CZ4041/CE4041: Machine Learning

Lesson 2a: Overview of Bayesian Classifiers

Sinno Jialin PAN

School of Computer Science and Engineering, NTU, Singapore

Uncertainty in Prediction

- Recall: in supervised learning, given a set of $\{\boldsymbol{x}_i, y_i\}$ for $i = 1, \ldots, N$, the goal is to learn a mapping $f: \boldsymbol{x} \rightarrow y$ by requiring $f(\boldsymbol{x}_i) = y_i$
- In many applications, the mapping or relationship f between the input features and the output labels is non-deterministic (uncertain)
- For example, suppose you are asked to predict a result of the final of a football cup between Team 1 and Team 2: which team will win?

Bayesian Classifiers

- From a probability point of view, the mapping $f: \boldsymbol{x} \rightarrow y$ can be modeled as a conditional probability $P(y|\boldsymbol{x})$
- Bayesian classifiers aim to learn the mapping $f: \boldsymbol{x} \rightarrow y$ for supervised learning in the form of conditional probability $P(y|\boldsymbol{x})$, such that for any input $\boldsymbol{x}^*$, one can use $P(y = c|\boldsymbol{x}^*)$ to predict the probability of $\boldsymbol{x}^*$ belonging to class c, where $c \in \{0, \ldots, C - 1\}$
  - How to estimate $P(y = c|\boldsymbol{x}^*)$ for different classes?
  - How to make use of $P(y = c|\boldsymbol{x}^*)$'s to make a prediction?
  - We first review some important probability concepts

Marginal Probability

- Let A be a random variable (an input feature / class label in machine learning)
- Marginal probability

$$P(A = a) \qquad 0 \leq P(A = a) \leq 1$$

refers to the probability that variable $A = a$

$$\sum_{a_i} P(A = a_i) = 1$$

- For example, let A denote who will be the winner in the final: $A = 1$ means Team 1 and $A = 0$ means Team 2. $P(A = 1) + P(A = 0) = 1$

Joint Probability

- Let A and B be a pair of random variables (features/labels in machine learning).
- Their joint probability

$$P(A = a, B = b)$$

refers to the probability that variable $A = a$ and variable $B = b$

Joint Probability (cont.)

- For example
  - A denotes the result of a match M_1 between Team 1 and Team 2: $A = 1$ means Team 1 wins, and $A = 0$ means Team 2 wins
  - B denotes the result of another match M_2 between Team 3 and Team 4: $B = 1$ means Team 3 wins, and $B = 0$ means Team 4 wins
  - $P(A = 1, B = 1)$ denotes the probability that Team 1 wins M_1 meanwhile Team 3 wins M_2
  - $P(A = 1, B = 0)$ denotes the probability that Team 1 wins M_1 meanwhile Team 3 loses M_2

Conditional Probability

- Conditional probability:

$$P(B = b|A = a)$$

refers to the probability that the variable B will take on the value b, given that the variable A is observed to have the value a

$$\sum_{b_i} P(B = b_i|A = a) = 1$$

Conditional Probability (cont.)

- Use the previous example
  - $A = 1$: Team 1 wins, and $A = 0$: Team 2 wins M_1
  - $B = 1$: Team 3 wins, and $B = 0$: Team 4 wins M_2
  - $P(B = 1|A = 1)$ denotes the probability that Team 3 will win <u>given</u> that Team 1 wins M_1
  - $P(B = 1|A = 0)$ denotes the probability that Team 3 will win <u>given</u> that Team 1 loses M_1

$$P(B = 1|A = 1) + P(B = 0|A = 1) = 1$$

$$P(B = 1|A = 0) + P(B = 0|A = 0) = 1$$

Sum Rule

- The connection between joint probability of A and B and marginal probability of A:

$$P(A=a)=\sum_{b_i} P(A=a, B=b_i) \quad \textbf{OR} \quad P(A)=\sum_{B} P(A,B)$$

$$P(A=a)=\sum_{c_j}\sum_{b_i} P(A=a, B=b_i, C=c_j) \quad \textbf{OR} \quad P(A)=\sum_{C}\sum_{B} P(A,B,C)$$

- Use the previous example

$$\boxed{P(A=1,B=1)}+\boxed{P(A=1,B=0)}=P(A=1)$$

Probability that Team 1 wins M_1 and at the same time Team 3 wins M_2

Probability that Team 1 wins M_1 and at the same time Team 3 loses M_2

Probability that Team 1 wins M_1

Product Rule

- The connections between joint, conditional and marginal probabilities for A and B:

$$P(A = a, B = b) = P(B = b|A = a) \times P(A = a)$$
$$= P(A = a|B = b) \times P(B = b)$$

OR

$$P(A, B) = P(B|A) \times P(A) = P(A|B) \times P(B)$$

- Use the previous example

$$\boxed{P(A = 1)} \times \boxed{P(B = 1|A = 1)} = P(A = 1, B = 1)$$

Probability that Team 1 wins M_1

Probability that Team 3 wins M_2 if Team 1 wins M_1

Probability that both Team 1 and Team 3 win the matches

Bayes Rule or Bayes Theorem

$$P(A|B) = \frac{P(A,B)}{P(B)} = \frac{P(B|A)P(A)}{P(B)}$$

- This is induced from product rule

$$P(A,B) = P(A|B) \times P(B) = P(B|A) \times P(A)$$

- Generalized to the case when **A** and **B** are a set of variables

$$P\left(A_1 \dots A_k \middle| B_1 \dots B_p\right) = \frac{P(B_1 \dots B_p, A_1 \dots A_k)}{P(B_1 \dots B_p)}$$

$$= \frac{P\left(B_1 \dots B_p \middle| A_1 \dots A_k\right) P(A_1 \dots A_k)}{P(B_1 \dots B_p)}$$

An Example

v.s.

- Suppose the next match between the two teams will be hosted by Manchester United
- Head to Head Statistics:

Team	Played	Win	Draw	Lose
Manchester United	151	59	47	45
Manchester City	151	45	47	59

  - Among the 59 victories for Manchester United, 32 of them come from playing at home
  - Among the games won by Manchester City, 20 of them are obtained while playing on Manchester United home ground
  - Among the draw games, 23 of them were played on Manchester United home ground

What result will likely be for the match?

Define Variables

- Let Y be the random variable that represents the result of the match (0, 1, 2)
  - $Y = 0$: Manchester United wins the match
  - $Y = 1$: Manchester City wins the match
  - $Y = 2$: Draw
- Let X be the random variable that represents the team hosting the match (0 or 1)
  - $X = 0$: Manchester United hosts the match
  - $X = 1$: Manchester City hosts the match
- To estimate $P(Y = 0|X = 0)$, $P(Y = 1|X = 0)$, and $P(Y = 2|X = 0)$

Estimate Probabilities

Team	Played	Win	Draw	Lose
Manchester United	151	59	47	45
Manchester City	151	45	47	59

- To calculate $P(Y = 0)$, $P(Y = 1)$, and $P(Y = 2)$
  - $P(Y = 0) = \frac{59}{151} \approx 39\%$
  - $P(Y = 1) = \frac{45}{151} \approx 30\%$
  - $P(Y = 2) = 1 - P(Y = 0) - P(Y = 1) = 31\%$

$$\sum_{y_i} P(Y = y_i) = 1$$

Estimate Probabilities (cont.)

- Among the 59 victories for Manchester United, 32 of them come from playing at home

$$P(X = 0|Y = 0) = \frac{32}{59} = 54\%$$

- Among the games won by Manchester City, 20 of them are obtained while playing on Manchester United home ground

$$P(X = 0|Y = 1) = \frac{20}{45} = 44\%$$

- Among the draw games, 23 of them were played on Manchester United home ground

$$P(X = 0|Y = 2) = \frac{23}{47} = 49\%$$

- However, the goal is to estimate

$$P(Y = 0|X = 0) \quad \text{v.s} \quad P(Y = 1|X = 0) \quad \text{v.s} \quad P(Y = 2|X = 0)$$

Apply Bayes Rule

- Probability that Manchester United wins: $P(Y = 0) = 0.39$
- Probability that Manchester City wins: $P(Y = 1) = 0.3$
- Probability of a draw game: $P(Y = 2) = 0.31$
- Probability that Manchester United hosted the match it won: $P(X = 0|Y = 0) = 0.54$
- Probability that Manchester United hosted the match won by Manchester City: $P(X = 0|Y = 1) = 0.44$
- Probability that Manchester United hosted the match that is a draw game: $P(X = 0|Y = 2) = 0.49$
- To use Bayes rule to compute $P(Y = 1|X = 0)$, $P(Y = 0|X = 0)$ and $P(Y = 2|X = 0)$

$$P(Y=1|X=0)$$

Bayes rule

$$= \frac{P(X=0|Y=1)\times P(Y=1)}{\boxed{P(X=0)}}$$

Sum rule: $P(X)=\sum_Y P(X,Y)$

$$= \frac{P(X=0|Y=1)\times P(Y=1)}{\boxed{P(X=0,Y=1)+P(X=0,Y=0)+P(X=0,Y=2)}}$$

$$= \frac{P(X=0|Y=1)\times P(Y=1)}{P(X=0|Y=1)\times P(Y=1)+P(X=0|Y=0)\times P(Y=0)+P(X=0|Y=2)\times P(Y=2)}$$

Product rule: $P(X,Y)=P(X|Y)P(Y)$

$$= \frac{0.44\times 0.3}{0.44\times 0.3+0.54\times 0.39+0.49\times 0.31} = \frac{0.132}{0.4945} = 0.267$$

Similarly

Tutorial

$$\boxed{P(Y=0|X=0)=0.426 \quad P(Y=2|X=0)=0.307}$$

Bayesian Classifiers (cont.)

- Bayesian classifiers aim to learn the mapping $f: \boldsymbol{x} \rightarrow y$ for supervised learning in the form of conditional probability $P(y|\boldsymbol{x})$ via Bayes rule

posterior

$$\boxed{P(y|\boldsymbol{x})} = \frac{P(y, \boldsymbol{x})}{P(\boldsymbol{x})} = \frac{P(\boldsymbol{x}|y)\boxed{P(y)}}{P(\boldsymbol{x})}$$

prior

- For a classification problem with C classes, given a test data instance $\boldsymbol{x}^*$, a Bayesian classifier computes

$$P(y = c|\boldsymbol{x}^*), c \in \{0, \dots, C-1\}, \text{ and } \textstyle\sum_c P(y = c|\boldsymbol{x}^*) = 1$$

- Make a prediction based on the maximum posterior

$$y^* = c^* \text{ if } c^* = \boxed{\arg\max_c} P(y = c|\boldsymbol{x}^*)$$

where $c \in \{0, \dots, C-1\}$

Return the value of c that maximizes $P(y = c|\boldsymbol{x}^*)$

Bayesian Classifiers (cont.)

- Based on Bayes rule

$$y^* = c^* \text{ if } c^* = \arg\max_c P(y = c|\boldsymbol{x}^*), c \in \{0, \dots, C-1\}$$

$$= \arg\max_c \frac{P(\boldsymbol{x}^*|y = c)P(y = c)}{\boxed{P(\boldsymbol{x}^*)}}$$

Constant w.r.t. diff. values of c

$$= \arg\max_c P(\boldsymbol{x}^*|y = c)P(y = c)$$

- Therefore, we make a prediction based on

$$y^* = c^* \text{ if } c^* = \arg\max_c P(\boldsymbol{x}^*|y = c)P(y = c), c \in \{0, \dots, C-1\}$$

- Take binary classification as an example: 0 vs 1

$$P(y = 0|\boldsymbol{x}) = \frac{\boxed{P(\boldsymbol{x}|y = 0)P(y = 0)}}{\cancel{P(\boldsymbol{x})}} \quad vs \quad P(y = 1|\boldsymbol{x}) = \frac{\boxed{P(\boldsymbol{x}|y = 1)P(y = 1)}}{\cancel{P(\boldsymbol{x})}}$$

Notes on Bayesian Classifiers

- Why not computing $P(y = c|\boldsymbol{x})$ for each class directly from the training data, but using Bayes rule to estimate $P(\boldsymbol{x}|y = c)$ and $P(y = c)$ instead?
  - To estimate $P(y = c|\boldsymbol{x})$, one needs to consider each combination of values of $\boldsymbol{x}$
  - E.g., suppose $\boldsymbol{x}$ is m-dimensional and each feature has binary values, then the total number of possible value combination of $\boldsymbol{x}$ is 2^m --- require a huge size of training dataset and time consuming, not practical!
  - The form of $P(\boldsymbol{x}|y = c)$ can be decomposed based on some assumptions and probability properties --- no need to consider all possible combination

Summary on Bayesian Classifiers

- Estimate $P(y|\boldsymbol{x})$ via Bayes rule

$$P(y=c|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|y=c)P(y=c)}{P(\boldsymbol{x})}$$

- Make predictions based on maximum posterior

$$y^* = c^* \text{ if } c^* = \arg\max_c \frac{P(\boldsymbol{x}|y=c)P(y=c)}{P(\boldsymbol{x})}$$

$$\Updownarrow$$

$$y^* = c^* \text{ if } c^* = \arg\max_c \boxed{P(\boldsymbol{x}|y=c)P(y=c)}$$

Advanced decision making: Bayesian Decision Theory (Lecture 2b)

How to estimate from training data?

- Two implementations will be introduced
  - Naïve Bayes Classifier (Lecture 3)
  - Bayesian Brief Networks (Lecture 4)

Thank you!