



UNIVERSITY of PENNSYLVANIA

RUTGERS
UNIVERSITY

Generative versus discriminative modeling frameworks

Albert A. Montillo, Ph.D.

Rutgers University, Computer Science

Piscataway, NJ, USA

Temple University, February 15, 2010

Why study modeling frameworks?

1. To simplify research

- What do these methods have in common?
 - Bayesian Networks
 - Gaussian Mixture models
 - Naïve Bayes methods
- Or these?
 - Logistic regression
 - Maximum Conditional Likelihood
- Or these?
 - Neural Networks
 - Decision trees
 - Support Vector Machines

2. To place new methods you learn about into context.

3. To design new methods.

Generative modeling

Focus

- Describe a system with a probabilistic model
- Facilitate generating new system configurations
- Facilitate making inferences and predictions

Illustrative example: Gaussian Naïve Bayes Classifier

- Joint PDF → Bayes rule

$$\begin{aligned} P(X,Y) &= P(Y|X) P(X) = P(X|Y) P(Y) \\ &P(Y|X) = P(X|Y) P(Y) / P(X) \end{aligned}$$

Discriminative modeling (type I)

Focus

- Assume task is classification (or regression) not generation
- Directly model the posterior:

$$P(Y|X) = P(X|Y) P(Y) / P(X)$$

Illustrative example: Logistic regression

Discriminative modeling (type II)

Focus

- Assume task is classification (or regression)
- Find the decision boundary
- Directly model the mapping $Y=f(X)$

Illustrative example: Decision trees Classifier

Outline

1. Gaussian Naïve Bayes (generative)
2. Logistic Regression (discriminative I)
3. Decision Trees (discriminative II)
4. Categorization of more sophisticated methods
5. Framework advantages & disadvantages
6. Heuristics for using the methods in practice
7. Summary and conclusions

Generative example

- Goal: estimate $P(X|Y)$ and $P(Y)$
- Unreasonable amount of data required for $P(X|Y)$
- Assumption: X_1, \dots, X_n conditionally independent of one another given Y
- Given these assumptions we have:

$$P(X_1 \dots X_n | Y) = \prod_{i=1}^n P(X_i | Y)$$

- For discrete Y , Naïve Bayes classifier is

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) P(X_1 \dots X_n | Y = y_k)}{\sum_j P(Y = y_j) P(X_1 \dots X_n | Y = y_j)}$$

- Using our conditional independence assumption:

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

- Calculate posterior given X_{test} and $P(Y)$ and $P(X|Y)$ learned from training data]

$$P(Y = y_k | X_1 \dots X_n) = \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

- To classify, use the most probable value of Y :

$$Y \leftarrow \arg \max_{y_k} \frac{P(Y = y_k) \prod_i P(X_i | Y = y_k)}{\sum_j P(Y = y_j) \prod_i P(X_i | Y = y_j)}$$

- or simply:

$$Y \leftarrow \arg \max_{y_k} P(Y = y_k) \prod_i P(X_i | Y = y_k)$$

Gaussian Naïve Bayes Classifier

- For continuous X_i , choose continuous valued distribution
- Assumption: For every y_k , distribution of each X_i is Gaussian and defined by μ and σ specific to X_i and y_k
- Training resolves the parameters of the conditional distributions:

$$\begin{aligned}\mu_{ik} &= E[X_i | Y = y_k] & \pi_k &= P(Y = y_k) \\ \sigma_{ik}^2 &= E[(X_i - \mu_{ik})^2 | Y = y_k]\end{aligned}$$

- Maximum likelihood estimation:

$$\begin{aligned}\hat{\mu}_{ik} &= \frac{1}{\sum_j \delta(Y^j = y_k)} \sum_j X_i^j \delta(Y^j = y_k) \\ \hat{\sigma}_{ik}^2 &= \frac{1}{(\sum_j \delta(Y^j = y_k)) - 1} \sum_j (X_i^j - \hat{\mu}_{ik})^2 \delta(Y^j = y_k)\end{aligned}$$

Discriminative I example

Logistic Regression: 2-category classification

- Goal: estimate posterior $P(Y|X)$ directly
- If Y is boolean (0/1), then Logistic regression assumes parametric form for $P(Y|X)$:

$$P(Y = 1|X) = \frac{1}{1 + \exp(w_0 + \sum_{i=1}^n w_i X_i)}$$

$$P(Y = 0|X) = \frac{\exp(w_0 + \sum_{i=1}^n w_i X_i)}{1 + \exp(w_0 + \sum_{i=1}^n w_i X_i)}$$

- To classify X , assign the Y that maximizes $P(Y|X)$. E.g. assign $Y=0$ if:

$$1 < \frac{P(Y = 0|X)}{P(Y = 1|X)}$$

- This expression can be re-written and we take the log:

$$1 < \exp(w_0 + \sum_{i=1}^n w_i X_i)$$

$$0 < w_0 + \sum_{i=1}^n w_i X_i$$

- Choose vector of parameters, $W = \langle w_0, \dots, w_n \rangle$ that maximizes conditional data likelihood:

$$W \leftarrow \arg \max_W \prod_l P(Y^l | X^l, W)$$

- After taking logs we have:

$$W \leftarrow \arg \max_W \sum_l \ln P(Y^l | X^l, W)$$

- Since Y is binary, express conditional data likelihood objective function as:

$$\begin{aligned} l(W) &= \sum_l Y^l \ln P(Y^l = 1 | X^l, W) + (1 - Y^l) \ln P(Y^l = 0 | X^l, W) \\ &= \sum_l Y^l \ln \frac{P(Y^l = 1 | X^l, W)}{P(Y^l = 0 | X^l, W)} + \ln P(Y^l = 0 | X^l, W) \\ &= \sum_l Y^l (w_0 + \sum_i^n w_i X_i^l) - \ln(1 + \exp(w_0 + \sum_i^n w_i X_i^l)) \end{aligned}$$

- No closed form solution to maximize conditional log likelihood wrt W
- Choose an optimization strategy, e.g. gradient ascent.
- Gradient can be expressed using:

$$\frac{\partial l(W)}{\partial w_i} = \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1 | X^l, W))$$

- To optimize, initial with zero weight, then iteratively update by walking in gradient direction:

$$w_i \leftarrow w_i + \eta \sum_l X_i^l (Y^l - \hat{P}(Y^l = 1 | X^l, W))$$

Logistic Regression: **multi**-category classification

- If Y takes on finite number of discrete values $\{y_1, \dots, y_K\}$ then posterior is modeled similarly as 2 category:

$$P(Y = y_k | X) = \frac{\exp(w_{k0} + \sum_{i=1}^n w_{ki} X_i)}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^n w_{ji} X_i)} \quad \text{for } Y=y_1, Y=y_2, \dots, Y=y_{K-1}$$
$$P(Y = y_K | X) = \frac{1}{1 + \sum_{j=1}^{K-1} \exp(w_{j0} + \sum_{i=1}^n w_{ji} X_i)} \quad \text{for } Y=y_K$$

- The gradient ascent weight update rule becomes:

$$w_{ji} \leftarrow w_{ji} + \eta \sum_l X_i^l (\delta(Y^l = y_j) - \hat{P}(Y^l = y_j | X^l, W))$$

Gaussian Naïve Bayes versus Logistic Regression (1/2)

- Ostensibly the difference is:
 - Logistic regression directly estimates $P(Y|X)$
 - Gaussian Naïve Bayes estimates $P(Y)$ and $P(X|Y)$
- But, ... the **essence of the difference** is:
 - Generative approach models the input, which reduces variance of parameter estimation, at the expense of possibly introducing model bias. [Liang & Jordan '08, Mitchell '10]
 - If the bias is appropriate, generative will be preferred, if not discriminative will be.

Gaussian Naïve Bayes versus Logistic Regression (2/2)

- Appropriate assumptions
 - LR & GNB yield same classification as # examples increases
 - Finite # examples, GNB reaches asymptotic accuracy sooner
 - Experimental comparison: [Ng & Jordan 02]
 - GNB converges by just $\log n$ examples
 - LR requires n examples

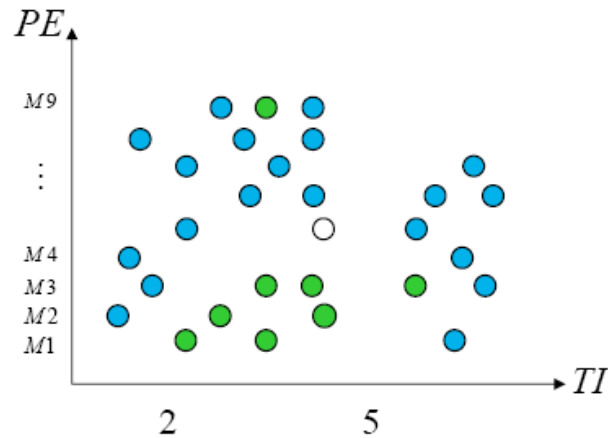
- Inappropriate assumptions:
 - LR's conditional likelihood maximization **adapts & selects** parameters that maximize fit to data
 - As # examples increases, LR reach or surpass GNB accuracy
 - Experimental comparison: [Ng & Jordan 02]
 - GNB outperforms LR when training data is scarce
 - LR outperforms GNB with many training examples

Discriminative II example

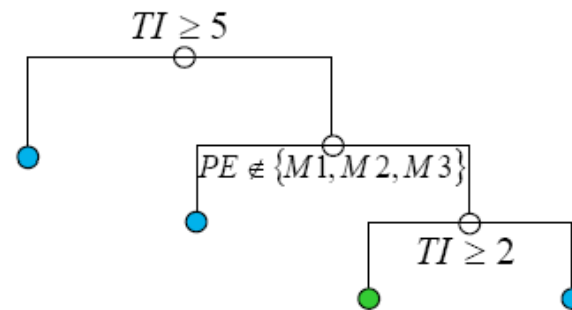
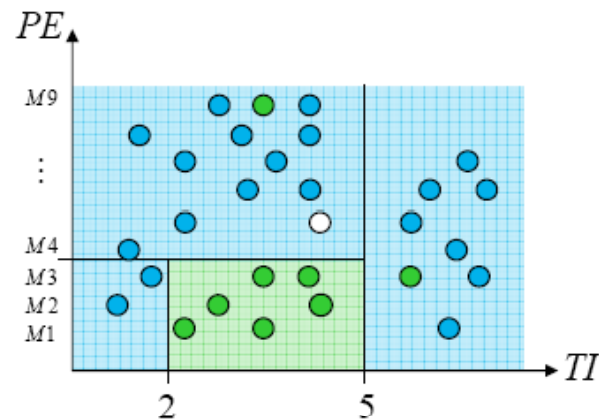
Decision trees involve greedy, recursive partitioning

Simple dataset with two predictors:

TI	PE	Response
1.0	$M2$	good
2.0	$M1$	bad
...
4.5	$M5$?



Greedy, recursive partitioning along TI and PE:



Decision trees: model, score criterion, search strategy

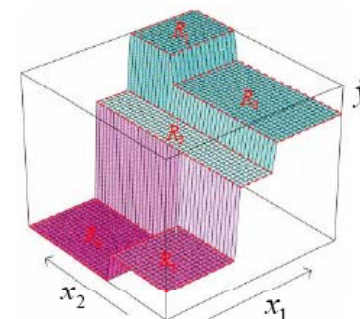
Training set: $D = \{y_i, x_{i1}, x_{i2}, \dots, x_{ik}\}_1^N = \{y_i, \mathbf{x}_i\}$
 where $k = \# \text{predictors}$ and $N = \# \text{samples}$

Model of underlying functional form sought from data

general \rightarrow decision trees

$$\hat{F}(\mathbf{x}) = \hat{F}(\mathbf{x}; \mathbf{a}) \in \mathcal{F}$$

$$\hat{y} = T(\mathbf{x}) = \sum_{m=1}^M \hat{c}_m I(\mathbf{x} \in \hat{R}_m)$$



Score criterion to judge quality of fit of model

$$\hat{R}(\mathbf{a}) = \frac{1}{N} \sum_{i=1}^N L(y_i, \hat{F}(\mathbf{x}_i; \mathbf{a}))$$

Squared error: $L(y, \hat{y}) = (y - \hat{y})^2$

Search strategy to minimize the score criterion

$$\hat{\mathbf{a}} = \arg \min_{\mathbf{a}} \hat{R}(\mathbf{a})$$

$$\{\hat{c}_m, \hat{R}_m\}_1^M = \arg \min_{\{c_m, R_m\}_1^M} \sum_{i=1}^N \left[y_i - \sum_{m=1}^M c_m I(\mathbf{x}_i \in R_m) \right]^2$$

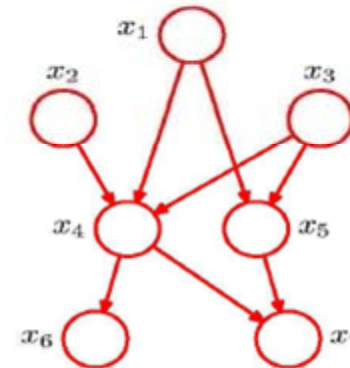
$\hat{y}_i = T(\mathbf{x}_i)$

Categorizing more sophisticated methods

Generative modeling: Other examples

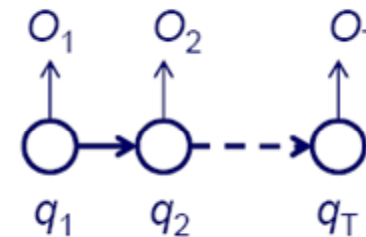
- Bayesian network (directed graph)

$$p(\mathbf{x}) = \prod_{i=1}^n p(x_i | pa_i)$$



- Hidden Markov model (HMM)

$$p(O | \lambda) = \sum_Q P(O | Q, \lambda) P(Q | \lambda)$$
$$= \sum_{q_1, q_2, \dots, q_T} \pi_{q_1} b_{q_1}(O_1) a_{q_1 q_2} b_{q_2}(O_2) \cdots a_{q_{T-1} q_T} b_{q_T}(O_T)$$



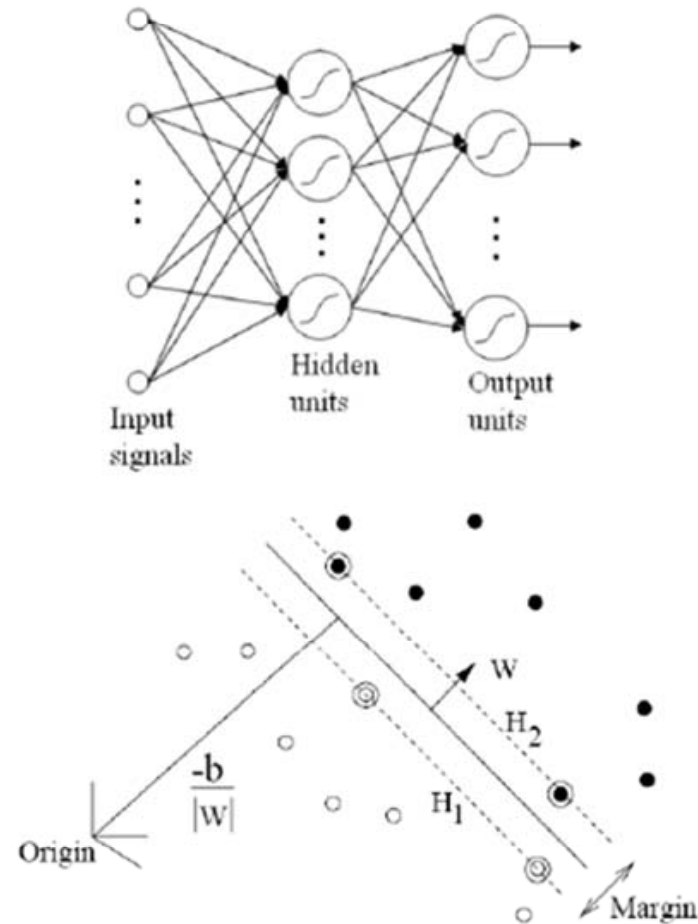
- Undirected graphs

- Markov random field (MRF)

Discriminative modeling: Other examples

- Artificial neural networks (ANN):
discriminant function regardless of probability distribution
- Support vector machine (SVM):
hyperplane classifier (2-class)
 - **Decision boundary**

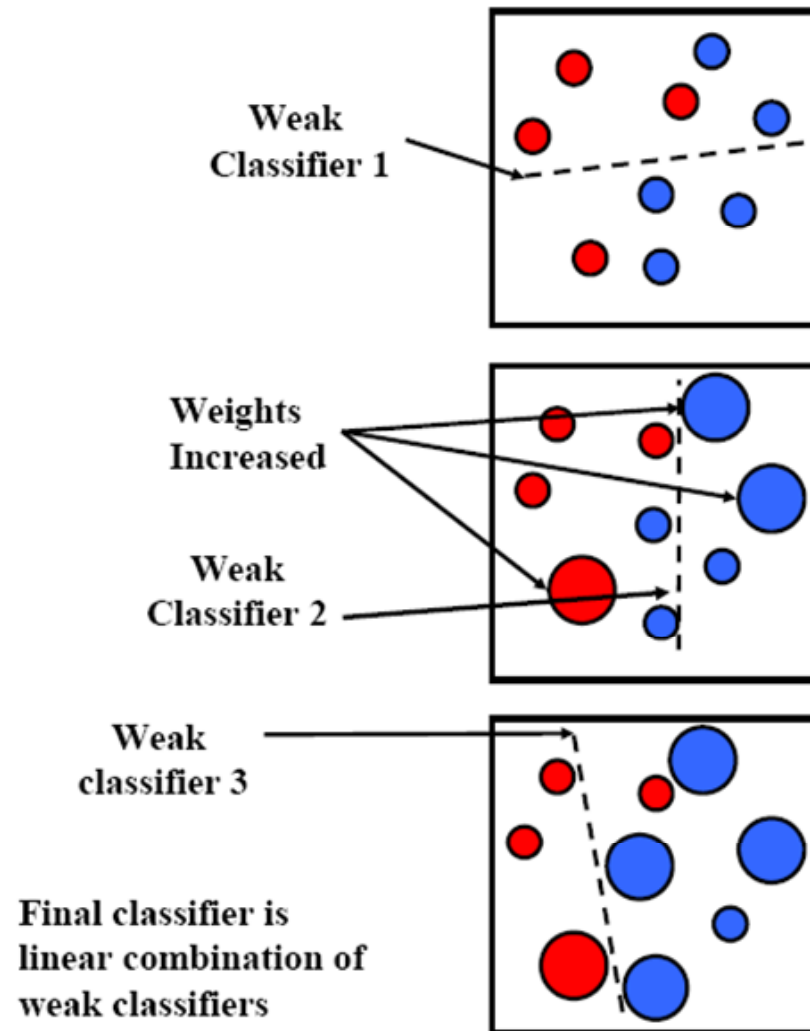
$$\mathbf{w} \cdot \mathbf{x} + b \gg 0$$



Credit Fei-Fei Li 2007

Discriminative modeling: Other examples

(Ada)Boost



Credit: P. Viola and C. Bishop, ICCV 2003 Tutorial

Generative modeling: advantages/disadvantages

- Advantages

- Ability to introduce prior knowledge
- Do not require large training sets
- Generation of synthetic inputs

- Disadvantages

- marred by generic optimization criteria
- Potentially wasteful modeling
- Reliant on domain expertise
- Don't scale well to large number of classes

Discriminative modeling: advantages/disadvantages

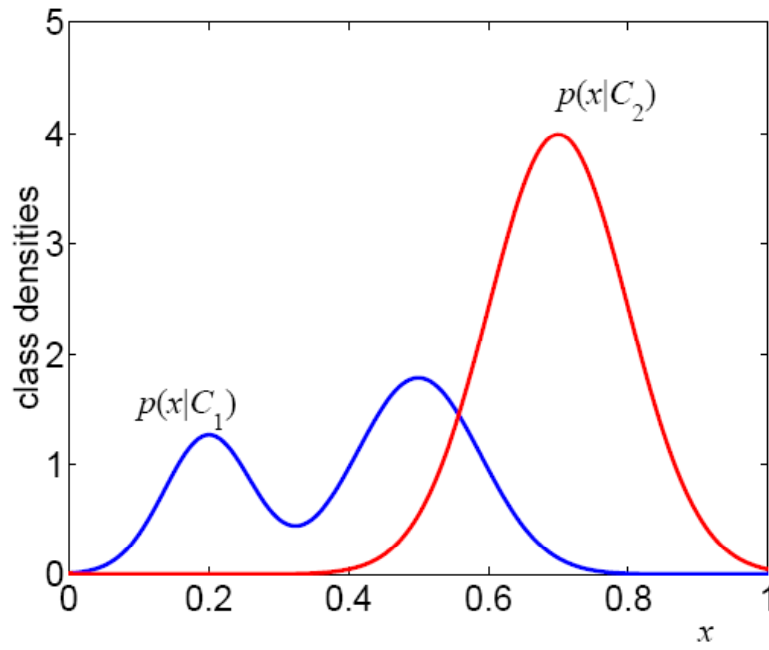
- Advantages
 - Fast prediction speed
 - Potentially more accurate prediction
- Disadvantages
 - Task specific
 - Long training time
 - Don't easily handle compositionality

When to use each type of model?

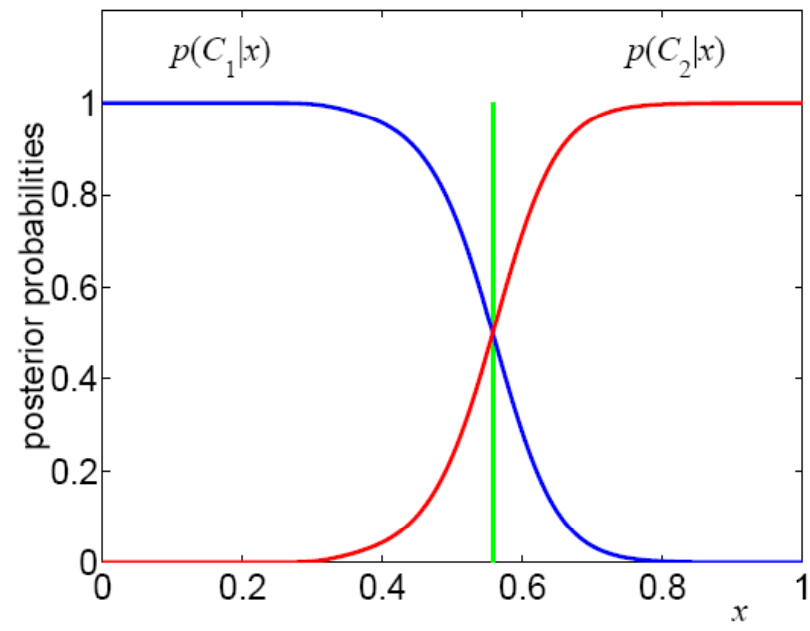
- Build a generative model if
 - Have expressible domain knowledge
 - Have small training dataset
 - Require instance generation
 - Interpretation of classifier output
- Build a discriminative model if
 - Little domain knowledge
 - Require fast (real-time) prediction
 - Large training dataset
- Trade off : training time vs prediction time
 - Generative: train fast, predict slow
 - Discriminative: train slow, predict fast

Summary and conclusions

Generative



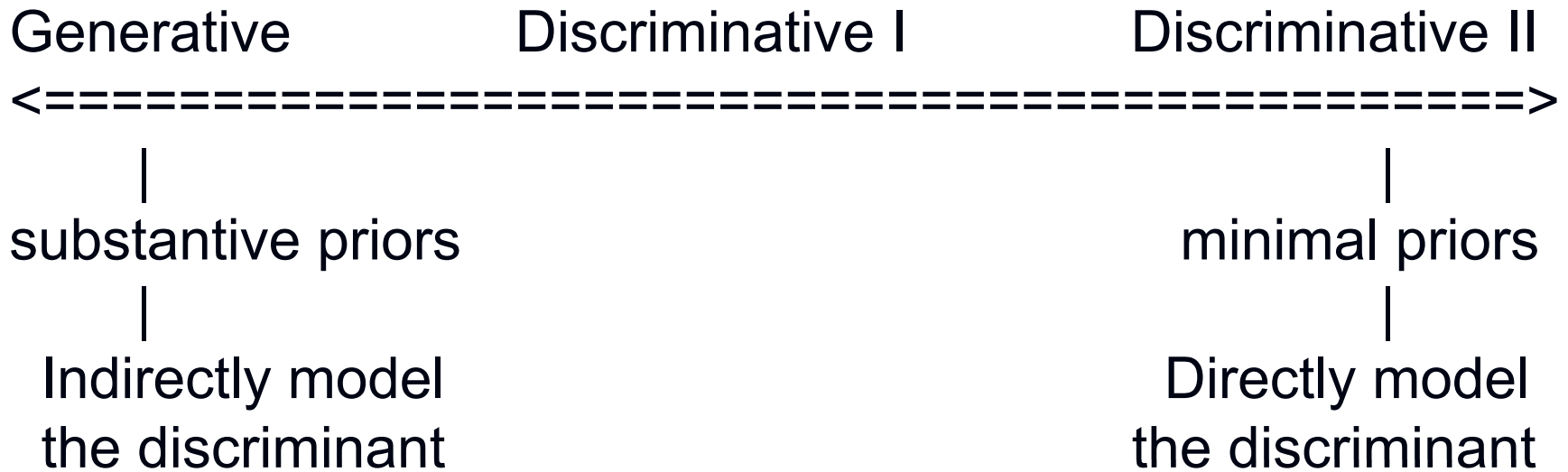
Discriminative



Credit: C. Bishop, ICPR 2004

Summary and conclusions

Coarse continuum between frameworks presented today:



Thank you

... Questions?

References

- T. Jebara, Machine Learning: Discriminative and Generative, Kluwer, 2004
ISBN: 1-4020-7647-9. Boston, MA, 2004.
- A. Y. Ng and M. I. Jordan, On Discriminative vs. Generative Classifiers: A comparison of logistic regression and Naive Bayes. In Advances in Neural Information Processing Systems 14, 2002
- T. Mitchell, Ch 1: Generative and Discriminative classifiers: Naïve Bayes and Logistic Regression, 2010
- P. Liang, M. Jordon, An Asymptotic Analysis of Generative, Discriminative, and Pseudolikelihood Estimators, ICML 2005
- C. Bishop, Pattern Recognition and Machine Learning, Springer, 2006
- C. Bishop, Generative vcs Discriminative Methods in Computer Vision, keynote at ICPR, 2004
- http://en.wikipedia.org/wiki/Logistic_regression
- Fei-Fei Li, Machine Learning in Computer Vision, Princeton Guest Lecture, 2007.
<http://www.cs.princeton.edu/courses/archive/spr07/cos424/lectures/li-guest-lecture.pdf>