Generative and Discriminative Models

Jie Tang
Department of Computer Science & Technology
Tsinghua University
2012

ML as Searching Hypotheses Space

- ML Methodologies are increasingly statistical
  - Rule-based expert systems being replaced by probabilistic generative models
  - Example: Autonomous agents in AI
  - Greater availability of data and computational power to migrate away from rule-based and manually specified models to probabilistic data-driven modes

Method	Hypothesis Space
Concept learning	Boolean expressions
Decision trees	All possible trees
Neural Networks	Weight space
Transfer learning	Different spaces

Generative and Discriminative Models

- An example task: determining the language that someone is speaking
- Generative approach:
  - is to learn each language and determine as to which language the speech belongs.
- Discriminative approach:
  - is determine the linguistic differences without learning any language.

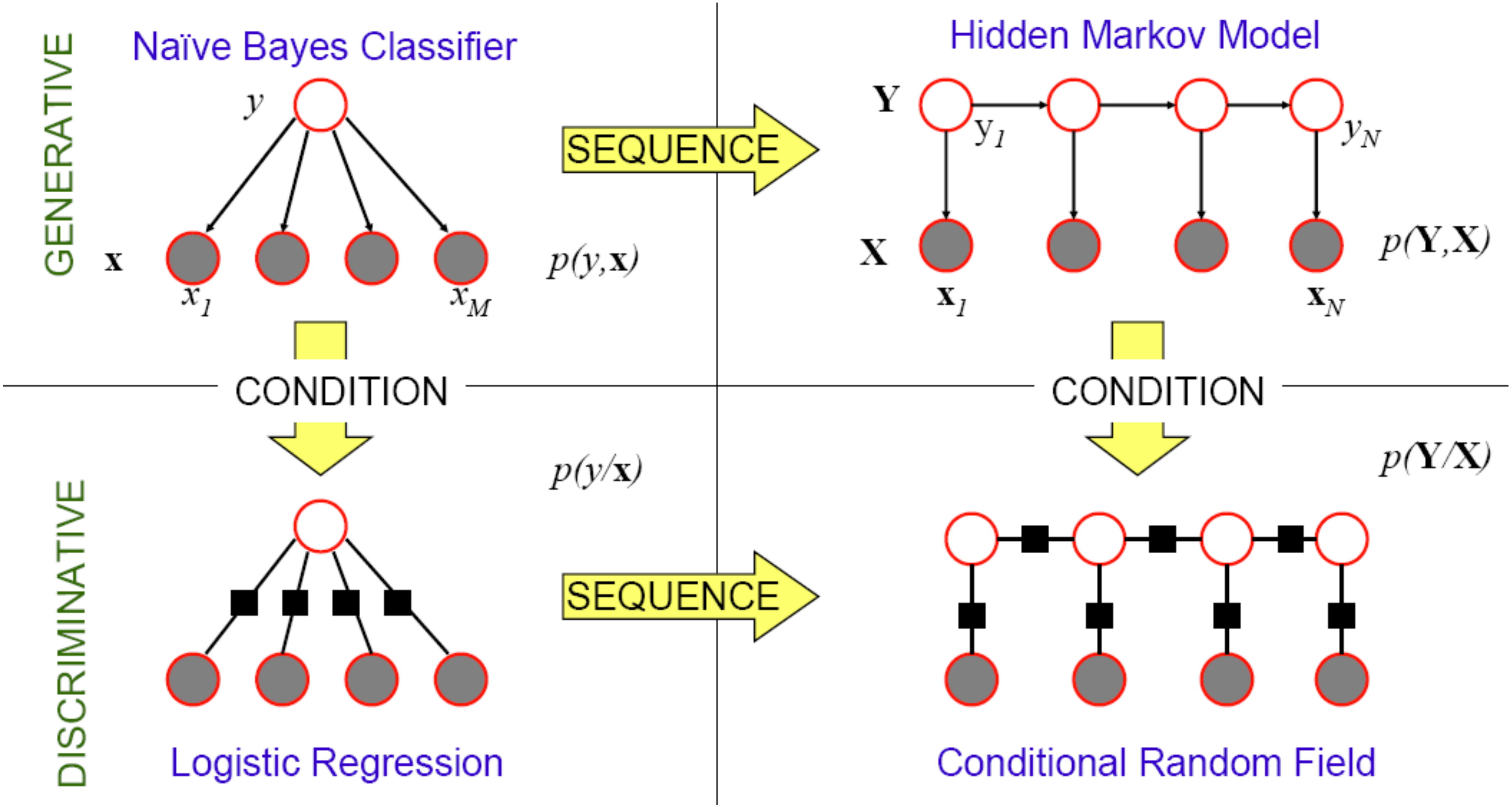
Generative and Discriminative Models

- **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)
    - Sigmoid 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 and Discriminative Pairs

- Data point-based
  - Naïve Bayes and Logistic Regression form a *generative-discriminative* pair for classification
- Sequence-based
  - HMMs and linear-chain CRFs for sequential data

Graphical Model Relationship



Generative Classifier: Naïve Bayes

- Given variables $x=(x_1,..,x_M)$ *and class variable y*
- Joint pdf is $p(x,y)$
  - Called generative model since we can generate more samples artificially
- Given a full joint pdf we can
  - Marginalize $p(y)=\sum_x p(x,y)$
  - Condition $p(y\,|\,x)=\dfrac{p(x,y)}{p(x)}$
  - 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*

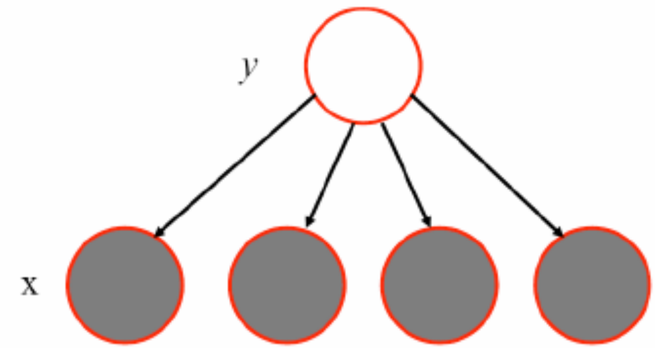
Naive Bayes Classifier

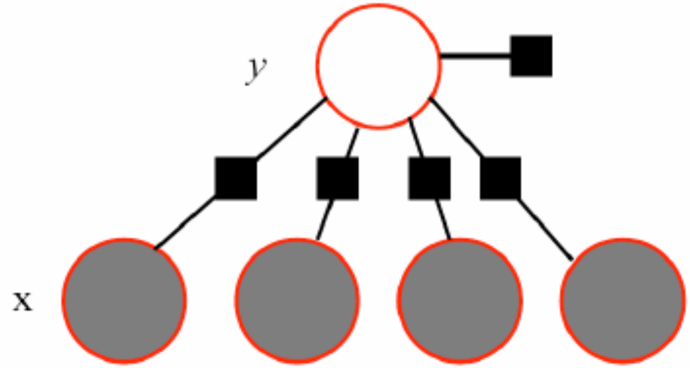
- Goal is to predict single class variable y given a vector of features $\mathrm{x}=(x_1,..,x_M)$
- Assume that once class labels are known the features are independent
- Joint probability model has the form

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

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

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



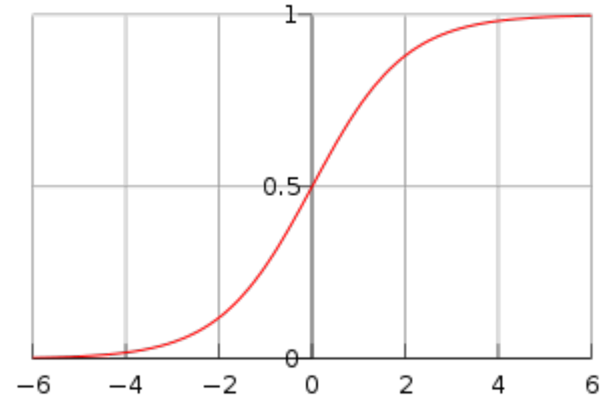


Discriminative Classifier: Logistic Regression

Binary logistic regression:

$$f(x,\mathbf{w}) = \frac{1}{1+e^{-\mathbf{w}^T\mathbf{x}}} \longleftarrow g(z) = \frac{1}{1+e^{-z}}$$

Logistic or sigmoid function



How to fit **w** for logistic regression model?

$$P(y=1\,|\,x;\mathbf{w}) = f(x,\mathbf{w})$$
$$P(y=0\,|\,x;\mathbf{w}) = 1 - f(x,\mathbf{w})$$

i.e.,

$$p(y\,|\,x;\mathbf{w}) = f(x,\mathbf{w})^y(1-f(x,\mathbf{w}))^{1-y}$$

Then we can obtain the log likelihood

$$\begin{aligned} L(\mathbf{w}) &= \log p(Y\,|\,X;\mathbf{w}) \\ &= \log\prod_{i=1}^{N} p(y_i\,|\,x_i;\mathbf{w}) \\ &= \log\prod_{i=1}^{N} f(x_i,\mathbf{w})^{y_i}(1-f(x_i,\mathbf{w}))^{1-y_i} \\ &= \sum_{i=1}^{N} y_i\log f(x_i,\mathbf{w}) + (1-y_i)\log(1-f(x_i,\mathbf{w})) \end{aligned}$$

Logistic Regression vs. Bayes Classifier

- Posterior probability of class variable *y is*

$$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x|y=1)p(y=1)+p(x|y=0)p(y=0)}$$

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

$$\text{where } \boxed{a = \ln\frac{p(x|y=1)p(y=1)}{p(x|y=0)p(y=0)}}$$

- 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

- For M-dimensional feature space logistic regression has M parameters $\mathbf{w}=(w_1,..,w_M)$
- 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 $p(y=1)$
  - Which can be reduced to $O(M)$ parameters by assuming independence via Naïve Bayes

Summary

- Generative and Discriminative methods are two basic approaches in machine learning
  - 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
- Generative models are more elegant, have explanatory power
- Discriminative models perform better in language related tasks

Thanks!

Jie Tang, DCST
http://keg.cs.tsinghua.edu.cn/jietang/
http://arnetminer.org
Email: jietang@tsinghua.edu.cn