UNIVERSITÀ PARTHENOPE
DiST
DIPARTIMENTO DI SCIENZE E TECNOLOGIE

Machine Learning (part II)

Graphical Models

Angelo Ciaramella

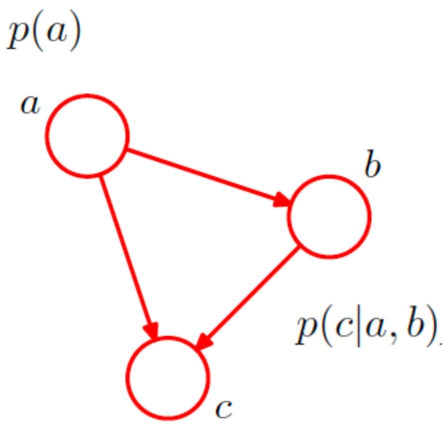
Introduction

- Probabilistic graphical models
  - simple way to visualize the structure of a probabilistic model
  - properties of the model, including conditional independence properties, can be obtained by inspection of the graph
  - Complex computations in terms of graphical manipulations
  - Main approaches
    - Bayesian networks
    - Markov random fields
    - Factor graph

Bayesian Networks

- Joint distribution

$$p(a, b, c) = p(c|a, b)p(a, b)$$



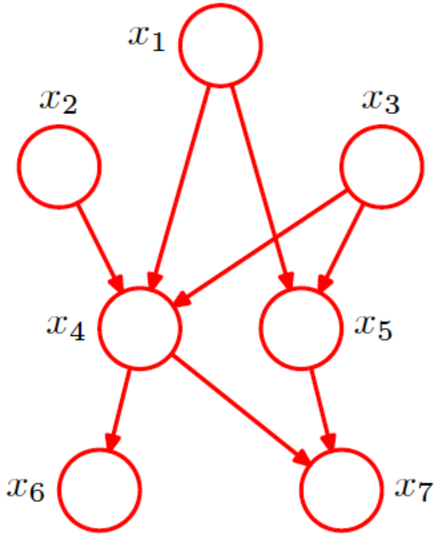
$$p(a, b, c) = p(c|a, b)p(b|a)p(a)$$

$$p(x_1, \ldots, x_K) = p(x_K|x_1, \ldots, x_{K-1}) \ldots p(x_2|x_1)p(x_1)$$ k-variables

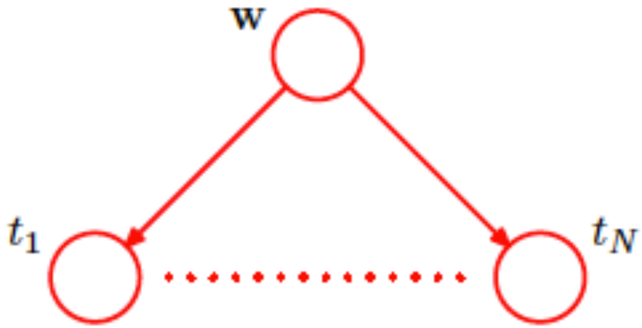
Bayesian Networks

- Joint distribution

$$p(x_1)p(x_2)p(x_3)p(x_4|x_1,x_2,x_3)p(x_5|x_1,x_3)p(x_6|x_4)p(x_7|x_4,x_5)$$



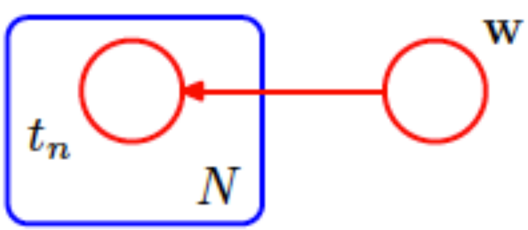
Polynomial regression



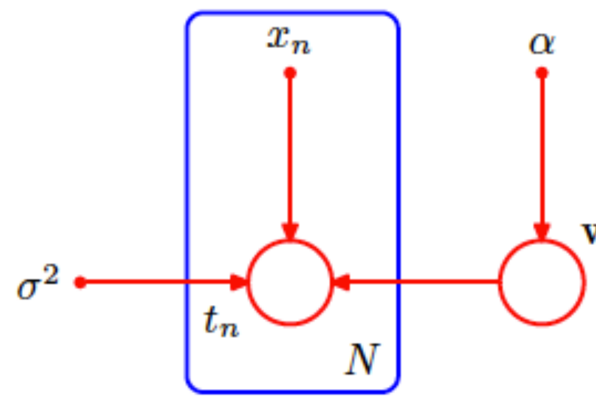
$$p(\mathbf{t}, \mathbf{w}) = p(\mathbf{w}) \prod_{n=1}^{N} p(t_n|\mathbf{w})$$

Parametric model

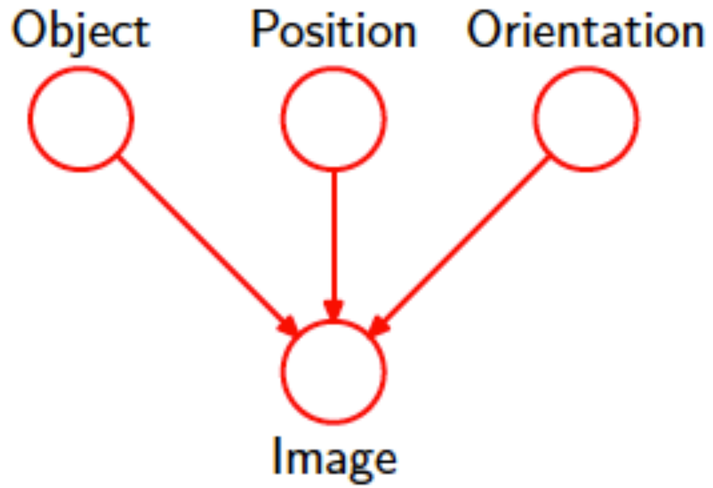
$$p(\mathbf{t}, \mathbf{w}|\mathbf{x}, \alpha, \sigma^2) = p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{w}, x_n, \sigma^2)$$



Compact representation

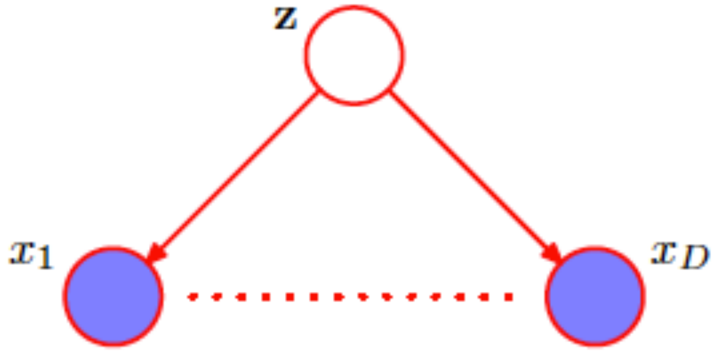


Generative models



A graphical model representing the process (causal process) by which images of objects are created. The image (a vector of pixel intensities) has a probability distribution that is dependent on the identity of the object as well as on its position and orientation.

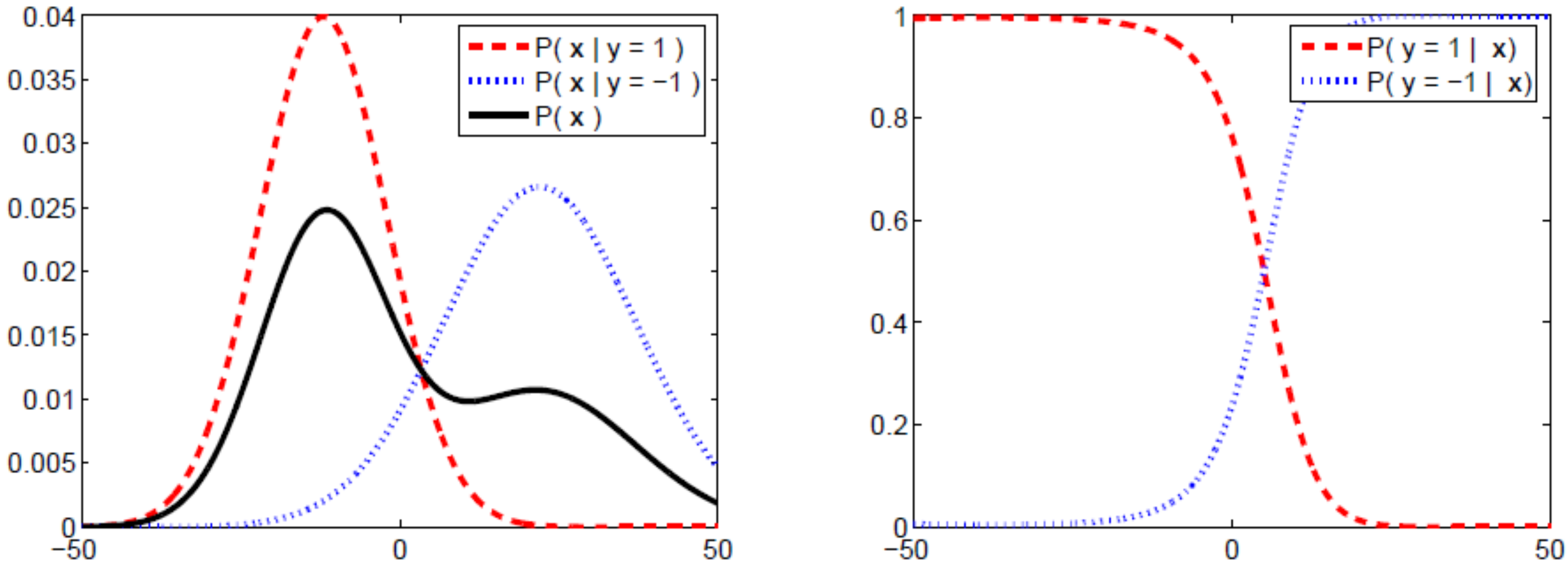
Naïve Bayes



Conditioned on the class label z, the components of the observed vector x = $(x_1, \ldots, x_D)^T$ are assumed to be independent

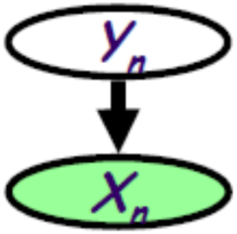
Bayes formula

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

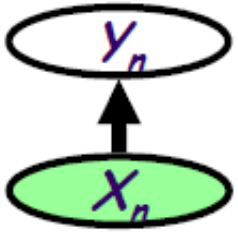


Generative vs Discriminative Models

- Generative models
  - Assume some functional form for P(X|Y), P(Y)
  - Estimate parameters of P(X|Y), P(Y) directly from training data
  - Use Bayes rule to calculate P(Y|X= x)



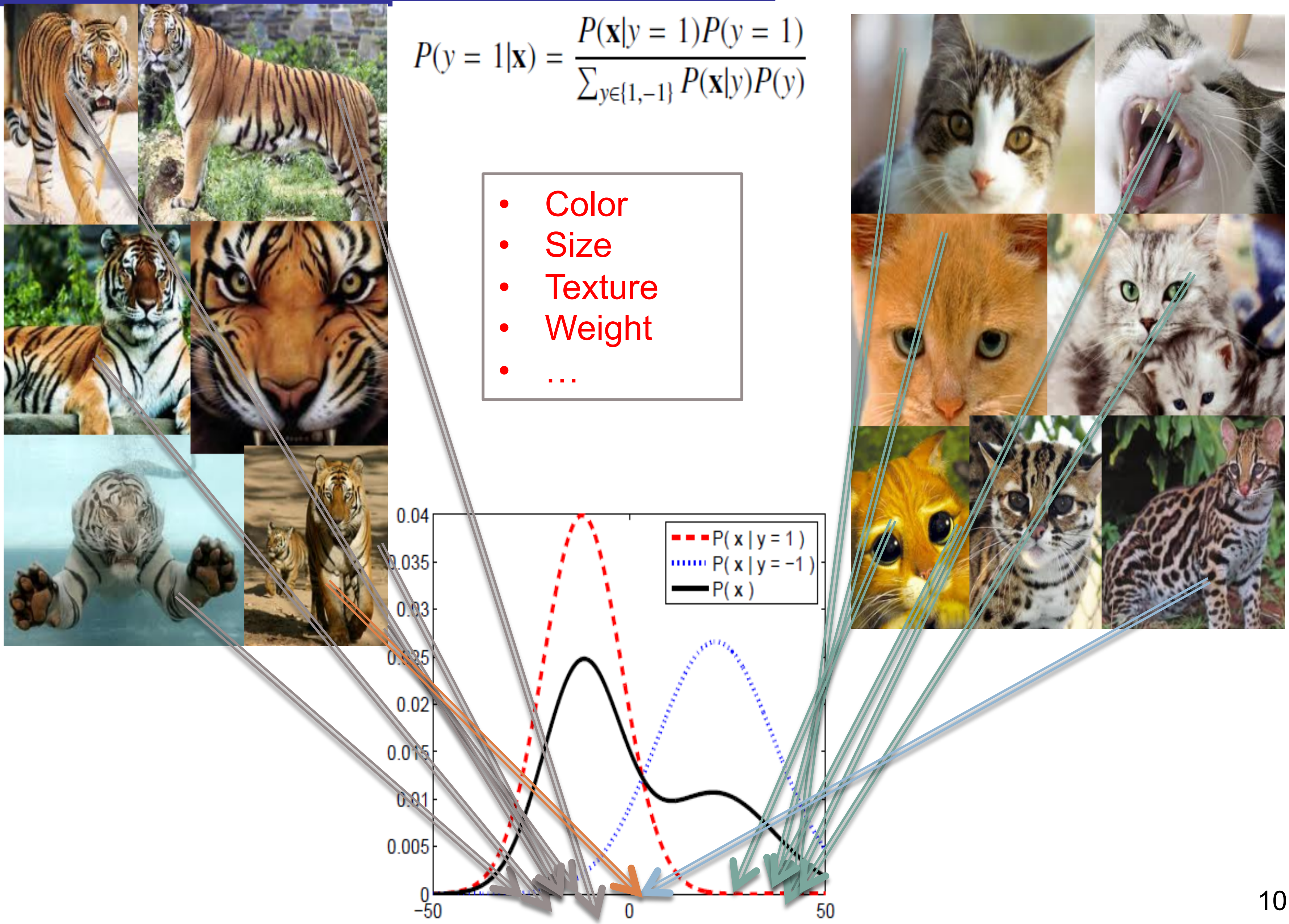
- Discriminative models
  - Directly assume some functional form for P(Y|X)
  - Estimate parameters of P(Y|X) directly from training data



Generative Model

$$P(y=1|\mathbf{x}) = \frac{P(\mathbf{x}|y=1)P(y=1)}{\sum_{y\in\{1,-1\}} P(\mathbf{x}|y)P(y)}$$

- Color
- Size
- Texture
- Weight
- …

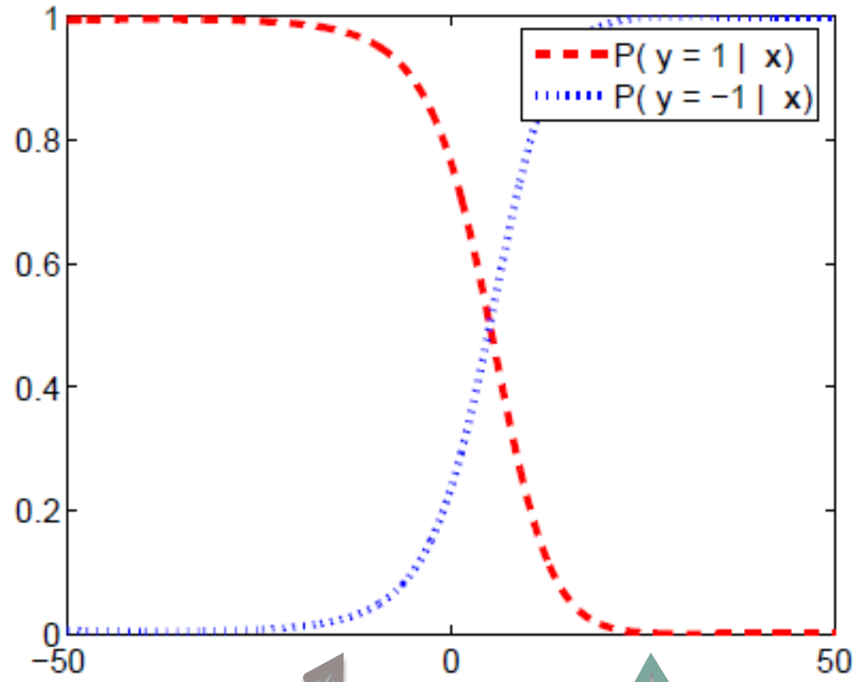


Discriminative Model

- Logistic regression

$$P(y = 1|\mathbf{x}) = \frac{1}{1 + \exp(yf(\mathbf{x}))}$$

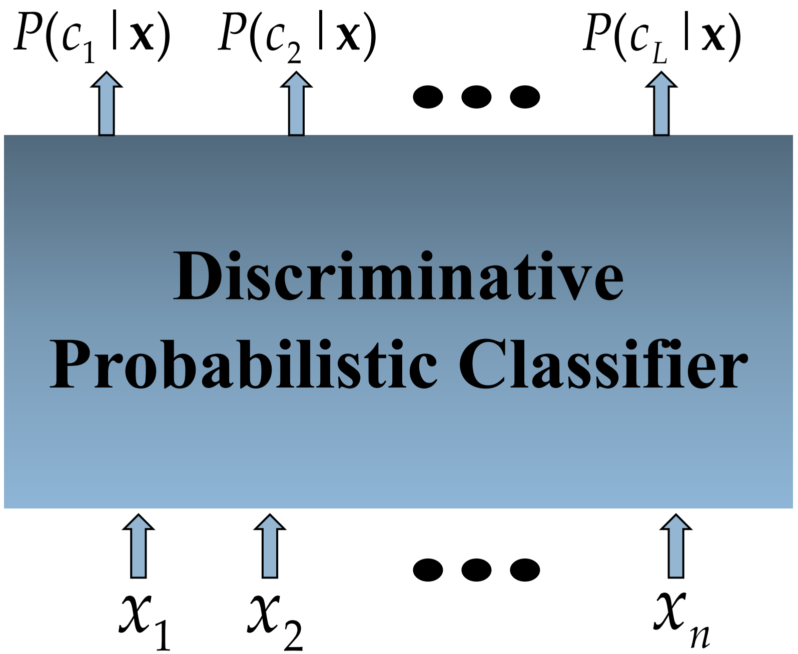
$$f^*(\mathbf{x}) = \begin{cases} +\infty & \Pr(y = 1|\mathbf{x}) > \frac{1}{2}, \\ -\infty & \Pr(y = -1|\mathbf{x}) < \frac{1}{2}, \\ \text{arbitrary} & \text{otherwise.} \end{cases}$$



- Color
- Size
- Texture
- Weight
- …

Discriminative model

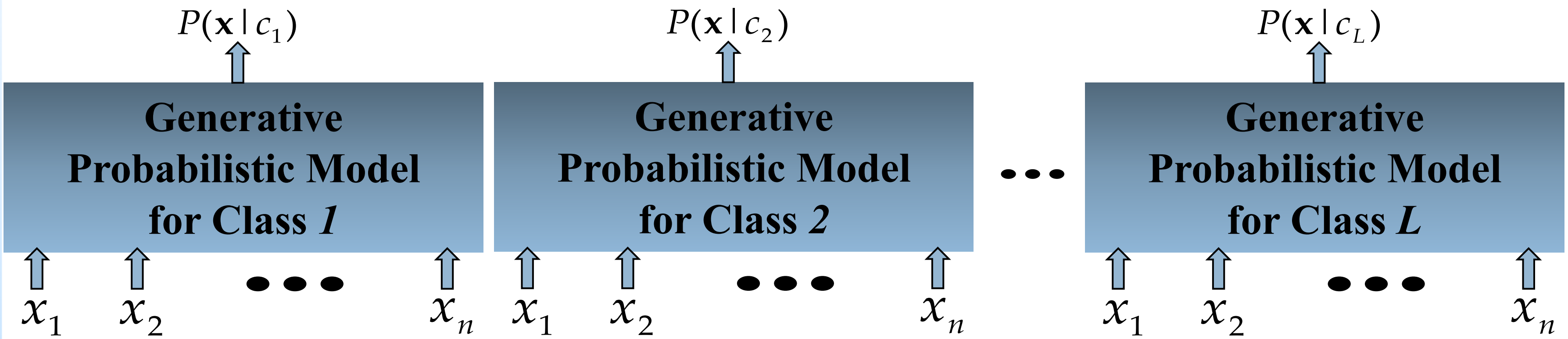
$$P(C \mid \mathbf{X}) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$



$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

Discriminative model

$$P(\mathbf{X} \mid C) \quad C = c_1, \cdots, c_L, \mathbf{X} = (X_1, \cdots, X_n)$$



$$\mathbf{x} = (x_1, x_2, \cdots, x_n)$$

Bayes classifier

- Bayes rule

Likelihood — Prior

$$P(Y|X_1,\ldots,X_n) = \frac{P(X_1,\ldots,X_n|Y)P(Y)}{P(X_1,\ldots,X_n)}$$

Normalization Constant

MAP classification rule

- Maximum A Posterior rule

$$P(C=c^{*}\mid \mathbf{X}=\mathbf{x}) > P(C=c\mid \mathbf{X}=\mathbf{x}) \quad c \neq c^{*},\ c=c_1,\cdots,c_L$$

- Generative classification

$$\begin{aligned} P(C=c_i \mid \mathbf{X}=\mathbf{x}) &= \frac{P(\mathbf{X}=\mathbf{x}\mid C=c_i)P(C=c_i)}{P(\mathbf{X}=\mathbf{x})} \\ &\propto P(\mathbf{X}=\mathbf{x}\mid C=c_i)P(C=c_i) \\ &\text{for } i=1,2,\cdots,L \end{aligned}$$

MAP classification rule

- Bayes classification

$$P(C \mid \mathbf{X}) \propto P(\mathbf{X} \mid C)P(C) = P(X_1, \cdots, X_n \mid C)P(C)$$

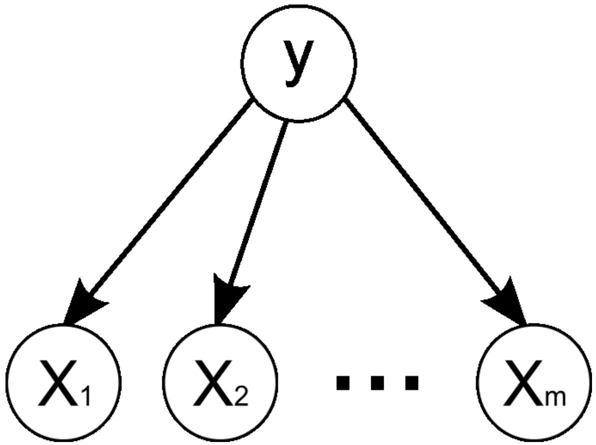
Difficulty for learning the joint probability

- Naïve Bayes

$$\begin{aligned} P(X_1, X_2, \cdots, X_n \mid C) &= P(X_1 \mid X_2, \cdots, X_n; C)P(X_2, \cdots, X_n \mid C) \\ &= P(X_1 \mid C)P(X_2, \cdots, X_n \mid C) \\ &= P(X_1 \mid C)P(X_2 \mid C) \cdots P(X_n \mid C) \end{aligned}$$

all input attributes are conditionally independent

Naïve Bayes



$$(\forall i, j) P(Y = y_i | X = x_j) = \frac{P(X = x_j | Y = y_i) P(Y = y_i)}{\sum_k P(X = x_j | Y = y_k) P(Y = y_k)}$$

Naïve Bayes

- Learning phase (given a training set S)

For each target value of c_i $(c_i = c_1, \cdots, c_L)$

$\hat{P}(C = c_i) \leftarrow$ estimate $P(C = c_i)$ with examples in **S**;

For every attribute value x_{jk} of each attribute X_j $(j = 1, \cdots, n; k = 1, \cdots, N_j)$

$\hat{P}(X_j = x_{jk} \mid C = c_i) \leftarrow$ estimate $P(X_j = x_{jk} \mid C = c_i)$ with examples in **S**;

- Test phase

$\mathbf{X}' = (a'_1, \cdots, a'_n)$ unknown istance

$$[\hat{P}(a'_1 \mid c^*) \cdots \hat{P}(a'_n \mid c^*)]\hat{P}(c^*) > [\hat{P}(a'_1 \mid c) \cdots \hat{P}(a'_n \mid c)]\hat{P}(c), \quad c \neq c^*, c = c_1, \cdots, c_L$$

Example

PlayTennis: training examples

Day	Outlook	Temperature	Humidity	Wind	PlayTennis
D1	Sunny	Hot	High	Weak	No
D2	Sunny	Hot	High	Strong	No
D3	Overcast	Hot	High	Weak	Yes
D4	Rain	Mild	High	Weak	Yes
D5	Rain	Cool	Normal	Weak	Yes
D6	Rain	Cool	Normal	Strong	No
D7	Overcast	Cool	Normal	Strong	Yes
D8	Sunny	Mild	High	Weak	No
D9	Sunny	Cool	Normal	Weak	Yes
D10	Rain	Mild	Normal	Weak	Yes
D11	Sunny	Mild	Normal	Strong	Yes
D12	Overcast	Mild	High	Strong	Yes
D13	Overcast	Hot	Normal	Weak	Yes
D14	Rain	Mild	High	Strong	No

$$\hat{P}(X_j = a_{jk} \mid C = c_i) = \frac{n_c + mp}{n + m}$$

n_c : number of training examples for which $X_j = a_{jk}$ and $C = c_i$

n : number of training examples for which $C = c_i$

p : prior estimate (usually, $p = 1/t$ for t possible values of X_j)

m : weight to prior (number of "virtual" examples, $m \geq 1$)

Learning phase

Outlook	Play=*Yes*	Play=*No*
Sunny	2/9	3/5
Overcast	4/9	0/5
Rain	3/9	2/5

Temperature	Play=*Yes*	Play=*No*
Hot	2/9	2/5
Mild	4/9	2/5
Cool	3/9	1/5

Humidity	Play=*Yes*	Play=N*o*
High	3/9	4/5
Normal	6/9	1/5

Wind	Play=*Yes*	Play=*No*
Strong	3/9	3/5
Weak	6/9	2/5

$P(\text{Play}=Yes) = 9/14$ $P(\text{Play}=No) = 5/14$

Test phase

- ## New istance

  x′=(Outlook=*Sunny*, Temperature=*Cool*, Humidity=*High*, Wind=*Strong*)

- ## Look up table

  P(Outlook=*Sunny*|Play=*Yes*) = 2/9
  P(Temperature=*Cool*|Play=*Yes*) = 3/9
  P(Huminity=*High*|Play=*Yes*) = 3/9
  P(Wind=*Strong*|Play=*Yes*) = 3/9
  P(Play=*Yes*) = 9/14

  P(Outlook=*Sunny*|Play=*No*) = 3/5
  P(Temperature=*Cool*|Play==*No*) = 1/5
  P(Huminity=*High*|Play=*No*) = 4/5
  P(Wind=*Strong*|Play=*No*) = 3/5
  P(Play=*No*) = 5/14

- ## MAP rule

  P(*Yes*|**x**′): [P(*Sunny*|*Yes*)P(*Cool*|*Yes*)P(*High*|*Yes*)P(*Strong*|*Yes*)]P(Play=*Yes*) = 0.0053

  P(*No*|**x**′): [P(*Sunny*|No) P(*Cool*|No)P(*High*|*No*)P(*Strong*|*No*)]P(Play=*No*) = 0.0206

Given the fact P(*Yes*|**x**′) < P(*No*|**x**′), we label **x**′ to be "*No*".

ML – Graphical Models

Continuous inputs

- Normal distribution

$$\hat{P}(X_j \mid C = c_i) = \frac{1}{\sqrt{2\pi}\sigma_{ji}} \exp\left(-\frac{(X_j - \mu_{ji})^2}{2\sigma_{ji}^2} \right)$$

μ_{ji} : mean (avearage) of attribute values X_j of examples for which $C = c_i$

σ_{ji} : standard deviation of attribute values X_j of examples for which $C = c_i$