 - Although a model for classification rather than for regression



Properties:

A. Symmetry

$$\sigma(-a) = 1 - \sigma(a)$$

B. Inverse

$$a = \ln(\sigma / 1 - \sigma)$$

known as *logit*.

Also known as
log odds since
it is the ratio

$$\ln[p(C_1/x)/p(C_2/x)]$$

C. Derivative 9

$$d\sigma/d\alpha = \sigma(1 - \sigma)$$

Relationship between Logistic Regression and Generative Classifier

- Posterior probability of class variable y is

$$\begin{aligned} p(C_1 | x) &= \frac{p(x | C_1)p(C_1)}{p(x | C_1)p(C_1) + p(x | C_2)p(C_2)} \\ &= \frac{1}{1 + \exp(-a)} = \sigma(a) \quad \text{where } a = \ln \frac{p(x | C_1)p(C_1)}{p(x | C_2)p(C_2)} \end{aligned}$$

- In a generative model we estimate the class-conditionals (which are used to determine a)
- In the discriminative approach we directly estimate a as a linear function of x i.e., $a = w^T x$

Logistic Regression Parameters

- With M variables logistic regression has M parameters $\mathbf{w} = (w_1, \dots, w_M)$
- By contrast, *generative approach*
 - by fitting Gaussian class-conditional densities will result in $2M$ parameters for means, $M(M+1)/2$ parameters for shared covariance matrix, and one for class prior $p(C_1)$
 - Which can be reduced to $O(M)$ parameters by assuming independence via Naïve Bayes

Learning Logistic Parameters

- For data set $\{x_n, t_n\}$, $t_n \in \{0, 1\}$ likelihood is where $t = (t_1, \dots, t_N)^T$ and $y_n = p(C_1 | x_n)$

$$p(t | w) = \prod_{n=1}^N y_n^{t_n} (1 - y_n)^{1-t_n}$$

- Parameter w specifies the y_n as follows

$$y_n = \sigma(a_n) \text{ and } a_n = w^T x_n$$

- Defining cross-entropy error function

$$E(w) = -\ln p(t | w) = -\sum_{n=1}^N \{t_n \ln y_n + (1 - t_n) \ln(1 - y_n)\}$$

- Taking gradient wrt w

$$\nabla E(w) = \sum_{n=1}^N (y_n - t_n) x_n$$

- Same form as sum-of-squares error for linear regression
 - Can use sequential algorithm where samples are presented one at a time using

$$w^{(t+1)} = w^{(t)} - \eta \nabla E_n$$

Multi-class Logistic Regression

- Case of $K > 2$ classes

$$\begin{aligned} p(C_k | \mathbf{x}) &= \frac{p(\mathbf{x} | C_k) p(C_k)}{\sum_j p(\mathbf{x} | C_j) p(C_j)} \\ &= \frac{\exp(a_k)}{\sum_j \exp(a_j)} \end{aligned}$$

- Known as **normalized exponential** where $a_k = \ln p(\mathbf{x} | C_k) p(C_k)$
- Normalized exponential also known as **softmax** since if $a_k \gg a_j$ then $p(C_k | \mathbf{x}) = 1$ and $p(C_j | \mathbf{x}) = 0$
- In logistic regression we assume **activations** given by $a_k = \mathbf{w}_k^T \mathbf{x}$

Learning Parameters for Multiclass

- Determining parameters $\{w_k\}$ by m.l.e. requires derivatives of y_k wrt all activations a_j

$$\frac{\partial y_k}{\partial a_j} = y_k (I_{kj} - y_j)$$

where I_{kj} are elements of the identity matrix

- Likelihood function written using 1-of- K coding
 - Target vector t_n for x_n belonging to C_k is a binary vector with all elements zero except for element k which is one

$$p(T | w_1, \dots, w_K) = \prod_{n=1}^N \prod_{k=1}^K p(C_k | x_n)^{t_{nk}} = \prod_{n=1}^N \prod_{k=1}^K y_{nk}^{t_{nk}}$$

where T is an $N \times K$ matrix of target variables with elements t_{nk}

- Cross entropy error function

$$E(w_1, \dots, w_K) = -\ln p(T | w_1, \dots, w_K) = -\sum_{n=1}^N \sum_{k=1}^K t_{nk} \ln y_{nk}$$

- Gradient of error function is

$$\sum_{n=1}^N (y_{nj} - t_{nj}) x_n$$

- which allows a sequential weight vector update algorithm

Graphical Model for Logistic Regression

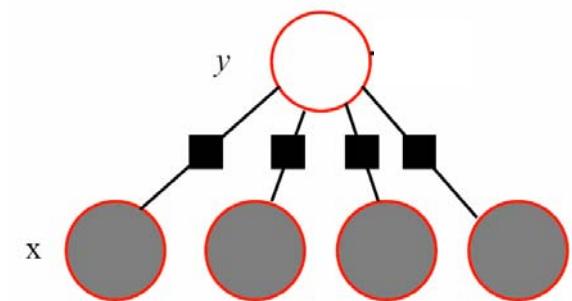
- Multiclass logistic regression can be written as

$$p(y | x) = \frac{1}{Z(x)} \exp \left\{ \lambda_y + \sum_{j=1}^K \lambda_{yj} x_j \right\} \text{ where}$$

$$Z(x) = \sum_y \exp \left\{ \lambda_y + \sum_{j=1}^K \lambda_{yj} x_j \right\}$$

- Rather than using one weight per class we can define feature functions that are nonzero only for a single class

$$p(y | x) = \frac{1}{Z(x)} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(y, x) \right\}$$



- This notation mirrors the usual notation for CRFs

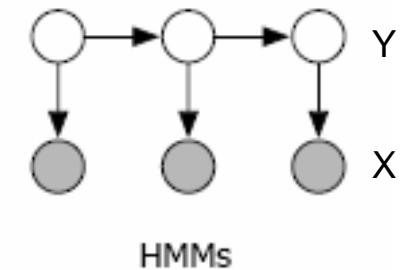
3. Sequence Models

- Classifiers predict only a single class variable
- Graphical Models are best to model many variables that are interdependent
- Given sequence of observations $X = \{x_n\}_{n=1}^N$
- Underlying sequence of states $Y = \{y_n\}_{n=1}^N$

Sequence Models

- Hidden Markov Model (HMM)

$$p(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^T p(y_t | y_{t-1}) p(x_t | y_t)$$

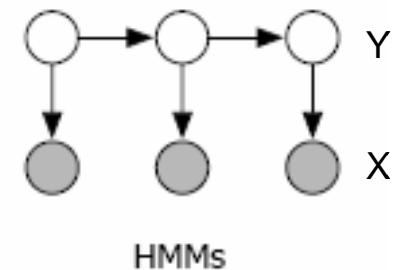


- Independence assumptions:
 - each state depends only on its immediate predecessor
 - each observation variable depends only on the current state
- Limitations:
 - Strong independence assumptions among the observations X.
 - Introduction of a large number of parameters by modeling the joint probability $p(\mathbf{y}, \mathbf{x})$, which requires modeling the distribution $p(\mathbf{x})$

Sequence Models

Hidden Markov Model (HMM)

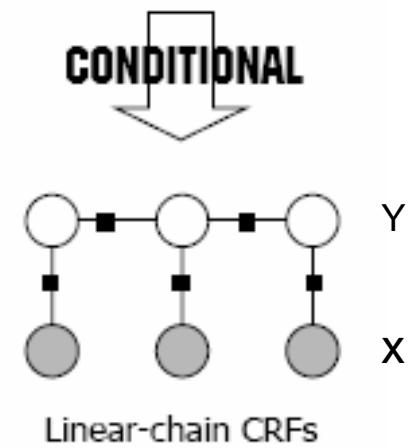
$$p(\mathbf{y}, \mathbf{x}) = \prod_{t=1}^T p(y_t | y_{t-1}) p(x_t | y_t)$$



Conditional Random Fields (CRF)

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left\{ \sum_{k=1}^K \lambda_k f_k(y_t, y_{t-1}, \mathbf{x}_t) \right\}$$

A key advantage of CRF is their great flexibility to include a wide variety of arbitrary, non-independent features of the observations.



Generative Model: HMM

- X is observed data sequence to be labeled,

Y is the random variable over the label sequences

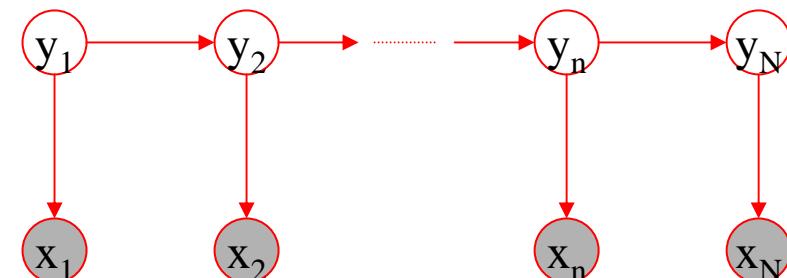
- HMM is a distribution that models $p(Y, X)$

- Joint distribution is

$$p(\mathbf{Y}, \mathbf{X}) = \prod_{n=1}^N p(y_n | y_{n-1}) p(x_n | y_n)$$

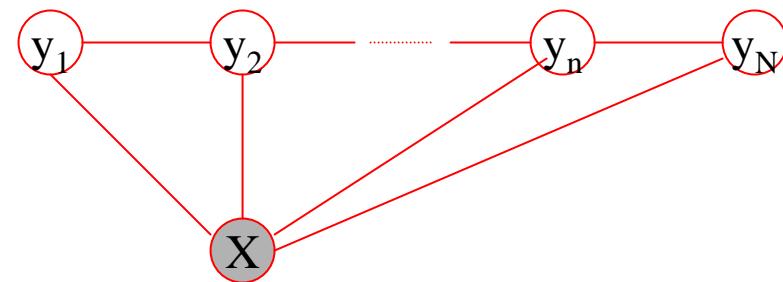
- Highly structured network indicates conditional independences,

- past states independent of future states
 - Conditional independence of observed given its state.



Discriminative Model for Sequential Data

- CRF models the conditional distribution $p(Y/X)$
- CRF is a random field globally conditioned on the observation X
- The conditional distribution $p(Y/X)$ that follows from the joint distribution $p(Y,X)$ can be rewritten as a *Markov Random Field*



Markov Random Field (MRF)

- Also called *undirected graphical model*
- Joint distribution of set of variables x is defined by an undirected graph as

$$p(x) = \frac{1}{Z} \prod_C \psi_C(x_C)$$

where C is a maximal clique

(each node connected to every other node),

x_C is the set of variables in that clique,

ψ_C is a *potential* function (or *local* or *compatibility* function)

such that $\psi_C(x_C) \geq 0$, typically $\psi_C(x_C) = \exp\{-E(x_C)\}$, and

$$Z = \sum_x \prod_C \psi_C(x_C)$$

is the *partition function* for normalization

- *Model* refers to a family of distributions and *Field* refers to a specific one

MRF with Input-Output Variables

- X is a set of input variables that are observed
 - Element of X is denoted x
- Y is a set of output variables that we predict
 - Element of Y is denoted y
- A are subsets of $X \cup Y$
 - Elements of A that are in $A \cap X$ are denoted x_A
 - Element of A that are in $A \cap Y$ are denoted y_A
- Then undirected graphical model has the form

$$p(x,y) = \frac{1}{Z} \prod_A \Psi_A(x_A, y_A)$$

where $Z = \sum_{x,y} \prod_A \Psi_A(x_A, y_A)$

MRF Local Function

- Assume each local function has the form

$$\Psi_A(x_A, y_A) = \exp \left\{ \sum_m \theta_{Am} f_{Am}(x_A, y_A) \right\}$$

where θ_A is a parameter vector, f_A are feature functions and $m=1..M$ are feature subscripts

From HMM to CRF

- In an HMM

$$p(\mathbf{Y}, \mathbf{X}) = \prod_{n=1}^N p(y_n | y_{n-1}) p(\mathbf{x}_n | y_n)$$

Indicator function:

$1_{\{x=x'\}}$ takes value 1 when $x=x'$ and 0 otherwise

- Can be rewritten as

$$p(\mathbf{Y}, \mathbf{X}) = \frac{1}{Z} \exp \left\{ \sum_n \sum_{i,j \in S} \lambda_{ij} 1_{\{y_n=i\}} 1_{\{y_{n-1}=j\}} + \sum_n \sum_{i \in S} \sum_{o \in O} \mu_{oi} 1_{\{y_n=i\}} 1_{\{x_n=o\}} \right\}$$

Parameters of the distribution:
 $\theta = \{\lambda_{ij}, \mu_{oi}\}$

- Further rewritten as

$$p(\mathbf{Y}, \mathbf{X}) = \frac{1}{Z} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, \mathbf{x}_n) \right\}$$

- Which gives us

$$p(\mathbf{Y} | \mathbf{X}) = \frac{p(\mathbf{Y}, \mathbf{X})}{\sum_{y'} p(y', \mathbf{X})} = \frac{\exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, \mathbf{x}_n) \right\}}{\sum_{y'} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, \mathbf{x}_n) \right\}}$$

Feature Functions have the form $f_m(y_n, y_{n-1}, x_n)$:

Need one feature for each state transition (i, j)

$f_{ij}(y, y', x) = 1_{\{y=i\}} 1_{\{y'=j\}}$ and one for each state-observation pair

$f_{io}(y, y', x) = 1_{\{y=i\}} 1_{\{x=o\}}$

- Note that Z cancels out

CRF definition

- A *linear chain* CRF is a distribution $p(\mathbf{Y}|\mathbf{X})$ that takes the form

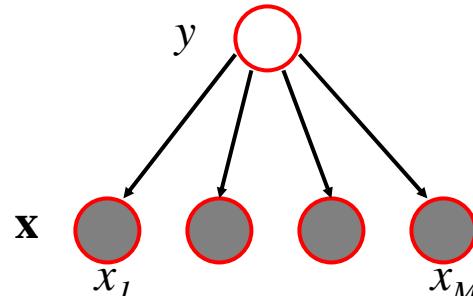
$$p(\mathbf{Y} | \mathbf{X}) = \frac{1}{Z(\mathbf{X})} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, \mathbf{x}_n) \right\}$$

- Where $Z(X)$ is an instance specific normalization function

$$Z(\mathbf{X}) = \sum_y \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, \mathbf{x}_n) \right\}$$

Functional Models

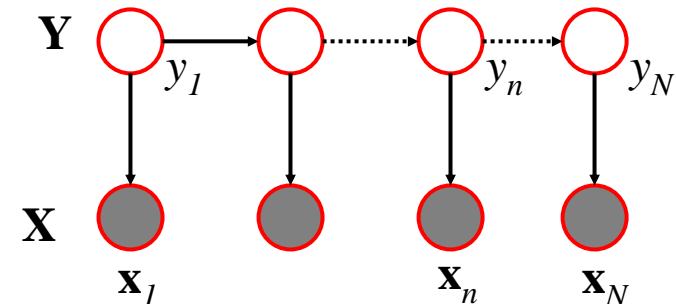
Naïve Bayes Classifier



$$p(y, \mathbf{x}) = p(y) \prod_{m=1}^M p(x_m | y)$$

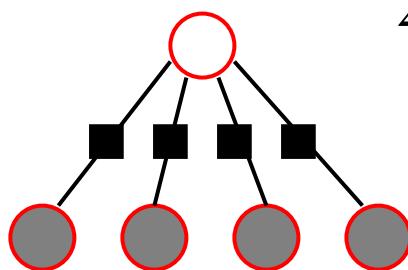
GENERATIVE

Hidden Markov Model



$$p(\mathbf{Y}, \mathbf{X}) = \prod_{n=1}^N p(y_n | y_{n-1}) p(x_n | y_n)$$

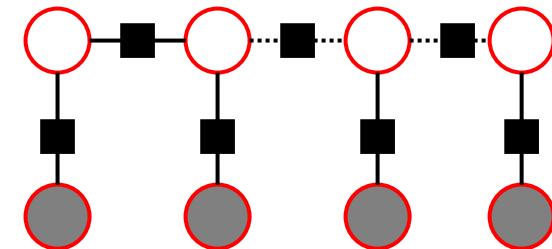
$$p(y | \mathbf{x}) = \frac{\exp \left\{ \sum_{m=1}^M \lambda_m f_m(y, \mathbf{x}) \right\}}{\sum_{y'} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y', \mathbf{x}) \right\}}$$



Logistic Regression

DISCRIMINATIVE

$$p(\mathbf{Y} | \mathbf{X}) = \frac{\exp \left\{ \sum_{m=1}^M \lambda_m f_m(y_n, y_{n-1}, \mathbf{x}_n) \right\}}{\sum_{y'} \exp \left\{ \sum_{m=1}^M \lambda_m f_m(y'_n, y'_{n-1}, \mathbf{x}_n) \right\}}$$



Conditional Random Field

4. CRF vs Other Models

- **Generative Models**
 - Relax assuming conditional independence of observed data given the labels
 - Can contain arbitrary feature functions
 - Each feature function can use entire input data sequence. Probability of label at observed data segment may depend on any past or future data segments.
- **Other Discriminative Models**
 - Avoid limitation of other discriminative Markov models biased towards states with few successor states.
 - Single exponential model for joint probability of entire sequence of labels given observed sequence.
 - Each factor depends only on previous label, and not future labels. $P(y | x) = \text{product of factors, one for each label.}$

NLP: Part Of Speech Tagging

For a sequence of words $w = \{w_1, w_2, \dots, w_n\}$ find syntactic labels s for each word:

$w =$ The quick brown fox jumped over the lazy dog

$s =$ DET VERB ADJ NOUN-S VERB-P PREP DET ADJ NOUN-S

Model	Error
HMM	5.69%
CRF	5.55%

Baseline is already 90%

- Tag every word with its most frequent tag
- Tag unknown words as nouns

Per-word error rates for POS tagging on the Penn treebank

Table Extraction

To label lines of text document:

Whether part of table and its role in table.

Finding tables and extracting information is necessary component of data mining, question-answering and IR tasks.

HMM	CRF
89.7%	99.9%

Shallow Parsing

- Precursor to full parsing or information extraction
 - Identifies non-recursive cores of various phrase types in text
- Input: words in a sentence annotated automatically with POS tags
- Task: label each word with a label indicating
 - word is outside a chunk (O), starts a chunk (B), continues a chunk (I) NP chunks

Rockwell International Corp. 's Tulsa unit said it signed a tentative agreement extending
its contract with Boeing Co. to provide structural parts for Boeing 's 747 jetliners.

CRFs beat all reported single-model NP chunking results on standard evaluation dataset

Model	F score
CRF	94.38%
Generalized winnow	93.89%
Voted perceptron	94.09%
MEMM	93.70%

Handwritten Word Recognition

Given word image and lexicon, find most probable lexical entry

Algorithm Outline

- Oversegment image
segment combinations are potential characters
- Given $y =$ a word in lexicon, $s =$ grouping of segments,
 $x =$ input word image features
- Find word in lexicon and segment grouping that maximizes $P(y, s | x)$,

CRF Model

$$P(y | x, \theta) = \frac{e^{\psi(y, x; \theta)}}{\sum_{y'} e^{\psi(y', x; \theta)}}$$

$$\psi(y, x; \theta) = \sum_{j=1}^m \left(A(j, y_j, x; \theta^s) + \sum_{(j,k) \in E} I(j, k, y_j, y_k, x, \theta^t) \right)$$

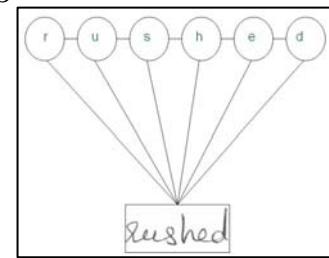
where $y_i \in \{a-z, A-Z, 0-9\}$, θ : model parameters

Association Potential (state term)

$$A(j, y_j, x; \theta^s) = \sum_i (f_i^s(j, y_j, x) \cdot \theta_{ij}^s)$$

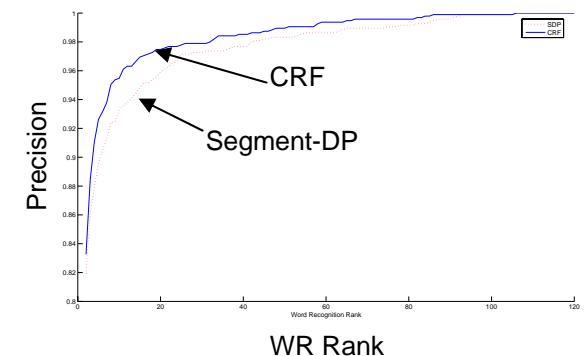
Interaction Potential

$$I(j, k, y_j, y_k, x; \theta^t) = \sum_i (f_i^t(j, k, y_j, y_k, x) \cdot \theta_{ijk}^t)$$



Feature	Description
Position	Position of character in the lexicon normalized by the length.
Place	Whether the character appears in the beginning, middle or at the end.
Height	Height(pixel)s) of the candidate character image
Width	Width(pixel)s) of the candidate character image
Aspect ratio	Ratio of the height of the character to its width
Euclidean Distance	Euclidean Distance of the character to its prototype cluster center
Manhattan Distance	Manhattan Distance of the character to its prototype cluster center
Tanimoto Distance	Tanimoto Distance of the character to its prototype cluster center
Inner Product	Inner Product of the character WMR features and its prototype cluster center features
KNN Distance	Distance of the character from its 5 nearest prototype images
Height Deviation	Deviation of the height of the character from its expected height
Top Deviation	Deviation of the position of the top of the character from its expected top position
Bottom Deviation	Deviation of the position of the top of the character from its expected top position

Feature	Description
Label	Label of the character pair eg. a,b or q,u etc.
Vertical overlap	Vertical overlap(pixel)s) between the two candidate character images
Height difference	Difference in Height(pixel)s) between the candidate character images
Width difference	Difference in Width(pixel)s) between the candidate character images
Aspect ratio difference	Difference in aspect ratio between the candidate character images
Bigram width	Sum of individual widths(pixel)s) of the candidate character images.



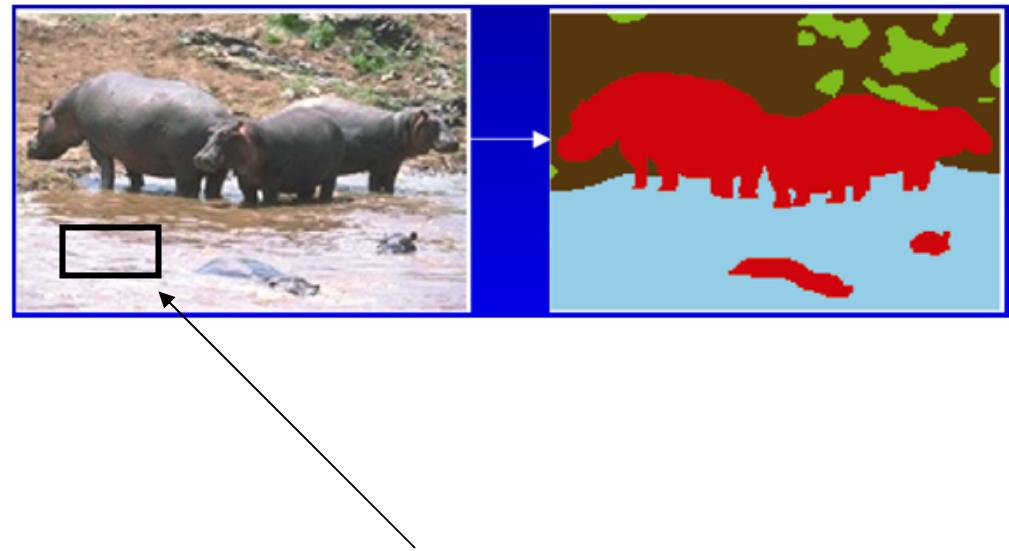
Document Zoning (labeling regions) error rates

	CRF	Neural Network	Naive Bayes
Machine Printed Text	1.64%	2.35%	11.54%
Handwritten Text	5.19%	20.90%	25.04%
Noise	10.20%	15.00%	12.23%
Total	4.25%	7.04%	12.58% 32

5. CRFs in Computer Vision

Applications in Computer Vision

- **Image Modeling**
 - Labeling regions in image (image segmentation)
 - Object detection and recognition (important areas)
 - Sign detection in natural images
 - Natural scene categorization
 - Gesture Recognition
- **Image retrieval**
- **Object/motion segmentation in video scene**



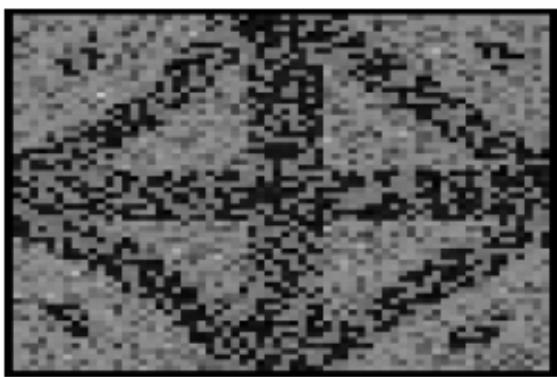
Small patch is water in one context and sky in another



(a)



(b)



(c)

Various tasks in computer vision that require explicit consideration of spatial dependencies!

(a) Segmentation and labeling of input image in meaningful regions.

(b) Detection of structured textures such as buildings.

(c) restore images corrupted by noise.

Image Modeling

Image Labeling:

classifying every image pixel/patch into a finite set of class

Object Recognition:

requires semantic image content understanding based on labeling

CRF models:

- Directly predict the segmentation/labeling given the observed image
- Incorporate arbitrary functions of the observed features into the training process

CRFs for Image Modeling

- **Discriminative Random Fields (DRF)**
 - S. Kumar and M. Hebert. Discriminative fields for modeling spatial dependencies in natural images. 2003
- **Multiscale Conditional Random Fields (MCRF)**
 - X. He, R. S. Zemel, and M. A' . Carreira-Perpin˜a'n. Multiscale conditional random fields for image labeling. 2004.
- **Hierarchical Conditional Random Field (HCRF)**
 - S. Kumar and M. Hebert. A hierarchical field framework for unified context-based classification. 2005
 - Jordan Reynolds and Kevin Murphy. Figure-ground segmentation using a hierarchical conditional random field. 2007
- **Tree Structured Conditional Random Fields (TCRF)**
 - P. Awasthi, A. Gagranı, and B. Ravindran, Image Modeling using Tree Structured Conditional Random Fields. 2007

Discriminative Random Fields (DRF)

CRF - normalized product of potential functions

$$p(\mathbf{y}|\mathbf{x}) = \frac{1}{Z(\mathbf{x})} \exp \left(\sum_j \lambda_j \underbrace{t_j(y_{i-1}, y_i, \mathbf{x}, i)}_{\text{transition feature function}} + \sum_k \mu_k \underbrace{s_k(y_i, \mathbf{x}, i)}_{\text{state feature function}} \right)$$

In the DRF framework → Interaction potential

Association potential

take into account the neighborhood interaction of the data

enforce adaptive data-dependent smoothing over the label field.

Proposed DRF model was applied to the task of detecting man-made structures in natural scenes.



(a) Input image



(b) Logistic



(c) MRF



(d) DRF

structure detection by DRF

(label each image site as *structured* or *nonstructured*)

For similar detection rates, DRF reduces the false positives considerably.

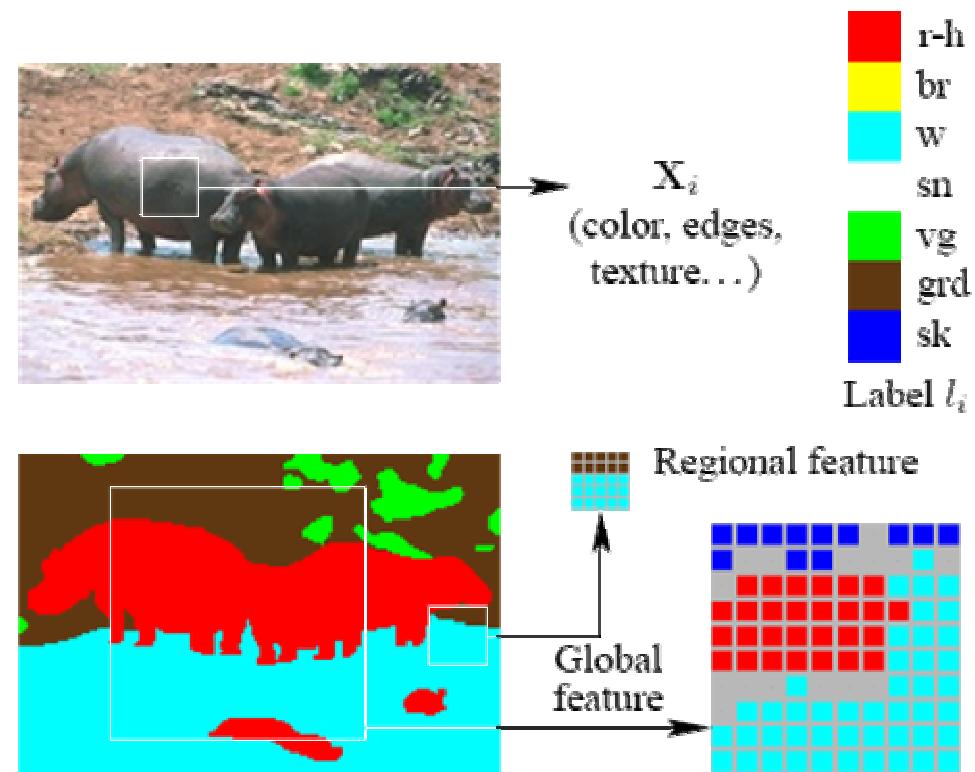
Though the model outperforms traditional MRFs, it is not strong enough to capture long range correlations among the labels due to the rigid lattice based structure which allows for only pairwise interactions

Multiscale Conditional Random Fields (MCRF)

Combination of three different classifiers operating at local, regional and global contexts respectively.

Two main drawbacks:

- Including additional classifiers operating at different scales into the mCRF framework introduces a large number of model parameters.
- The model assumes conditional independence of hidden variables given the label field.



Hierarchical CRF (HCRF)



Scene context is important in different domains to achieve good classification even though the local appearance is impoverished

- region-region interaction
- object-region interaction
- object-object interaction

HCRF: Incorporate the local as well as the global context of any of the three types in a single model

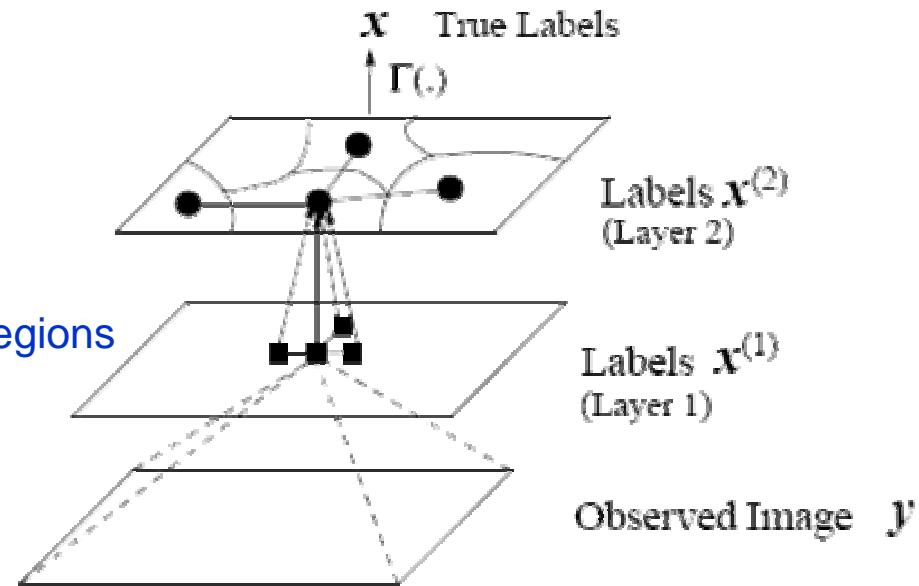
Hierarchical CRF (HCRF)

Layer 1

short-range: pixelwise label

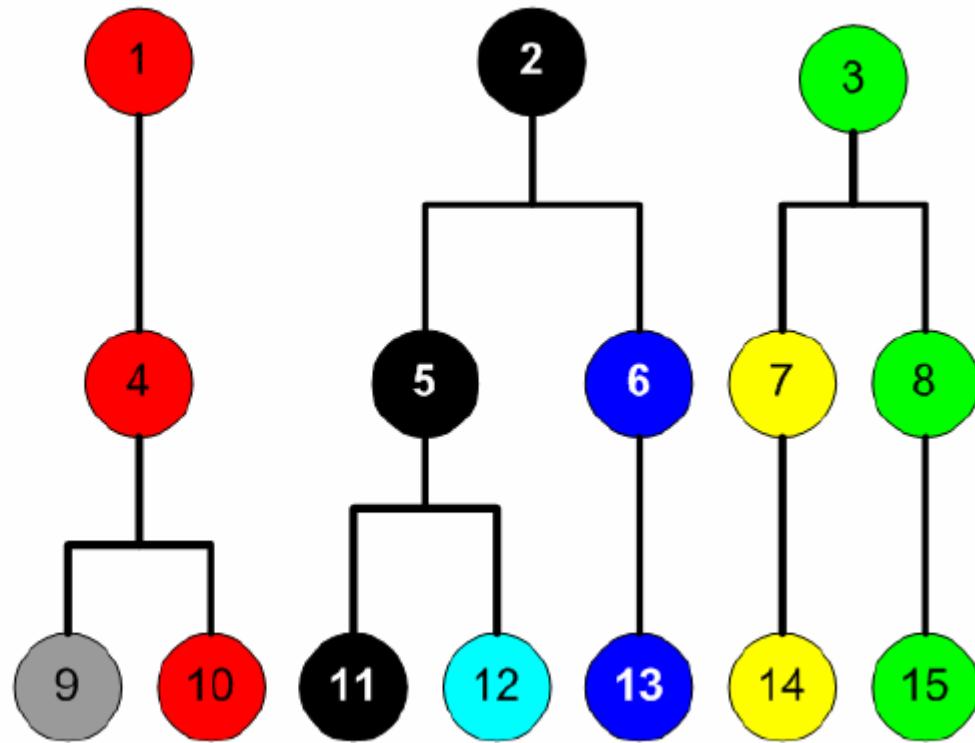
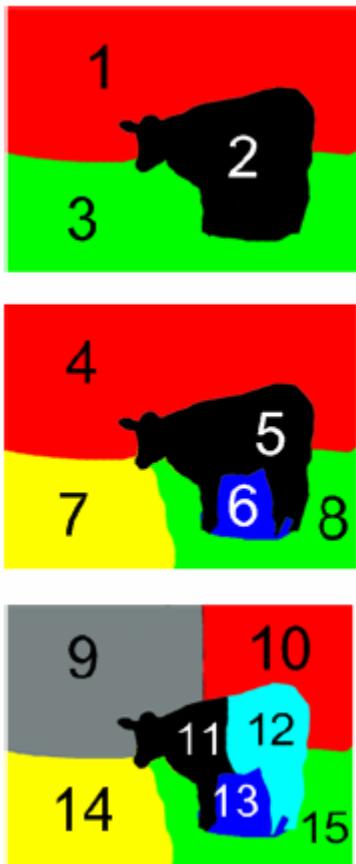
Layer 2

long-range : contextual object or regions

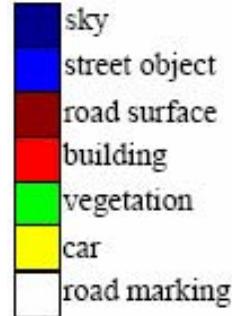


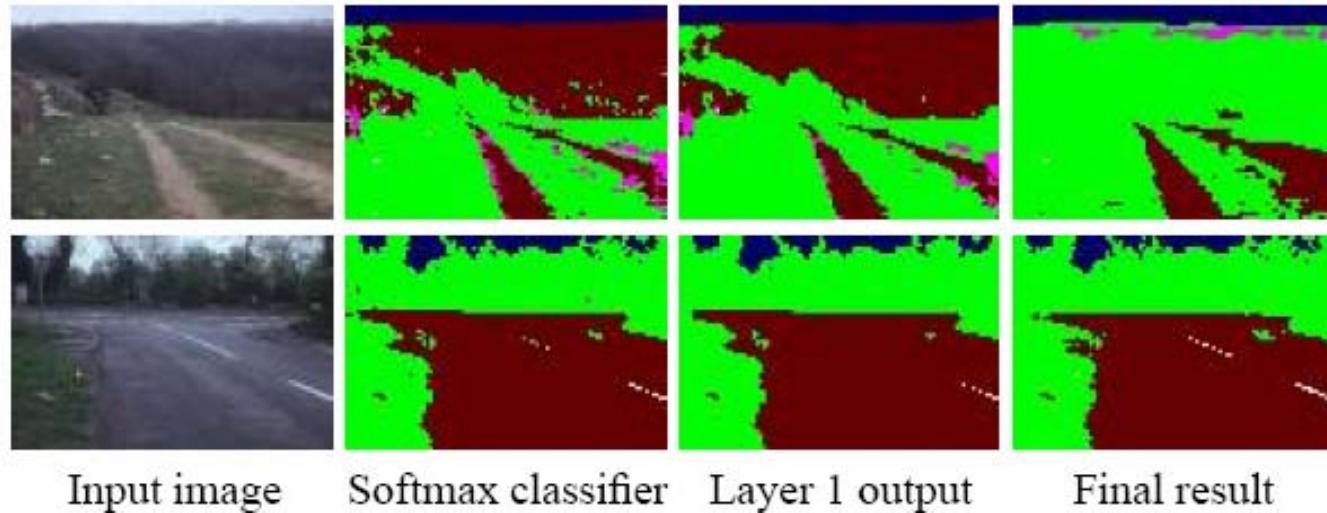
- exploit different levels of contextual information in images for robust classification.
- Each layer is modeled as a conditional field that allows one to capture arbitrary observation dependent label interactions

Hierarchical CRF (HCRF)

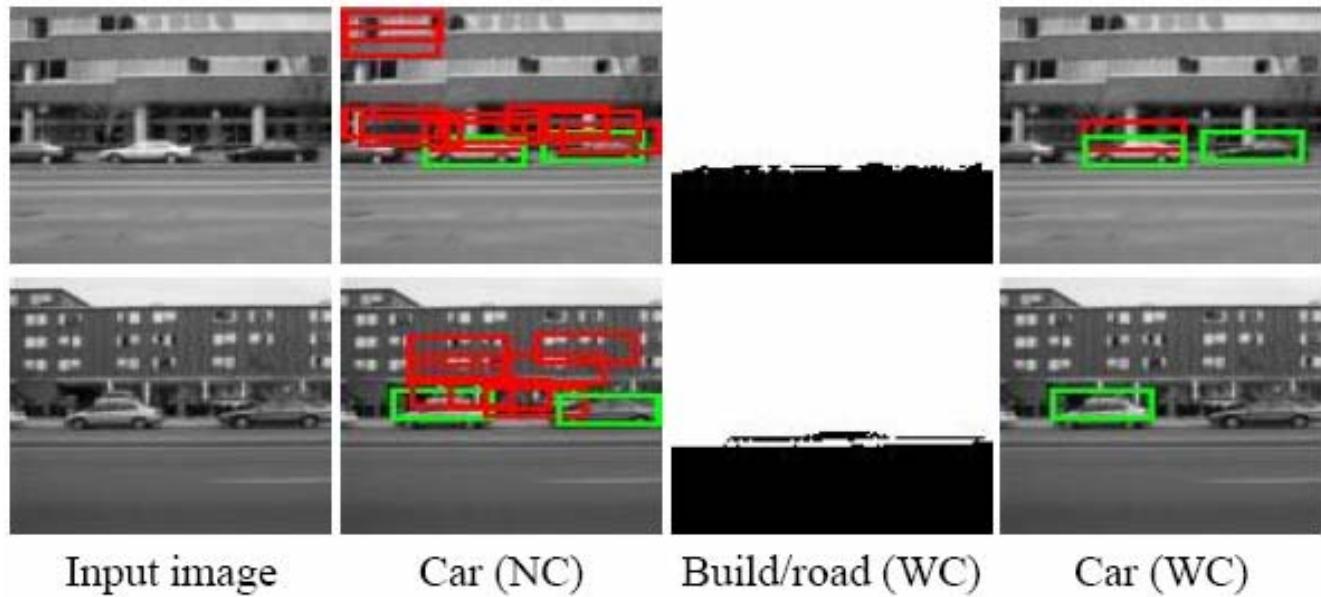


region-region interaction


sky
street object
road surface
building
vegetation
car
road marking



object-region interaction



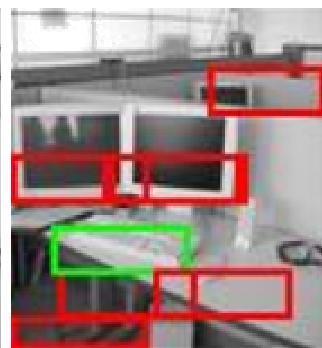
object-object interaction



Input image



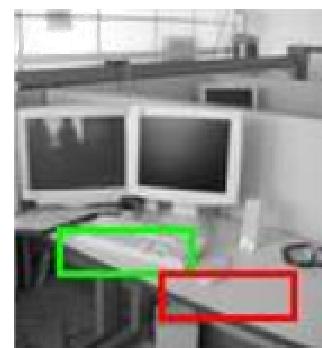
Monitor (NC)



Keyboard (NC)



Mouse (NC)



Keyboard (WC)

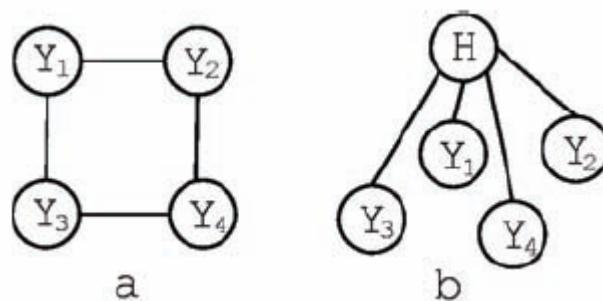


Mouse (WC)

Tree-Structured CRF (TCRF)

Combine advantages of hierarchical models and discriminative approaches in a single framework

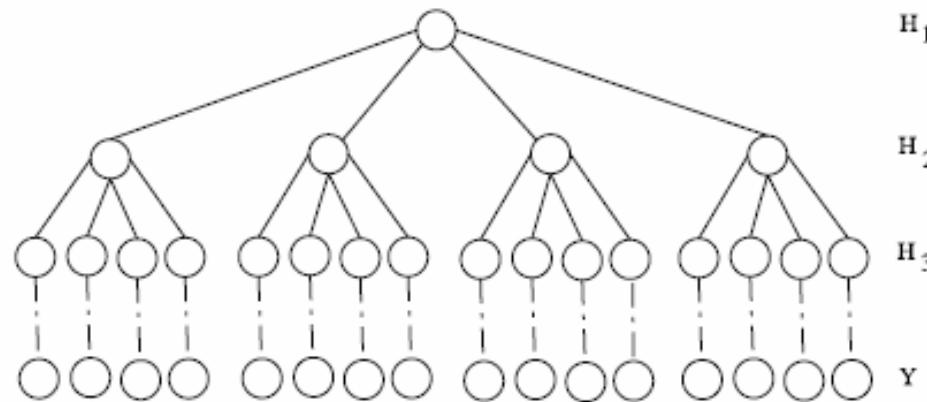
model the association between the labels of 2×2 neighborhood of pixels



- introduce a weight vector for every edge which represents the compatibility between the labels of the pairwise nodes
- introduce a hidden variable H which is connected to all the 4 nodes. For every value which variable H takes, it induces a probability distribution over the labels of the nodes connected to it

Tree-Structured CRF (TCRF)

- Dividing the whole image into regions of size $m \times m$
- introducing a hidden variable for each of them gives a layer of hidden variables over the given label field
- each label node is associated with a hidden variable in the layer above



Tree structured CRF with $m=2$

Similar to CRFs, the conditional probability $p(y, H|x)$ of a TCRF factors into a product of potential functions

Tree-Structured CRF (TCRF)



Model	Classification Accuracy (%)
LR	69.82
DRF	72.54 ¹
TCRF	85.83

Object detection

Model	Classification Accuracy (%)
LR	65.03
mCRF	74.3
TCRF	77.76

Image labeling

- long range correlations among non-neighboring pixels can be easily modeled as associations among the hidden variables in the layer above.
- tree structure allows inference to be carried out in a time linear in the number of pixels

Sign Detection in Natural Images

Weinman, J. Hanson, A. McCallum, A. Sign detection in natural images with conditional random fields. 2004.

Calculates a joint labeling of image patches, rather than labeling patches independently and use CRF to learn the characteristics of regions that contain text.



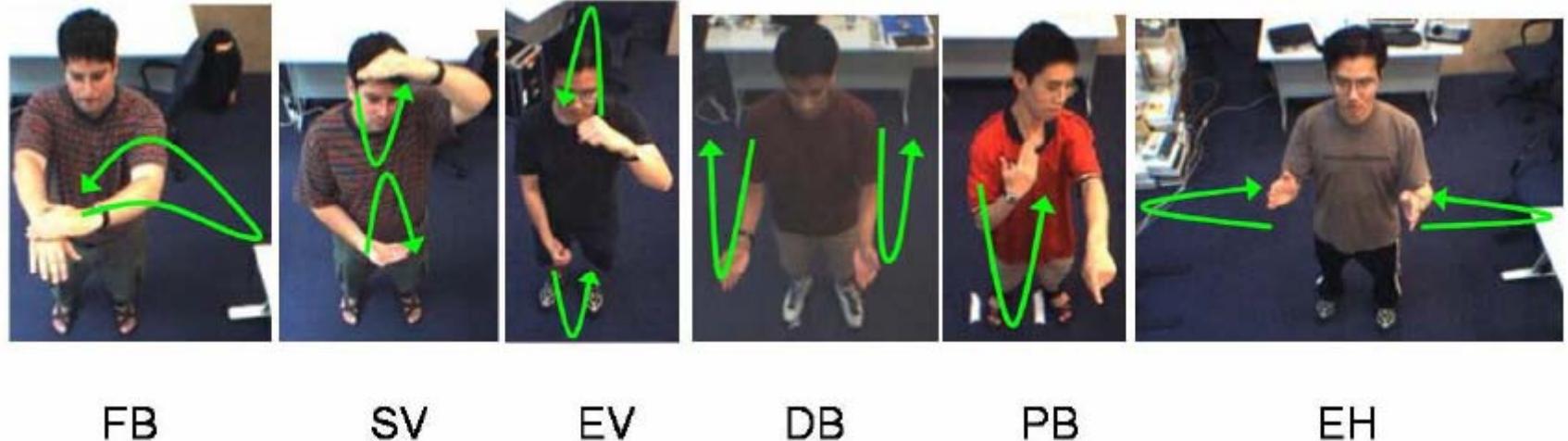
MaxEnt



adding CRF

Human Gesture Recognition

Wang Quattoni, A. Morency, L.-P. Demirdjian, D. Darrell, T..Hidden Conditional Random Fields for Gesture Recognition. Computer Science and Artificial Intelligence Laboratory, MIT. 2006.

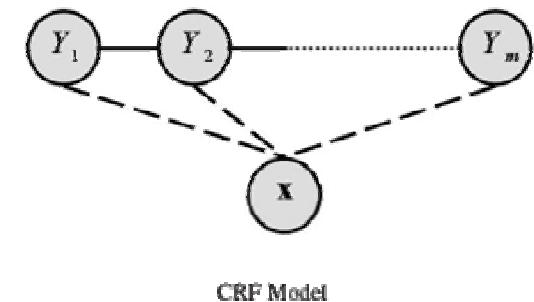


These gesture classes are: FB - Flip Back, SV - Shrink Vertically, EV - Expand Vertically, DB - Double Back, PB - Point and Back, EH – Expand Horizontally

Human Gesture Recognition

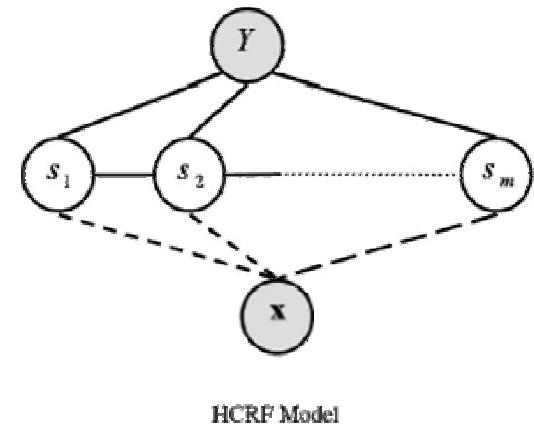
Hidden Conditional Random Fields

$$P(y \mid \mathbf{x}, \theta) = \sum_{\mathbf{s}} P(y, \mathbf{s} \mid \mathbf{x}, \theta) = \frac{\sum_{\mathbf{s}} e^{\Psi(y, \mathbf{s}, \mathbf{x}; \theta)}}{\sum_{y' \in \mathcal{Y}, \mathbf{s} \in S^m} e^{\Psi(y', \mathbf{s}, \mathbf{x}; \theta)}}$$

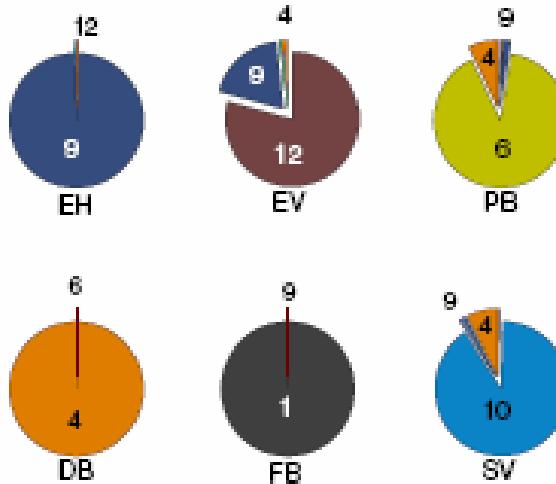


$\mathbf{s} = \{s_1, s_2, \dots, s_m\}$, is the set of hidden states in the model, captures certain underlying structure of each class.

$\Psi(y, \mathbf{s}, \mathbf{x}; \theta)$ is potential function, measures the compatibility between a label, a set of observations and a configuration of the hidden states.



Hidden states distribution

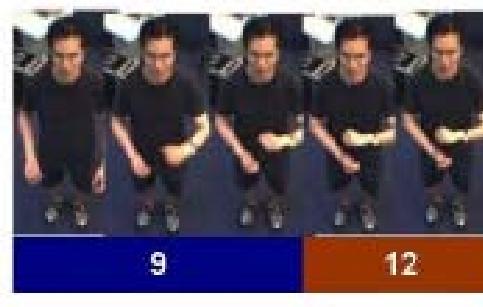


EH



Hidden States

EV



Hidden States

PB



Hidden States

DB



Hidden States

FB



Hidden States

SV

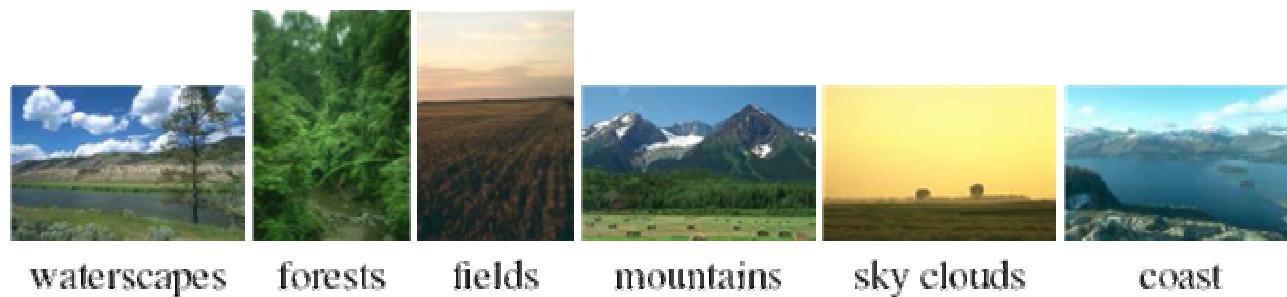
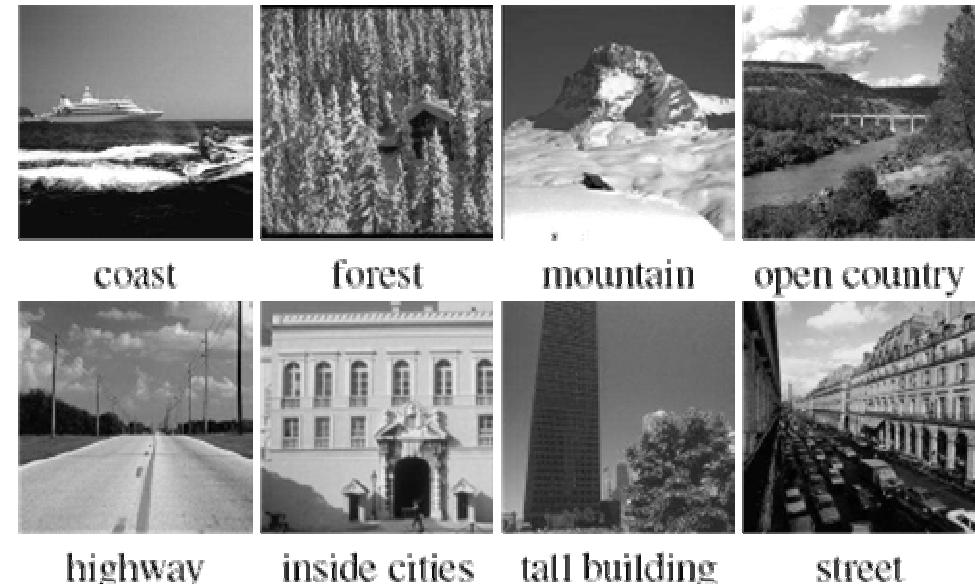


Hidden States

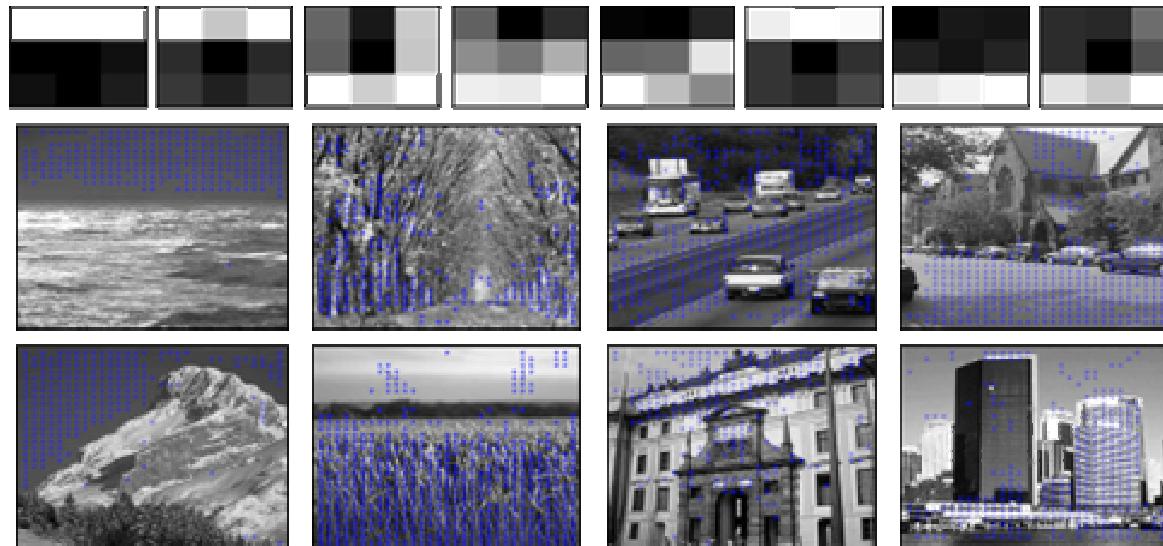
Natural Scene Categorization

Wang, Y., Gong, S., Conditional Random Field for Natural Scene Categorization. 2007.

Classification Oriented
Conditional Random Field
(COCRFS)

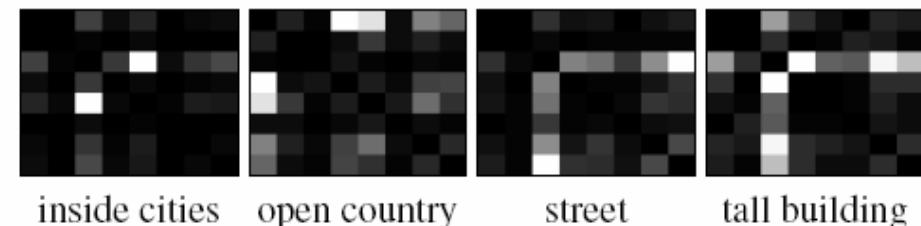


Natural Scene Categorization



COCRF discovers the spatial layout distribution of local patches and their pairwise interaction for a category

coast	forest	highway	street
mountain	open country	inside cities	tall building
topic=6	topic=3	topic=8	topic=5



54

pairwise interaction potential between topics for four categories

6. Summary

- Generative and Discriminative methods are two broad approaches: former involve modeling, latter directly solve classification
- For classification
 - Naïve Bayes and Logistic Regression form a generative discriminative pair
- For sequential data
 - HMM and CRF are corresponding pair
- CRF performs better in language related tasks
- Generative models are more elegant, have explanatory power

7. References

1. C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006
2. C. Sutton and A. McCallum, *An Introduction to Conditional Random Fields for Relational Learning*
3. S. Shetty, H. Srinivasan and S. N. Srihari, *Handwritten Word Recognition using CRFs*, ICDAR 2007
4. S. Shetty, H. Srinivasan and S. N. Srihari, *Segmentation and Labeling of Documents using CRFs*, SPIE-DRR 2007
5. X. He, R. Zennel and M.A. Carreira-Perpinan, *Multiscale Conditional Random Fields for Image Labeling*

Other Topics in Sequential Data

- **Sequential Data and Markov Models:**

<http://www.cedar.buffalo.edu/~srihari/CSE574/Chap11/Ch1.1-MarkovModels.pdf>

- **Hidden Markov Models:**

<http://www.cedar.buffalo.edu/~srihari/CSE574/Chap11/Ch1.2-HiddenMarkovModels.pdf>

- **Extensions of HMMs:**

<http://www.cedar.buffalo.edu/~srihari/CSE574/Chap11/Ch1.3-HMMExtensions.pdf>

- **Linear Dynamical Systems:**

<http://www.cedar.buffalo.edu/~srihari/CSE574/Chap11/Ch1.4-LinearDynamicalSystems.pdf>