CS540 Introduction to Artificial Intelligence Lecture 9

Young Wu

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

June 23, 2022

Discriminative Model vs Generative Model Motivation

Generative Models

Motivation

- In probability terms, discriminative models are estimating $\mathbb{P}\{Y|X\}$, the conditional distribution. For example, $a_i \approx \mathbb{P}\{y_i = 1|x_i\}$ and $1 a_i \approx \mathbb{P}\{y_i = 0|x_i\}$.
- Generative models are estimating $\mathbb{P}\{Y,X\}$, 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

Bayes Rule Example 1 Distribution Quiz

Bayes Rule Example 2

Bayesian Network

- 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:

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

 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

- 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

- 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

- 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

• 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 $P(X_j)$.

$$\mathbb{P}\left\{ x_{j}|\rho\left(X_{j}\right)\right\} ,\rho\left(X_{j}\right)\in P\left(X_{j}\right)$$

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

$$\widehat{\mathbb{P}}\left\{x_{j}|p\left(X_{j}\right)\right\} = \frac{c_{x_{j},p}(x_{j})}{c_{p}(x_{i})}$$

Bayesian Network Diagram Quiz

Bayesian Network Diagram CPT Count

Bayes Net Training Example, Training Quiz

Bayes Net Training Example, Training 1

Bayes Net Training Example, Training 3 Quiz

Bayes Net Training Example, Training 4 Quiz

Bayes Net Training Example, Training 5

Bayes Net Training Example, Training 5

Laplace Smoothing

 Recall that the MLE estimation can incorporate Laplace smoothing.

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

- 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.

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

Bayes Net Inference 2

Definition

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

$$\mathbb{P}\left\{x_{j}|x_{j'},x_{j''},...\right\} = \frac{\mathbb{P}\left\{x_{j},x_{j'},x_{j''},...\right\}}{\mathbb{P}\left\{x_{j'},x_{j''},...\right\}}
\mathbb{P}\left\{x_{j},x_{j'},x_{j''},...\right\} = \sum_{X_{k}:k\neq j,j',j'',...} \mathbb{P}\left\{x_{1},x_{2},...,x_{m}\right\}
\mathbb{P}\left\{x_{j'},x_{j''},...\right\} = \sum_{X_{k}:k\neq j',j'',...} \mathbb{P}\left\{x_{1},x_{2},...,x_{m}\right\}$$

Bayes Net Inference Example 1

Bayes Net Inference Example 1 Computation 1

Bayes Net Inference Example 1 Computation 2

Bayes Net Inference Example 2

Bayes Net Inference Example 2 Derivation Quiz

Bayes Net Inference Example 3

Bayes Net Inference Example 3 Derivation Quiz

Bayes Net Inference Example 4

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

- 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 $\mathbb{P}\{Y = y | X = x\}$, and the prediction \hat{y} is y that maximizes the conditional probability.

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

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.

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

Gaussian Naive Bayes Training

Discussion

- Training involves estimating $\mu_y^{(j)}$ and $\sigma_y^{(j)}$ since they completely determine the distribution of $X_i|Y=y$.
- 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.

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

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

- 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