



# Discriminative Model vs Generative Model

## Motivation





# Conditional Distribution

## Motivation

- Suppose the joint distribution is given.

$$\mathbb{P} \{X_j = x_j, X_{j'} = x_{j'}\}$$

- The conditional distribution of  $X_j$  given  $X_{j'} = x_{j'}$  is ratio between the joint distribution and the marginal distribution.

$$\mathbb{P} \{X_j = x_j | X_{j'} = x_{j'}\} = \frac{\mathbb{P} \{X_j = x_j, X_{j'} = x_{j'}\}}{\mathbb{P} \{X_{j'} = x_{j'}\}}$$





# Bayes Rule Example 2

## Quiz





# Conditional Independence

## Definition

- Recall two events  $A, B$  are independent if:

$$\mathbb{P}\{A, B\} = \mathbb{P}\{A\} \mathbb{P}\{B\} \text{ or } \mathbb{P}\{A|B\} = \mathbb{P}\{A\}$$

- In general, two events  $A, B$  are conditionally independent, conditional on event  $C$  if:

$$\mathbb{P}\{A, B|C\} = \mathbb{P}\{A|C\} \mathbb{P}\{B|C\} \text{ or } \mathbb{P}\{A|B, C\} = \mathbb{P}\{A|C\}$$

# Causal Chain

## Definition

- For three events  $A, B, C$ , the configuration  $A \rightarrow B \rightarrow C$  is called causal chain.
- In this configuration,  $A$  is not independent of  $C$ , but  $A$  is conditionally independent of  $C$  given information about  $B$ .
- Once  $B$  is observed,  $A$  and  $C$  are independent.



# Common Effect

## Definition

- For three events  $A, B, C$ , the configuration  $A \rightarrow B \leftarrow C$  is called common effect.
- In this configuration,  $A$  is independent of  $C$ , but  $A$  is not conditionally independent of  $C$  given information about  $B$ .
- Once  $B$  is observed,  $A$  and  $C$  are not independent.

# Training Bayes Net

## Definition

- Training a Bayesian network given the DAG is estimating the conditional probabilities. Let  $P(X_j)$  denote the parents of the vertex  $X_j$ , and  $p(X_j)$  be realizations (possible values) of  $P(X_j)$ .

$$\mathbb{P}\{x_j | p(X_j)\}, p(X_j) \in P(X_j)$$

- It can be done by maximum likelihood estimation given a training set.

$$\hat{\mathbb{P}}\{x_j | p(X_j)\} = \frac{c_{x_j, p(X_j)}}{c_{p(X_j)}}$$

# Bayesian Network Diagram

## Quiz

# Bayesian Network Diagram CPT Count

## Quiz



# Bayes Net Training Example, Training Quiz

# Bayes Net Training Example, Training 1

## Quiz

# Bayes Net Training Example, Training 2

## Quiz

# Bayes Net Training Example, Training 3

## Quiz

# Bayes Net Training Example, Training 4

## Quiz

# Bayes Net Training Example, Training 5

## Quiz

# Bayes Net Training Example, Training 5

## Quiz

# Laplace Smoothing

## Definition

- Recall that the MLE estimation can incorporate Laplace smoothing.

$$\hat{\mathbb{P}}\{x_j | p(X_j)\} = \frac{c_{x_j, p(X_j)} + 1}{c_p(X_j) + |X_j|}$$

- Here,  $|X_j|$  is the number of possible values (number of categories) of  $X_j$ .
- Laplace smoothing is considered regularization for Bayesian networks because it avoids overfitting the training data.



# Bayes Net Inference 1

## Definition

- Given the conditional probability table, the joint probabilities can be calculated using conditional independence.

$$\begin{aligned}\mathbb{P}\{x_1, x_2, \dots, x_m\} &= \prod_{j=1}^m \mathbb{P}\{x_j | x_1, x_2, \dots, x_{j-1}, x_{j+1}, \dots, x_m\} \\ &= \prod_{j=1}^m \mathbb{P}\{x_j | p(x_j)\}\end{aligned}$$

## Bayes Net Inference 2

### Definition

- Given the joint probabilities, all other marginal and conditional probabilities can be calculated using their definitions.

$$\mathbb{P}\{x_j | x_{j'}, x_{j''}, \dots\} = \frac{\mathbb{P}\{x_j, x_{j'}, x_{j''}, \dots\}}{\mathbb{P}\{x_{j'}, x_{j''}, \dots\}}$$

$$\mathbb{P}\{x_j, x_{j'}, x_{j''}, \dots\} = \sum_{x_k: k \neq j, j', j'', \dots} \mathbb{P}\{x_1, x_2, \dots, x_m\}$$

$$\mathbb{P}\{x_{j'}, x_{j''}, \dots\} = \sum_{x_k: k \neq j', j'', \dots} \mathbb{P}\{x_1, x_2, \dots, x_m\}$$

# Bayes Net Inference Example 1

## Quiz

# Bayes Net Inference Example 1 Computation 1

## Quiz

# Bayes Net Inference Example 1 Computation 2

## Quiz

# Bayes Net Inference Example 2

## Quiz

# Bayes Net Inference Example 2 Derivation

## Quiz

# Bayes Net Inference Example 3

## Quiz



# Bayes Net Inference Example 3 Derivation

## Quiz

# Bayes Net Inference Example 4

## Quiz

# Bayes Net Inference Example 4 Derivation

## Quiz

# Network Structure

## Discussion

- Selecting from all possible structures (DAGs) is too difficult.
- Usually, a Bayesian network is learned with a tree structure.
- Choose the tree that maximizes the likelihood of the training data.

# Chow Liu Algorithm

## Discussion

- Add an edge between features  $X_j$  and  $X_{j'}$  with edge weight equal to the information gain of  $X_j$  given  $X_{j'}$  for all pairs  $j, j'$ .
- Find the maximum spanning tree given these edges. The spanning tree is used as the structure of the Bayesian network.





# Multinomial Naive Bayes

## Discussion

- The implicit assumption for using the counts as the maximum likelihood estimate is that the distribution of  $X_j | Y = y$ , or in general,  $X_j | P(X_j) = p(X_j)$  has the multinomial distribution.

$$\mathbb{P}\{X_j = x | Y = y\} = p_x$$
$$\hat{p}_x = \frac{c_{x,y}}{c_y}$$



# Gaussian Naive Bayes

## Discussion

- If the features are not categorical, continuous distributions can be estimated using MLE as the conditional distribution.
- Gaussian Naive Bayes is used if  $X_j|Y = y$  is assumed to have the normal distribution.

$$\lim_{\varepsilon \rightarrow 0} \frac{1}{\varepsilon} \mathbb{P} \{x < X_j \leq x + \varepsilon | Y = y\} = \frac{1}{\sqrt{2\pi}\sigma_y^{(j)}} \exp \left( -\frac{(x - \mu_y^{(j)})^2}{2(\sigma_y^{(j)})^2} \right)$$

# Gaussian Naive Bayes Training

## Discussion

- Training involves estimating  $\mu_y^{(j)}$  and  $\sigma_y^{(j)}$  since they completely determine the distribution of  $X_j|Y = y$ .
- The maximum likelihood estimates of  $\mu_y^{(j)}$  and  $(\sigma_y^{(j)})^2$  are the sample mean and variance of the feature  $j$ .

$$\hat{\mu}_y^{(j)} = \frac{1}{n_y} \sum_{i=1}^n x_{ij} \mathbb{1}_{\{y_i=y\}}, \quad n_y = \sum_{i=1}^n \mathbb{1}_{\{y_i=y\}}$$

$$\left(\hat{\sigma}_y^{(j)}\right)^2 = \frac{1}{n_y} \sum_{i=1}^n \left(x_{ij} - \hat{\mu}_y^{(j)}\right)^2 \mathbb{1}_{\{y_i=y\}}$$

sometimes  $\left(\hat{\sigma}_y^{(j)}\right)^2 \approx \frac{1}{n_y - 1} \sum_{i=1}^n \left(x_{ij} - \hat{\mu}_y^{(j)}\right)^2 \mathbb{1}_{\{y_i=y\}}$



# Tree Augmented Network Algorithm Diagram

## Discussion