CS540 Introduction to Artificial Intelligence

Lecture 8

Young Wu

Based on lecture slides by Jerry Zhu, Yingyu Liang, and Charles Dyer

June 20, 2022

Discriminative Model vs Generative Model

Motivation

Generative Models

Motivation

- In probability terms, discriminative models are estimating $\mathbb{P}\left\{Y|X\right\}$, the conditional distribution. For example, $a_i \approx \mathbb{P}\left\{y_i = 1|x_i\right\}$ and $1 - a_i \approx \mathbb{P}\left\{y_i = 0|x_i\right\}$.
- Generative models are estimating $\mathbb{P}\left\{Y, X\right\}$, the joint distribution.
- Bayes rule is used to perform classification tasks.

$$\mathbb{P}\left\{Y|X\right\} = \frac{\mathbb{P}\left\{Y, X\right\}}{\mathbb{P}\left\{X\right\}} = \frac{\mathbb{P}\left\{X|Y\right\}\mathbb{P}\left\{Y\right\}}{\mathbb{P}\left\{X\right\}}$$

Joint Distribution

Motivation

- The joint distribution of X_j and $X_{j'}$ provides the probability of $X_j = x_j$ and $X_{j'} = x_{j'}$ occur at the same time.
$$\mathbb{P}\left\{X_j = x_j, X_{j'} = x_{j'}\right\}$$
- The marginal distribution of X_j can be found by summing over all possible values of $X_{j'}$.
$$\mathbb{P}\left\{X_j = x_j\right\} = \sum_{x \in X_{j'}} \mathbb{P}\left\{X_j = x_j, X_{j'} = x\right\}$$

Conditional Distribution

Motivation

- Suppose the joint distribution is given.

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

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

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

Bayes Rule Example 1

Quiz

- Two documents A and B. Suppose A contains 1 "Groot" and 9 other words, and B contains 8 "Groot" and 2 other words. One document is taken out A with probably $\frac{2}{3}$ and B with probably $\frac{1}{3}$, and one word is picked out at random with equal probabilities. The word is "Groot". What is the probability that the document is A?

Bayes Rule Example 1 Distribution

Quiz

Bayes Rule Example 2

Quiz

- Two documents A and B. Suppose A contains 1 "Groot" and 9 other words, and B contains 8 "Groot" and 2 other words. One document is taken out at random (with equal probability), and one word is picked out at random (all words with equal probability). The word is "Groot". What is the probability that the document is A?
- A : $\frac{1}{9}$, B: $\frac{1}{20}$, C: $\frac{2}{5}$, D: $\frac{9}{20}$, E: I don't understand

Bayesian Network

Definition

- A Bayesian network is a directed acyclic graph (DAG) and a set of conditional probability distributions.
- Each vertex represents a feature X_j.
- Each edge from X_j to $X_{j'}$ represents that X_j directly influences $X_{j'}$.
- No edge between X_j and $X_{j'}$ implies independence or conditional independence between the two features.

Conditional Independence

Definition

- Recall two events A, B are independent if:

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

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

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

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 Cause

Definition

- For three events A, B, C, the configuration $A \leftarrow B \rightarrow C$ is called common cause.
- 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 $P\left(X_j\right)$ denote the parents of the vertex X_j, and $p\left(X_j\right)$ be realizations (possible values) of $P\left(X_j\right)$.

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

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

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

Bayesian Network Diagram

Quiz

- Story: either Amber (H) or Johnny's dog (D) stepped on a bee, and put something on Johnny's bed (B), and given there is something on Johnny's bed (B), Johnny (J) and Amber (A) can be unhappy.

H	D	B	J	A
0	0	0	1	0
0	0	0	0	1
1	0	0	1	1
0	1	0	1	1
0	0	1	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0

Bayesian Network Diagram CPT Count

Quiz

Bayes Net Training Example, Training

Quiz

- Given a network and the training data.
  $H \rightarrow B, D \rightarrow B, B \rightarrow J, B \rightarrow A$.

H	D	B	J	A
0	0	0	1	0
0	0	0	0	1
1	0	0	1	1
0	1	0	1	1
0	0	1	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0

Bayes Net Training Example, Training 1

Quiz

- Compute $\hat{\mathbb{P}}\{D = 1\}$.

H	D	B	J	A
0	0	0	1	0
0	0	0	0	1
1	0	0	1	1
0	1	0	1	1
0	0	1	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0

Bayes Net Training Example, Training 2

Quiz

- Compute $\hat{\mathbb{P}}\left\{J = 1 \middle| B = 1\right\}$.

H	D	B	J	A
0	0	0	1	0
0	0	0	0	1
1	0	0	1	1
0	1	0	1	1
0	0	1	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0

Bayes Net Training Example, Training 3

Quiz

- What is the conditional probability $\hat{\mathbb{P}}\{J = 1 | B = 0\}$?
- A : I don't understand, B: $\frac{1}{4}$, C: $\frac{1}{2}$, D: $\frac{3}{4}$, E: 1

H	D	B	J	A
0	0	0	1	0
0	0	0	0	1
1	0	0	1	1
0	1	0	1	1
0	0	1	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0

Bayes Net Training Example, Training 4

Quiz

- Compute $\hat{\mathbb{P}}\left\{B = 1 \middle| H = 0, D = 1\right\}$.

H	D	B	J	A
0	0	0	1	0
0	0	0	0	1
1	0	0	1	1
0	1	0	1	1
0	0	1	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0

Bayes Net Training Example, Training 5

Quiz

- What is the conditional probability $\hat{\mathbb{P}}\left\{B = 1 \middle| H = 0, D = 0\right\}$?
- A : I don't understand, B: $\frac{1}{4}$, C: $\frac{1}{2}$, D: $\frac{3}{4}$, E: 1

H	D	B	J	A
0	0	0	1	0
0	0	0	0	1
1	0	0	1	1
0	1	0	1	1
0	0	1	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0

Bayes Net Training Example, Training 5

Quiz

- What is the conditional probability $\hat{\mathbb{P}}\left\{A = 0 \middle| H = 1, D = 1\right\}$?
- A : I don't understand, B: 0 , C: $\dfrac{1}{2}$, D: 1 , E: NA

H	D	B	J	A
0	0	0	1	0
0	0	0	0	1
1	0	0	1	1
0	1	0	1	1
0	0	1	1	0
0	0	1	0	1
1	0	1	1	0
0	1	1	1	0

Laplace Smoothing

Definition

- Recall that the MLE estimation can incorporate Laplace smoothing.

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

- Here, $\left|X_j\right|$ 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.

$$\mathbb{P}\left\{x_1, x_2, ..., x_m\right\} = \prod_{j=1}^{m} \mathbb{P}\left\{x_j \middle| x_1, x_2, ..., x_{j-1}, x_{j+1}, ..., x_m\right\}$$
$$= \prod_{j=1}^{m} \mathbb{P}\left\{x_j \middle| p\left(X_j\right)\right\}$$

Bayes Net Inference 2

Definition

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

$$\mathbb{P}\left\{x_j \middle| x_{j'}, x_{j''}, \ldots\right\} = \frac{\mathbb{P}\left\{x_j, x_{j'}, x_{j''}, \ldots\right\}}{\mathbb{P}\left\{x_{j'}, x_{j''}, \ldots\right\}}$$

$$\mathbb{P}\left\{x_j, x_{j'}, x_{j''}, \ldots\right\} = \sum_{X_k : k \neq j, j', j'', \ldots} \mathbb{P}\left\{x_1, x_2, \ldots, x_m\right\}$$

$$\mathbb{P}\left\{x_{j'}, x_{j''}, \ldots\right\} = \sum_{X_k : k \neq j', j'', \ldots} \mathbb{P}\left\{x_1, x_2, \ldots, x_m\right\}$$

Bayes Net Inference Example 1

Quiz

- Assume the network is trained on a larger set with the following CPT. Compute $\hat{\mathbb{P}}\left\{H = 0, D = 1 | J = 1, A = 0\right\}$?

$$\hat{\mathbb{P}}\left\{H = 1\right\} = 0.001, \hat{\mathbb{P}}\left\{D = 1\right\} = 0.001$$

$$\hat{\mathbb{P}}\left\{B = 1 | H = 1, D = 1\right\} = 0.95, \hat{\mathbb{P}}\left\{B = 1 | H = 1, D = 0\right\} = 0.94$$

$$\hat{\mathbb{P}}\left\{B = 1 | H = 0, D = 1\right\} = 0.29, \hat{\mathbb{P}}\left\{B = 1 | H = 0, D = 0\right\} = 0.00$$

$$\hat{\mathbb{P}}\left\{J = 1 | B = 1\right\} = 0.9, \hat{\mathbb{P}}\left\{J = 1 | B = 0\right\} = 0.05$$

$$\hat{\mathbb{P}}\left\{A = 1 | B = 1\right\} = 0.7, \hat{\mathbb{P}}\left\{A = 1 | B = 0\right\} = 0.01$$

Bayes Net Inference Example 1 Computation 1

Quiz

Bayes Net Inference Example 1 Computation 2

Quiz

Bayes Net Inference Example 2

Quiz

- Compute $\hat{\mathbb{P}}\left\{D = 1 | H = 0\right\}$?

$$\hat{\mathbb{P}}\left\{H = 1\right\} = 0.001, \hat{\mathbb{P}}\left\{D = 1\right\} = 0.001$$

$$\hat{\mathbb{P}}\left\{B = 1 | H = 1, D = 1\right\} = 0.95, \hat{\mathbb{P}}\left\{B = 1 | H = 1, D = 0\right\} = 0.94$$

$$\hat{\mathbb{P}}\left\{B = 1 | H = 0, D = 1\right\} = 0.29, \hat{\mathbb{P}}\left\{B = 1 | H = 0, D = 0\right\} = 0.00$$

- A : 0, B: 0.001, C: 0.0094, D: 0.0095, E: 1

Bayes Net Inference Example 2 Derivation

Quiz

Bayes Net Inference Example 3

Quiz

- Compute $\hat{\mathbb{P}}\left\{H = 0, D = 1 \middle| B = 1\right\}$?

$$\hat{\mathbb{P}}\left\{H = 1\right\} = 0.001, \hat{\mathbb{P}}\left\{D = 1\right\} = 0.001$$

$$\hat{\mathbb{P}}\left\{B = 1 \middle| H = 1, D = 1\right\} = 0.95, \hat{\mathbb{P}}\left\{B = 1 \middle| H = 1, D = 0\right\} = 0.94$$

$$\hat{\mathbb{P}}\left\{B = 1 \middle| H = 0, D = 1\right\} = 0.29, \hat{\mathbb{P}}\left\{B = 1 \middle| H = 0, D = 0\right\} = 0.00$$

- A : 0, B: 0.001, C: 0.0094, D: 0.0095, E: 1

Bayes Net Inference Example 3 Derivation

Quiz

Bayes Net Inference Example 4

Quiz

- Compute $\hat{\mathbb{P}}\left\{B = 1 \middle| J = 1, A = 0\right\}$?

$$\hat{\mathbb{P}}\left\{J = 1 \middle| B = 1\right\} = 0.9, \hat{\mathbb{P}}\left\{J = 1 \middle| B = 0\right\} = 0.05$$

$$\hat{\mathbb{P}}\left\{A = 1 \middle| B = 1\right\} = 0.7, \hat{\mathbb{P}}\left\{A = 1 \middle| B = 0\right\} = 0.01$$

Given

$\mathbb{P}\left\{B = 1\right\} = 0.001 \cdot 0.001 \cdot 0.95 + 0.001 \cdot 0.999 \cdot (0.94 + 0.29)$.

- A : 0, B: 0.001, C: 0.0094, D: 0.0095, E: 1

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.

Classification Problem

Discussion

- Bayesian networks do not have a clear separation of the label Y and the features $X_1, X_2, ..., X_m$.
- The Bayesian network with a tree structure and Y as the root and $X_1, X_2, ..., X_m$ as the leaves is called the Naive Bayes classifier.
- Bayes rules is used to compute $\mathbb{P}\left\{Y = y | X = x\right\}$, and the prediction $\hat{y}$ is y that maximizes the conditional probability.

$$\hat{y}_i = \underset{y}{\operatorname{argmax}}\, \mathbb{P}\left\{Y = y | X = x_i\right\}$$

Naive Bayes Diagram

Discussion

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.

$$\mathbb{P}\left\{X_j = x | Y = y\right\} = p_x$$

$$\hat{p}_x = \frac{c_{x,y}}{c_y}$$

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.

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

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 $\left(\sigma_y^{(j)}\right)^2$ are the sample mean and variance of the feature j.

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

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

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

Tree Augmented Network Algorithm

Discussion

- It is also possible to create a Bayesian network with all features $X_1, X_2, ..., X_m$ connected to Y (Naive Bayes edges) and the features themselves form a network, usually a tree (MST edges).
- Information gain is replaced by conditional information gain (conditional on Y) when finding the maximum spanning tree.
- This algorithm is called TAN: Tree Augmented Network.

Tree Augmented Network Algorithm Diagram

Discussion