Bayesian Approaches

Data Mining Selected Technique

Henry Xiao

`xiao@cs.queensu.ca`

School of Computing

Queen's University

Probabilistic Bases

Review the fundamentals of Probability Theory.

- A is a *Boolean-valued variable* if A denotes an event, and there is some degree of uncertainty as to whether A occurs.
- $P(A)$ denotes the fraction of possible worlds in which A is true.
- The **axioms** of *general* probability:
  - $0 \leq P(A) \leq 1$.
  - $P(true) = 1$.
  - $P(false) = 0$.
  - $P(\neg A) + P(A) = 1$.
  - $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.

Probabilistic Bases

Multivalued Random Variables.

- A is a *random variable with arity k* if it can take on exactly one value out of $v_1, v_2, \ldots, v_k$.
- $P(A = v_i \cap A = v_j) = 0$ if $i \neq j$.
- $P(A = v_1 \cup A = v_2 \cup \ldots A = v_k) = \sum_{i=1}^{k} P(A = v_i) = 1$.
- $P(B) = \sum_{i=1}^{k} P(B \cap A = v_i)$.

Bayes Rule

Conditional Probability and Bayes Rule.

- $P(A|B)$ is a fraction of worlds in which A is true given B is true.
- $P(A|B) = \frac{P(A \cap B)}{P(B)} \Longrightarrow P(A \cap B) = P(A|B)P(B)$ *(Chain Rule)*
- $P(B|A) = \frac{P(A|B)P(B)}{P(A)}$ *(Bayes Rule)*
- $P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B)+P(A|\neg B)P(\neg B)}$.
- $P(B|A \cap C) = \frac{P(A|B \cap C)P(B \cap C)}{P(A \cap C)}$.

Density Estimation

- A Density Estimator M learns a mapping from a set of attributes to a Probability

Queen's UNIVERSITY

Density Estimation

- A Density Estimator M learns a mapping from a set of attributes to a Probability
- Given a record x, M can tell you how likely the record is: $P(x|M)$.

Queen's UNIVERSITY

Density Estimation

- A Density Estimator M learns a mapping from a set of attributes to a Probability
- Given a record x, M can tell you how likely the record is: $P(x|M)$.
- Given a dataset with R records, M can tell you how likely the dataset is: $P(dataset|M) = P(x_1 \cap x_2 \cap \ldots x_R|M) = \prod_{k=1}^{R} P(x_k|M)$.
  - Joint Distribution Estimator.
  - Naïve Estimator.

Density Estimation

- A Density Estimator M learns a mapping from a set of attributes to a Probability
- Given a record x, M can tell you how likely the record is: $P(x|M)$.
- Given a dataset with R records, M can tell you how likely the dataset is:
  $P(dataset|M) = P(x_1 \cap x_2 \cap \dots x_R|M) = \prod_{k=1}^{R} P(x_k|M)$.
  - Joint Distribution Estimator.
  - Naïve Estimator.
- Problem:
  - The Joint Estimator just mirrors the training data - overfitting.
  - Naïve Estimator assumes each attribute is independent.

Bayes Classifiers

Build a Bayes Classifier.

Bayes Classifiers

Build a Bayes Classifier.

- Assume we want to predict Y which has arity n and values $v_1, v_2, \ldots, v_n$.
- Assume there are m input attributes $X_1, X_2, \ldots, X_m$.

Bayes Classifiers

Build a Bayes Classifier.

- Assume we want to predict Y which has arity n and values $v_1, v_2, \ldots, v_n$.
- Assume there are m input attributes $X_1, X_2, \ldots, X_m$.
- Break dataset into n smaller datasets $DS_1, DS_2, \ldots, DS_n$.

Bayes Classifiers

Build a Bayes Classifier.

- Assume we want to predict Y which has arity n and values $v_1, v_2, \ldots, v_n$.
- Assume there are m input attributes $X_1, X_2, \ldots, X_m$.
- Break dataset into n smaller datasets $DS_1, DS_2, \ldots, DS_n$.
- Let DS_i contains the records where $Y = v_i$.

Bayes Classifiers

Build a Bayes Classifier.

- Assume we want to predict Y which has arity n and values $v_1, v_2, \ldots, v_n$.
- Assume there are m input attributes $X_1, X_2, \ldots, X_m$.
- Break dataset into n smaller datasets $DS_1, DS_2, \ldots, DS_n$.
- Let DS_i contains the records where $Y = v_i$.
- For each DS_i, learn Density Estimator M_i to model the input distribution among the $Y = v_i$ records.

Bayes Classifiers

Build a Bayes Classifier.

- Assume we want to predict Y which has arity n and values $v_1, v_2, \ldots, v_n$.
- Assume there are m input attributes $X_1, X_2, \ldots, X_m$.
- Break dataset into n smaller datasets $DS_1, DS_2, \ldots, DS_n$.
- Let DS_i contains the records where $Y = v_i$.
- For each DS_i, learn Density Estimator M_i to model the input distribution among the $Y = v_i$ records.
- M_i estimates $P(X_1, X_2, \ldots, X_m | Y = v_i)$.

Bayes Classifiers

When a new set of input values $(X_1 = u_1, X_2 = u_2, \ldots, X_m = u_m)$ come to be evaluated, predict the value of Y.

Bayes Classifiers

When a new set of input values $(X_1 = u_1, X_2 = u_2, \ldots, X_m = u_m)$ come to be evaluated, predict the value of Y.

- $Y^{predict} = arg\max_v P(X_1 = u_1, \ldots, X_m = u_m | Y = v) \Longrightarrow$ *Maximum Likelihood Estimation (MLE)*.

Bayes Classifiers

When a new set of input values $(X_1 = u_1, X_2 = u_2, \ldots, X_m = u_m)$ come to be evaluated, predict the value of Y.

- $Y^{predict} = arg\max_v P(X_1 = u_1, \ldots, X_m = u_m | Y = v) \Longrightarrow$ *Maximum Likelihood Estimation (MLE)*.
- $Y^{predict} = arg\max_v P(Y = v | X_1 = u_1, \ldots, X_m = u_m) \Longrightarrow$ *Bayes Estimation (BE)*

Bayes Classifiers

When a new set of input values $(X_1 = u_1, X_2 = u_2, \ldots, X_m = u_m)$ come to be evaluated, predict the value of Y.

- $Y^{predict} = arg\max_v P(X_1 = u_1, \ldots, X_m = u_m | Y = v) \Longrightarrow$ *Maximum Likelihood Estimation (MLE)*.
- $Y^{predict} = arg\max_v P(Y = v | X_1 = u_1, \ldots, X_m = u_m) \Longrightarrow$ *Bayes Estimation (BE)*
- Posterior probability:

$$
\begin{aligned}
& P(Y = v | X_1 = u_1, \ldots, X_m = u_m) \\
= \quad & \frac{P(X_1 = u_1, \ldots, X_m = u_m | Y = v) P(Y = v)}{P(X_1 = u_1, \ldots, X_m = u_m)} \\
= \quad & \frac{P(X_1 = u_1, \ldots, X_m = u_m | Y = v) P(Y = v)}{\sum_{i=1}^{n} P(X_1 = u_1, \ldots, X_m = u_m | Y = v_i) P(Y = v_i)}.
\end{aligned}
$$

Bayes Classifiers

Bayes Classifier Model.

- Learn the distribution over inputs for each value Y *.
- Calculate $P(X_1, X_2, \ldots, X_m | Y = v_i)$
- Estimate $P(Y = v_i)$ as a fraction of records with $Y = v_i$.
- For a new prediction:
  $Y^{predict} = arg \max_v P(X_1 = u_1, \ldots, X_m = u_m | Y = v) P(Y = v)$.

* Learn the distribution is done by a Density Estimator here.

Gaussian Bayes Classifier

Gaussian Bayes Classifier Bases.

- Generate the output by drawing $y_i \sim Multinomial(p_1, p_2, \ldots, p_n)$
- Generate the inputs from a Gaussian PDF that depends on the value of y_i: $x_i \sim N(\mu_i, \Sigma_i)$.
- $P(y = i|\mathbf{X}) = \frac{p(\mathbf{X}|y=i)p(y=i)}{p(\mathbf{X})}$.
- $P(y = i|\mathbf{X}) = \frac{\frac{1}{(2\pi)^{m/2}\|S_i\|^{1/2}} e^{-\frac{1}{2}(x_k-\mu_i)^T S_i (x_k-\mu_i)} p_i}{p(\mathbf{X})}$
  where μ_i and S_i is the mean and variance by taking *MLE Gaussian* $(\mu_i^{mle}, \Sigma_i^{mle})$.

Gaussian Bayes Classifier

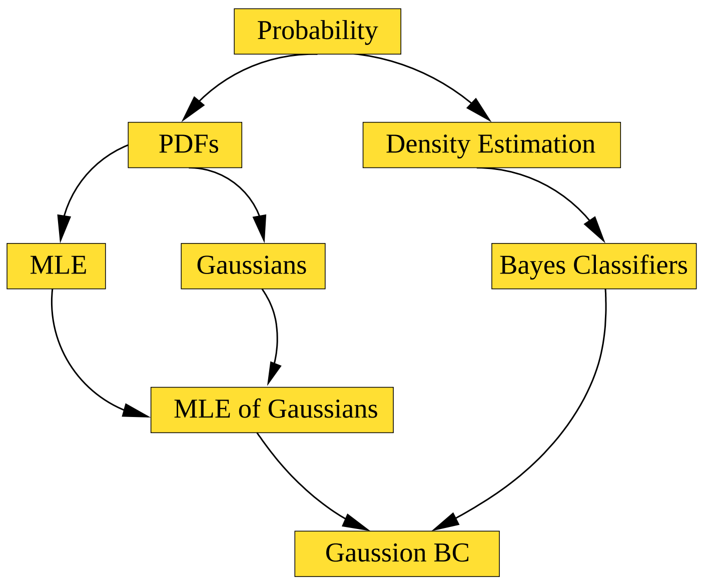
Gaussian BC Discussion.

- Number of parameters is quadratic with number of dimensions generally because of the covariance matrix.
- Normalize Gaussian may give *O(1)* covariance parameters.
- Gaussian DE and BC inputs require to be *Real-valued*.
- Gaussian Naïve or Gaussian Joint can accept both *categorical* and *real-valued* inputs.

Queen's UNIVERSITY

Bayes Classifiers Methods

Different Bayes Classifiers have been developed.



BC Road Map

- Joint Density Bayes Classifier.
- Naïve Bayes Classifier.
- Gaussian Bayes Classifier.
- Other joint Bayes Classifier possible.

Bayesian Networks

Introduction to Bayes Nets. (A Bayes Net Example)

- A Bayes net (or Belief network) is an augmented directed acyclic graph, represented by the vertex set V and directed edge set E.
- There is no loops allowed. And each vertex $v \in V$ represents a random variable.
- A probability distribution table indicating how the probability of this variable's values depends on all possible combinations of parental values.
- Two variable v_i and v_j may still correlated even if they are not connected.
- Each variable v_i is conditionally independent of all non-descendants, given its parents.

Bayesian Networks

Building a Bayes Net.

Bayesian Networks

Building a Bayes Net.

- Choose a set of relevant variables.
- Choose an ordering for them.

Bayesian Networks

Building a Bayes Net.

- Choose a set of relevant variables.
- Choose an ordering for them.
- Assume the variables are $X_1, X_2, \ldots, X_n$ (where X_1 is the first, X_i is the ith).
- for $i = 1\ to\ n$:
  - Add the X_i vertex to the network
  - Set $Parent(X_i)$ to be a minimal subset of $X_1, \ldots, X_{i-1}$, such that we have conditional independence of X_i and all other members of $X_1, \ldots, X_{i-1}$ given $Parents(X_i)$
  - Define the probability table of $P(X_i = k \mid Assignments\ of\ Parent(X_i))$.

Bayesian Networks

Bayesian Networks Discussion.

- Independence and conditional independence.
- Inference can be calculated.
- Enumerating entries is exponential in the number of variables.
- The stochastic simulation and likelihood weighting.

Conclusion

Topics we have discussed here:

- Probabilistic bases and Bayes Rule
- Density Estimation
- Bayes Model
- Bayes Classifiers
- Bayesian Networks

Bayes, Thomas(1763)

Ending

Questions regarding Bayesian Approaches?

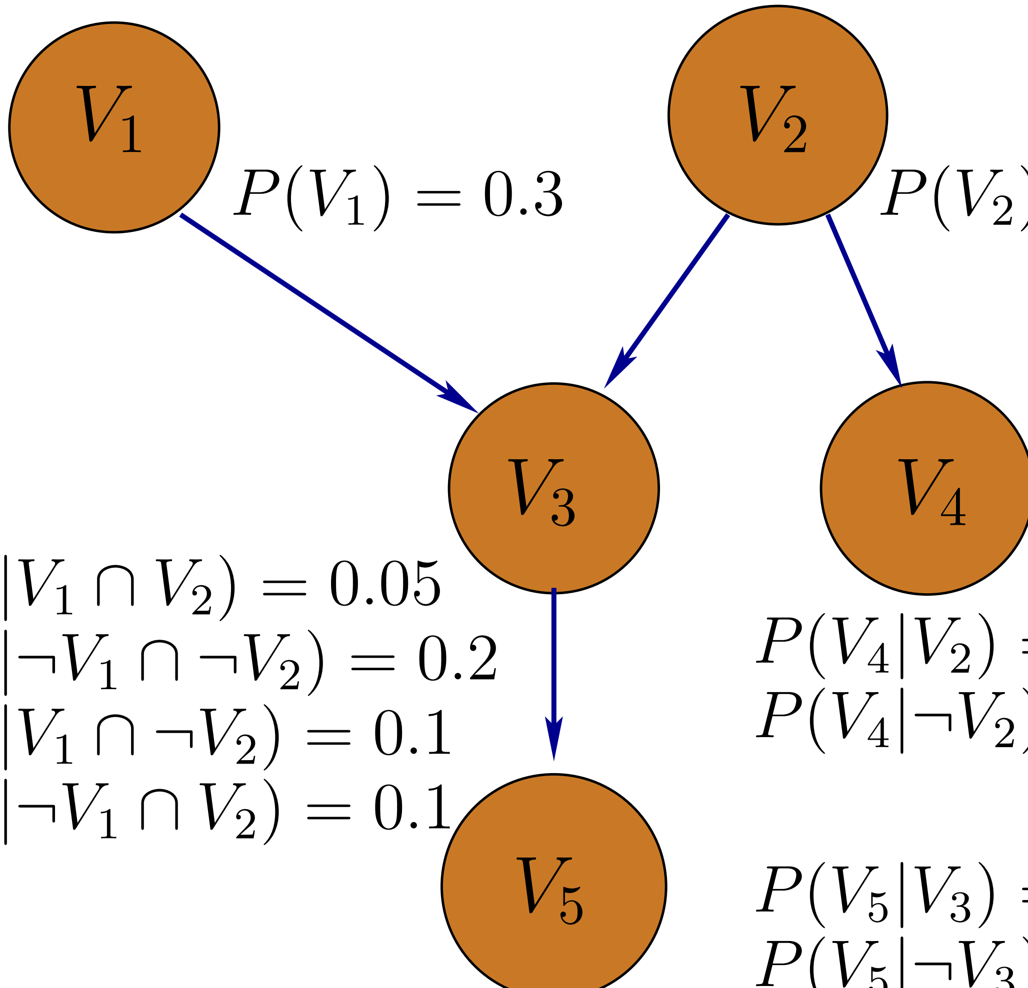
Information Site: http://www.cs.queensu.ca/home/xiao/dm.html

E-mail: xiao@cs.queens.ca

Thank you

A Bayes Net

A Bayes net example (Back to Bayes Nets)



$P(V_1) = 0.3$

$P(V_2) = 0.6$

$$P(V_3|V_1 \cap V_2) = 0.05$$
$$P(V_3|\neg V_1 \cap \neg V_2) = 0.2$$
$$P(V_3|V_1 \cap \neg V_2) = 0.1$$
$$P(V_3|\neg V_1 \cap V_2) = 0.1$$

$$P(V_4|V_2) = 0.3$$
$$P(V_4|\neg V_2) = 0.6$$

$$P(V_5|V_3) = 0.3$$
$$P(V_5|\neg V_3) = 0.8$$