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

- Previous weeks' focus is on discriminative models.
- Given a training set $(x_i, y_i)_{i=1}^{n}$, the task is classification (machine learning) or regression (statistics), *i.e.* finding a function $\hat{f}$ such that given new instances x_i', y can be predicted as $\hat{y}_i = \hat{f}(x_i')$.
- The function $\hat{f}$ is usually represented by parameters w and b. These parameters can be learned by methods such as gradient descent by minimizing some cost objective function.

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.

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

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.

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

Notation

Motivation

- The notations for joint, marginal, and conditional distributions will be shortened as the following.

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

- When the context is not clear, for example when $x_j = a, x_{j'} = b$ with specific constants a, b, subscripts will be used under the probability sign.

$$\mathbb{P}_{x_j, x_{j'}}\left\{a, b\right\}, \mathbb{P}_{x_j}\left\{a\right\}, \mathbb{P}_{x_j | x_{j'}}\left\{a \middle| b\right\}$$

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:

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

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

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

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 $A \to B \leftarrow C$ 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.

Storing Distribution

Definition

- If there are m binary variables with k edges, there are 2^m joint probabilities to store.
- There are significantly less conditional probabilities to store. For example, if each node has at most 2 parents, then there are less than $4m$ conditional probabilities to store.
- Given the conditional probabilities, the joint probabilities can be recovered.

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.

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

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.

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

$$= \prod_{j=1}^{m} \mathbb{P}\left\{x_j \middle| p\left(X_j\right)\right\}$$

Bayes Net Inference 2

Definition

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

$$\mathbb{P}\left\{x_j | x_{j'}, x_{j''}, ...\right\} = \dfrac{\mathbb{P}\left\{x_j, x_{j'}, x_{j''}, ...\right\}}{\mathbb{P}\left\{x_{j'}, x_{j''}, ...\right\}}$$

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

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

Bayesian Network

Algorithm

- Input: instances: $\{x_i\}_{i=1}^{n}$ and a directed acyclic graph such that feature X_j has parents $P(X_j)$.
- Output: conditional probability tables (CPTs): $\hat{\mathbb{P}}\{x_j | p(X_j)\}$ for $j = 1, 2, ..., m$.
- Compute the transition probabilities using counts and Laplace smoothing.

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

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.

Aside: Prim's Algorithm

Discussion

- To find the maximum spanning tree, start with an arbitrary vertex, a vertex set containing only this vertex, V, and an empty edge set, E.
- Choose an edge with the maximum weight from a vertex $v \in V$ to a vertex $v' \notin V$ and add v' to V, add an edge from v to v' to E
- Repeat this process until all vertices are in V. The tree (V, E) is the maximum spanning tree.

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.

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

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.