Machine Learning: Generative and Discriminative Models

Sargur N. Srihari

srihari@cedar.buffalo.edu

Machine Learning Course:

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

Outline of Presentation

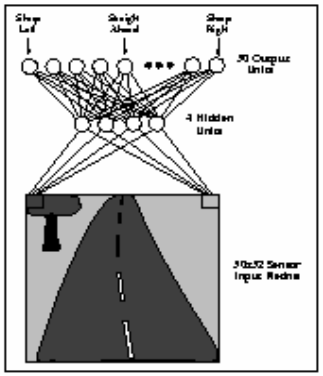
1. What is Machine Learning?
   - ML applications, ML as Search
2. Generative and Discriminative Taxonomy
3. Generative-Discriminative Pairs
   - Classifiers: Naïve Bayes and Logistic Regression
   - Sequential Data: HMMs and CRFs
4. Performance Comparison in Sequential Applications
   - NLP: Table extraction, POS tagging, Shallow parsing,
   - Handwritten word recognition, Document analysis
5. Advantages, disadvantages
6. Summary
7. References

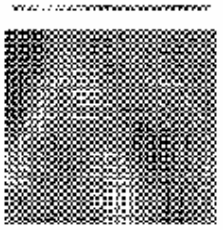
1. Machine Learning

- Programming computers to use example data or past experience
- Well-Posed Learning Problems
  - A computer program is said to learn from *experience E*
  - with respect to *class of tasks T* and *performance measure P*,
  - if its performance at tasks *T*, as measured by *P*, improves with experience *E*.

Problems Too Difficult To Program by Hand

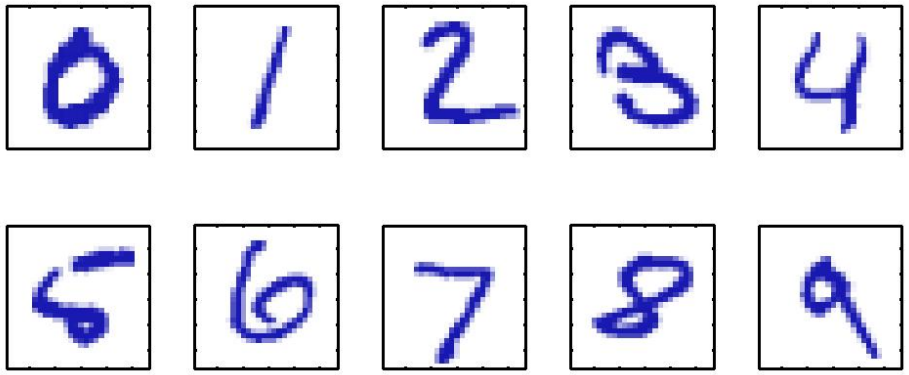
- Learning to drive an autonomous vehicle
  - Train computer-controlled vehicles to steer correctly
  - Drive at 70 mph for 90 miles on public highways
  - Associate steering commands with image sequences





Task *T*: driving on public, 4-lane highway using vision sensors
Perform measure *P*: average distance traveled before error
(as judged by human overseer)
Training *E*: sequence of images and steering commands recorded while observing a human driver

Example Problem: Handwritten Digit Recognition



Wide variability of same numeral

- Handcrafted rules will result in large no of rules and exceptions
- Better to have a machine that learns from a large training set

Other Applications of Machine Learning

- Recognizing spoken words
  - Speaker-specific strategies for recognizing phonemes and words from speech
  - Neural networks and methods for learning HMMs for customizing to individual speakers, vocabularies and microphone characteristics
- Search engines
  - Information extraction from text
- Data mining
  - Very large databases to learn general regularities implicit in data
  - Classify celestial objects from image data
    - Decision tree for objects in sky survey: 3 terabytes

ML as Searching Hypotheses Space

- Very large space of possible hypotheses to fit:
  - observed data and
  - any prior knowledge held by the observer

Method	Hypothesis Space
Concept Learning	Boolean Expressions
Decision Trees	All Possible Trees
Neural Networks	Weight Space

ML Methodologies are increasingly statistical

- Rule-based expert systems being replaced by probabilistic generative models
- Example: Autonomous agents in AI
  - ELIZA : natural language rules to emulate therapy session
  - Manual specification of models, theories are increasingly difficult
- Greater availability of data and computational power to migrate away from rule-based and manually specified models to probabilistic data-driven modes

The Statistical ML Approach

1. Data Collection

   Large sample of data of how humans perform the task

2. Model Selection

   Settle on a parametric statistical model of the process

3. Parameter Estimation

   Calculate parameter values by inspecting the data

Using learned model perform:

4. Search

   Find optimal solution to given problem

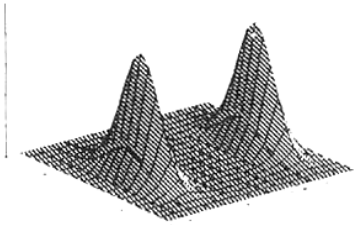
2. Generative and Discriminative Models: An analogy

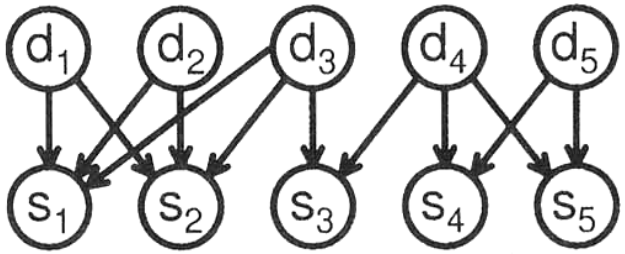
- The task is to determine the language that someone is speaking
- Generative approach:
  - is to learn each language and determine as to which language the speech belongs to
- Discriminative approach:
  - is determine the linguistic differences without learning any language– a much easier task!

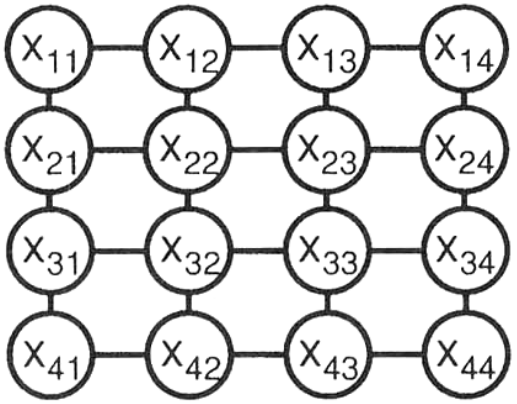
Taxonomy of ML Models

- **Generative Methods**
  - Model class-conditional pdfs and prior probabilities
  - "Generative" since sampling can generate synthetic data points
  - Popular models
    - Gaussians, Naïve Bayes, Mixtures of multinomials
    - Mixtures of Gaussians, Mixtures of experts, Hidden Markov Models (HMM)
    - Sigmoidal belief networks, Bayesian networks, Markov random fields
- **Discriminative Methods**
  - Directly estimate posterior probabilities
  - No attempt to model underlying probability distributions
  - Focus computational resources on given task– better performance
  - Popular models
    - Logistic regression, SVMs
    - Traditional neural networks, Nearest neighbor
    - Conditional Random Fields (CRF)

Generative Models (graphical)







(a) Mixture Model (b) Directed Graphical Model (c) Undirected Graphical Model

Parent node selects between components

Quick Medical Reference -DT

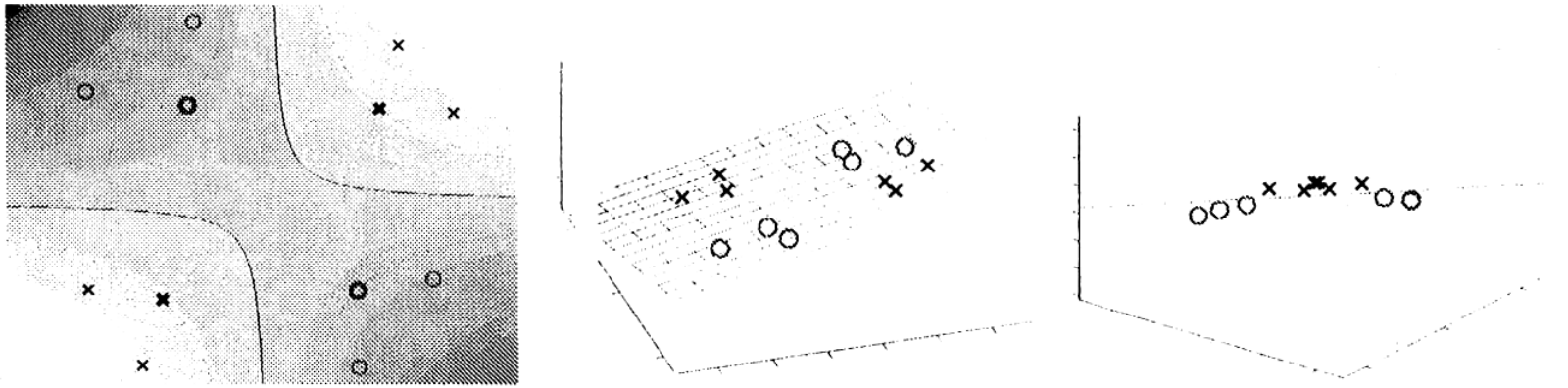
Diagnosing Diseases from Symptoms

Markov Random Field

Successes of Generative Methods

- NLP
  - Traditional rule-based or Boolean logic systems (eg Dialog and Lexis-Nexis) are giving way to statistical approaches (Markov models and stochastic context free grammars)
- Medical Diagnosis
  - QMR knowledge base, initially a heuristic expert systems for reasoning about diseases and symptoms has been augmented with decision theoretic formulation
- Genomics and Bioinformatics
  - Sequences represented as generative HMMs

Discriminative Classifier: SVM



(a) Support Vector Machine (b) Mapped Data (c) Separating Hyperplane

Nonlinear decision boundary

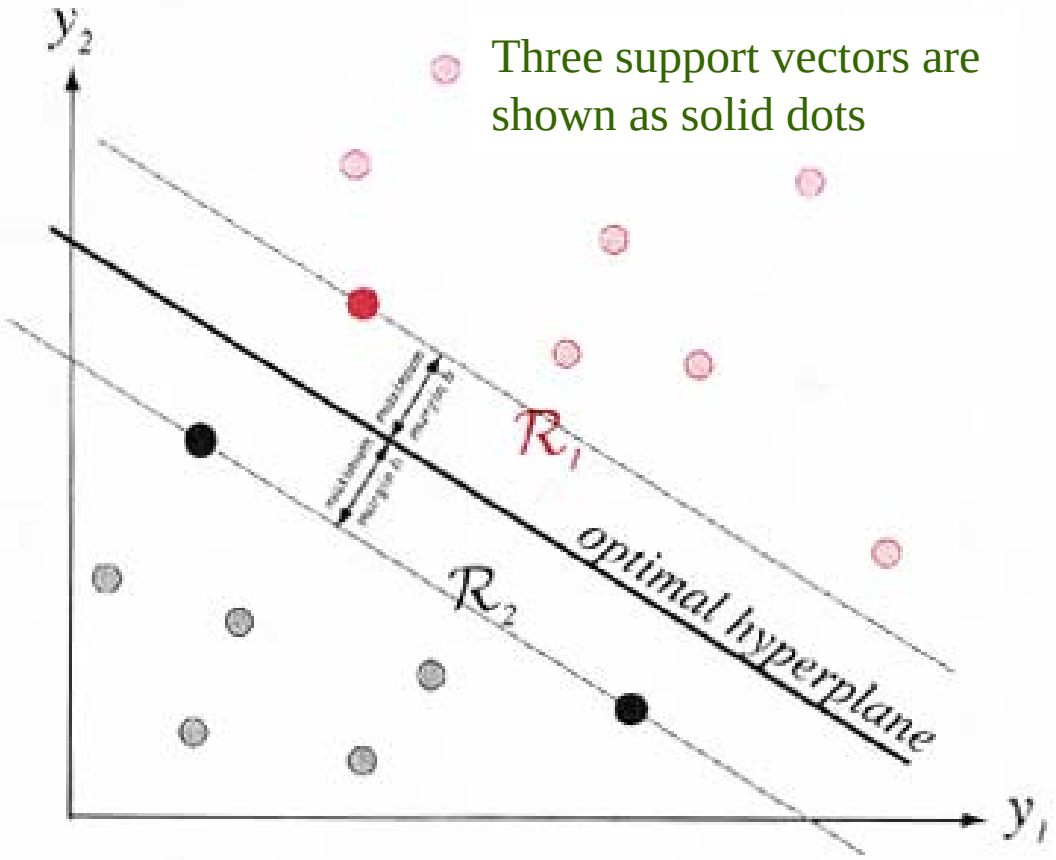
(x1, x2) → (x1, x2, x1x2)

Linear boundary in high-dimensional space

Support Vector Machines

- Support vectors are those nearest patterns at distance *b* from hyperplane
- SVM finds hyperplane with maximum distance from nearest training patterns
- For full description of SVMs see

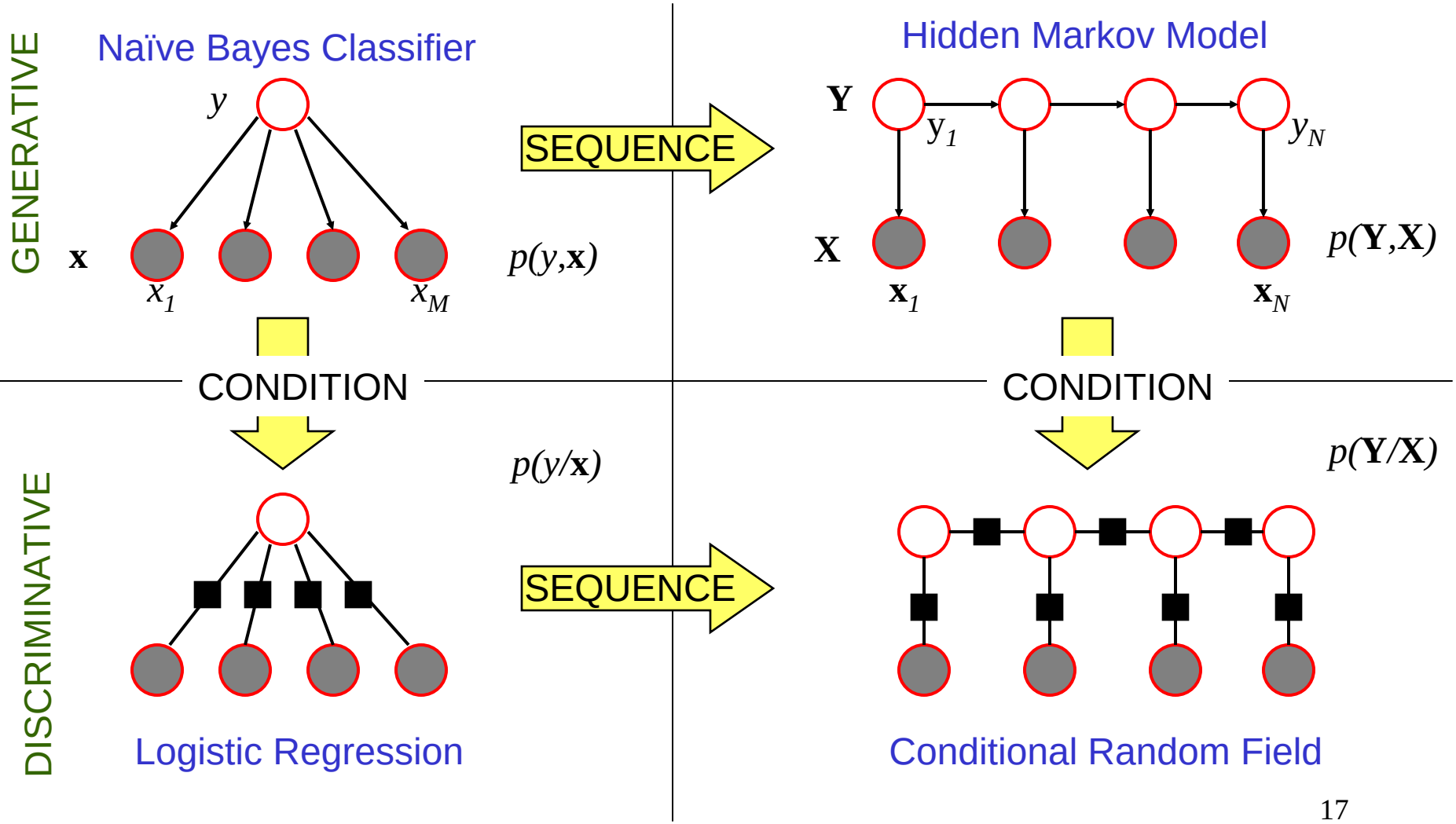
http://www.cedar.buffalo.edu/~srihari/CSE555/SVMs.pdf



3. Generative-Discriminative Pairs

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

Graphical Model Relationship



Generative Classifier: Bayes

- Given variables $\mathrm{x} = (x_1,..,x_M)$ and class variable y
- Joint pdf is $p(\mathrm{x},y)$
  - Called generative model since we can generate more samples artificially
- Given a full joint pdf we can
  - Marginalize $p(y) = \sum_{\mathrm{x}} p(\mathrm{x}, y)$
  - Condition $\boxed{p(y \mid \mathrm{x}) = \frac{p(\mathrm{x}, y)}{p(\mathrm{x})}}$
  - By conditioning the joint pdf we form a classifier
- Computational problem:
  - If x is binary then we need 2^M values
  - If *100* samples are needed to estimate a given probability, $M=10$, and there are two classes then we need *2048* samples

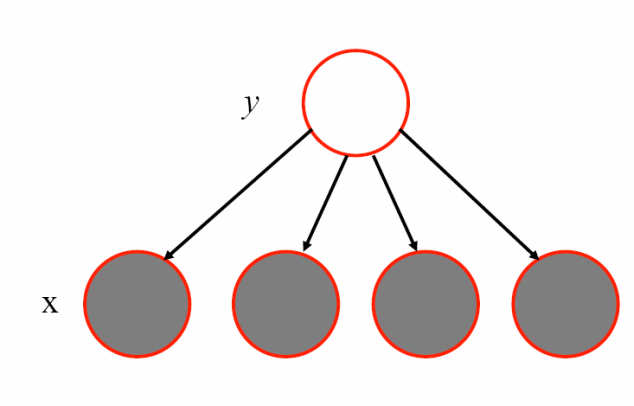
Naïve Bayes Classifier

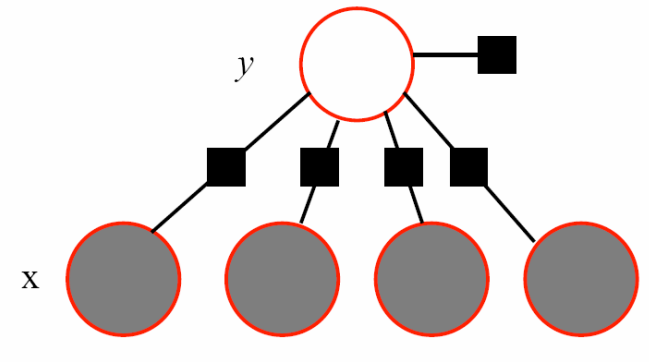
- Goal is to predict single class variable y given a vector of features $\mathrm{x}=(x_1,..,x_M)$
- Assume that once class labels are known the features are independent
- Joint probability model has the form

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

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

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





Discriminative Classifier: Logistic Regression

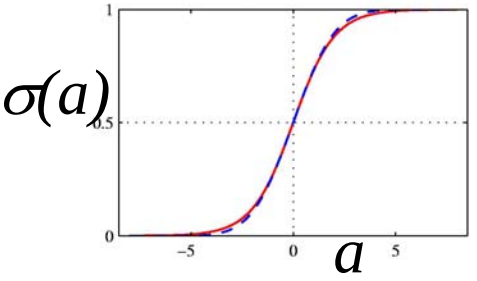
- Feature vector x
- Two-class classification: class variable y has values C_1 and C_2
- *A posteriori* probability $p(C_1|\mathrm{x})$ written as

$$p(C_1|\mathrm{x}) = f(\mathrm{x}) = \sigma(\mathrm{w}^T\mathrm{x}) \quad \text{where}$$

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

- It is known as logistic regression in statistics
  - Although it is a model for classification rather than for regression

Logistic Sigmoid



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

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

Logistic Regression versus Generative Bayes Classifier

- Posterior probability of class variable y is

$$p(C_1 \mid \mathrm{x}) = \frac{p(\mathrm{x} \mid C_1)p(C_1)}{p(\mathrm{x} \mid C_1)p(C_1) + p(\mathrm{x} \mid C_2)p(C_2)}$$

$$= \frac{1}{1+\exp(-a)} = \sigma(a) \quad \text{where} \quad a = \ln \frac{p(\mathrm{x} \mid C_1)p(C_1)}{p(\mathrm{x} \mid C_2)p(C_2)}$$

- 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., $a = \mathrm{w}^T\mathrm{x}$

Logistic Regression Parameters

- For M-dimensional feature space logistic regression has M parameters $\mathrm{w}=(w_1,..,w_M)$
- 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

Multi-class Logistic Regression

- Case of $K>2$ classes

$$p(C_k \mid \mathrm{x}) = \frac{p(x \mid C_k)\,p(C_k)}{\sum_j p(x \mid C_j)\,p(C_j)} = \frac{\exp(a_k)}{\sum_j \exp(a_j)}$$

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

Graphical Model for Logistic Regression

- Multiclass logistic regression can be written as

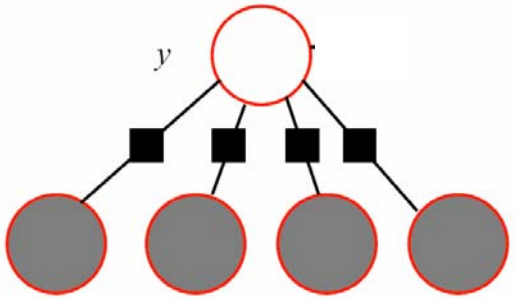
$$p(y \mid \mathrm{x}) = \frac{1}{Z(\mathrm{x})} \exp\left\{ \lambda_y + \sum_{j=1}^{K} \lambda_{yj} x_j \right\} \text{ where}$$

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

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

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

- This notation mirrors the usual notatic for CRFs



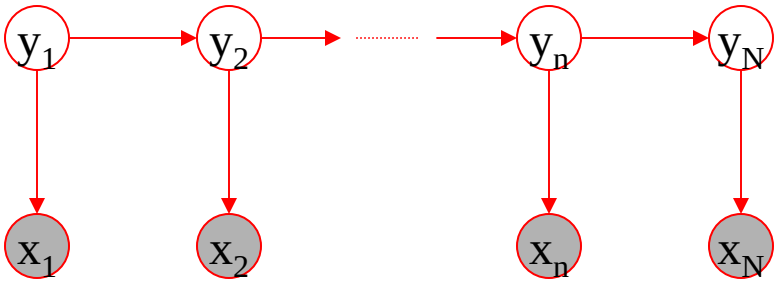
4. Sequence Models

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

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

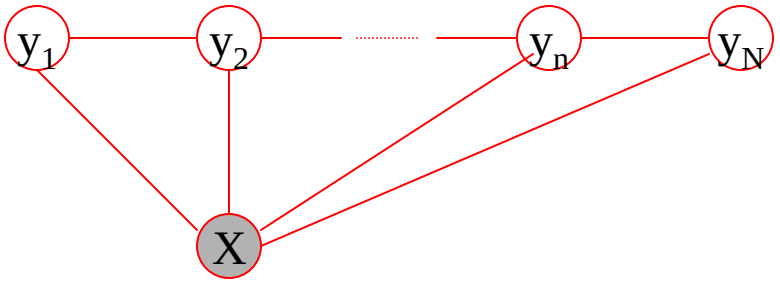


$$p(\mathbf{Y},\mathbf{X}) = \prod_{n=1}^{N} p(y_n \mid y_{n-1})\, p(\mathbf{x}_n \mid y_n)$$

- 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

$$p(\mathrm{x}) = \frac{1}{Z}\prod_{C}\psi_C(\mathrm{x}_C)$$

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 $\psi_C(\mathrm{x}_C) \geq 0$, typically $\psi_C(\mathrm{x}_C) = exp\{-E(x_C)\}$, and

$$Z = \sum_{\mathrm{x}}\prod_{C}\psi_C(\mathrm{x}_C)$$

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 U Y
  - Elements of A that are in A ^ X are denoted x_A
  - Element of A that are in A ^ Y are denoted y_A
- Then undirected graphical model has the form

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

MRF Local Function

- Assume each local function has the form

$$\Psi_A(\mathrm{x}_A, \mathrm{y}_A) = \exp\left\{ \sum_m \theta_{Am} f_{Am}(\mathrm{x}_A, \mathrm{y}_A) \right\}$$

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

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

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

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

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

- Which gives us

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

  - Note that Z cancels out

CRF definition

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

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

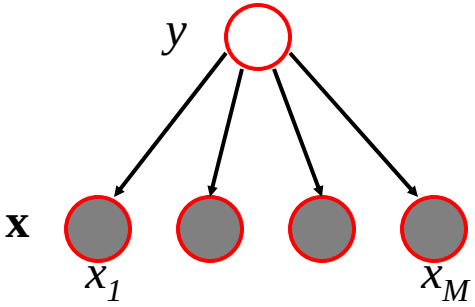
- Where $Z(X)$ is an instance specific normalization function

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

Functional Models

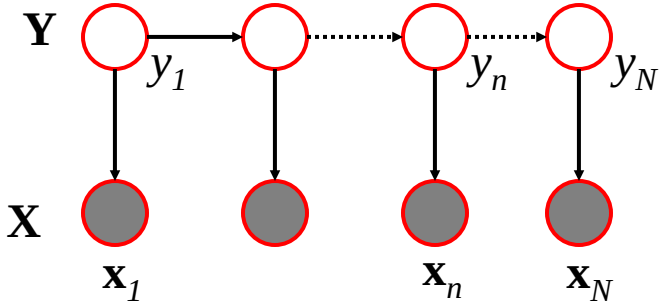
GENERATIVE

Naïve Bayes Classifier



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

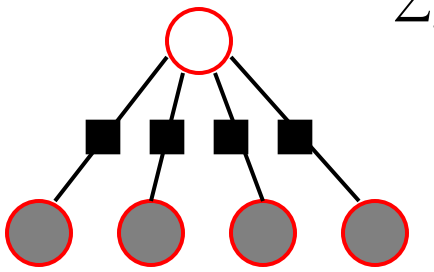
Hidden Markov Model



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

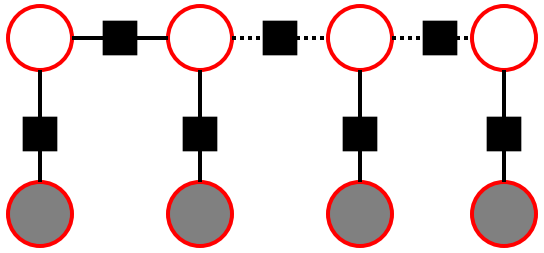
DISCRIMINATIVE

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



Logistic Regression

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



Conditional Random Field

NLP: Part Of Speech Tagging

For a sequence of words $w = \{w_1, w_2, .. w_n\}$ 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 $P(y,s \mid x)$,

CRF Model

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

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

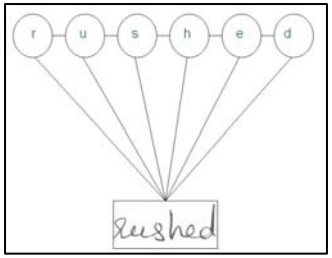
where $y_i \, \varepsilon \, (a\text{-}z, A\text{-}Z, 0\text{-}9\}$, θ: model parameters

Association Potential (state term)

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

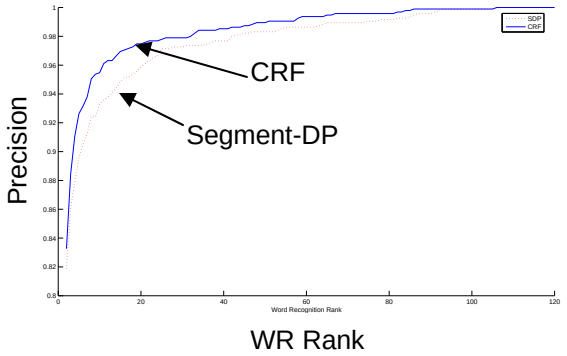
Interaction Potential

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



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(pixels) of the *candidate character image*
Width	Width(pixels) 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(pixels) between the two *candidate character images*
Height difference	Difference in Height(pixels) between the *candidate character images*
Width difference	Difference in Width(pixels) between the *candidate character images*
Aspect ratio difference	Difference in aspect ratio between the *candidate character images*
Bigram width	Sum of individual widths(pixels) of the *candidate character images.*



Document Analysis (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%

5. Advantage of CRF over Other Models

- Other 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. $P(\mathbf{y} \mid \mathbf{x})$ = product of factors, one for each label.

Disadvantages of Discriminative Classifiers

- Lack elegance of generative
  - Priors, structure, uncertainty
- Alternative notions of penalty functions, regularization, kernel functions
- Feel like black-boxes
  - Relationships between variables are not explicit and visualizable

Bridging Generative and Discriminative

- Can performance of SVMs be combined elegantly with flexible Bayesian statistics?
- Maximum Entropy Discrimination marries both methods
  - Solve over a distribution of parameters (a distribution over solutions)

6. Summary

- Machine learning algorithms have great practical value in a variety of application domains
  - A well-defined learning problem requires a well-specified task, performance metric, and source of experience
- Generative and Discriminative methods are two-broad approaches:
  - former involve modeling, latter directly solve classification
- Generative and Discriminative Method Pairs
  - Naïve Bayes and Logistic Regression are a corresponding pair for classification
  - HMM and CRF are a corresponding pair for sequential data
- CRF performs better in language related tasks
- Generative models are more elegant, have explanatory power

7. References

1. T. Mitchell, *Machine Learning*, McGraw-Hill, 1997
2. C. Bishop, *Pattern Recognition and Machine Learning*, Springer, 2006
3. T. Jebarra, *Machine Learning: Discriminative and Generative*, Kluwer, 2004
4. R.O. Duda, P.E. Hart and D. Stork, *Pattern Classification*, 2nd Ed, Wiley 2002
5. C. Sutton and A. McCallum, *An Introduction to Conditional Random Fields for Relational Learning*
6. S. Shetty, H. Srinivasan and S. N. Srihari, *Handwritten Word Recognition using CRFs*, ICDAR 2007
7. S. Shetty, H.Srinivasan and S. N. Srihari, *Segmentation and Labeling of Documents using CRFs*, SPIE-DRR 2007