Lecture 3: 21 January, 2025

Madhavan Mukund

https://www.cmi.ac.in/~madhavan

Data Mining and Machine Learning
January–April 2024

Market-basket analysis

- Set of items $I = \{i_1, i_2, \ldots, i_N\}$
- A transaction is a set $t \subseteq I$ of items, set of transactions $T = \{t_1, t_2, \ldots, t_M\}$

Market-basket analysis

- Set of items $I = \{i_1, i_2, \ldots, i_N\}$
- A transaction is a set $t \subseteq I$ of items, set of transactions $T = \{t_1, t_2, \ldots, t_M\}$
- Identify association rules $X \to Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - If $X \subseteq t_j$ then it is likely that $Y \subseteq t_j$

Market-basket analysis

- Set of items $I = \{i_1, i_2, \ldots, i_N\}$
- A transaction is a set $t \subseteq I$ of items, set of transactions $T = \{t_1, t_2, \ldots, t_M\}$
- Identify association rules $X \to Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - If $X \subseteq t_j$ then it is likely that $Y \subseteq t_j$
- How frequently does $X \subseteq t_j$ imply $Y \subseteq t_j$?
  - Want $\dfrac{(X \cup Y).count}{X.count} \geq \chi$ (Confidence)

Market-basket analysis

- Set of items $I = \{i_1, i_2, \ldots, i_N\}$
- A transaction is a set $t \subseteq I$ of items, set of transactions $T = \{t_1, t_2, \ldots, t_M\}$
- Identify association rules $X \to Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - If $X \subseteq t_j$ then it is likely that $Y \subseteq t_j$
- How frequently does $X \subseteq t_j$ imply $Y \subseteq t_j$?
  - Want $\dfrac{(X \cup Y).count}{X.count} \geq \chi$ (Confidence)
- How significant is this pattern overall?
  - Want $\dfrac{(X \cup Y).count}{M} \geq \sigma$ (support)

Market-basket analysis

- Set of items $I = \{i_1, i_2, \ldots, i_N\}$
- A transaction is a set $t \subseteq I$ of items, set of transactions $T = \{t_1, t_2, \ldots, t_M\}$
- Identify association rules $X \to Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - If $X \subseteq t_j$ then it is likely that $Y \subseteq t_j$
- How frequently does $X \subseteq t_j$ imply $Y \subseteq t_j$?
  - Want $\dfrac{(X \cup Y).count}{X.count} \geq \chi$ (Confidence)

$X \cup Y = Z$

$Z.count \geq \sigma . M$

- How significant is this pattern overall?
  - Want $\dfrac{(X \cup Y).count}{M} \geq \sigma$ (support)
- Given sets of items I and transactions T, with confidence χ and support σ, find all valid association rules $X \to Y$

Apriori

Apriori observation

If Z is not a frequent itemset, no superset $Y \supseteq Z$ can be frequent

$\{x, y\}$
is

- For any frequent pair $\{x, y\}$, both $\{x\}$ and $\{y\}$ must be frequent
- Build frequent itemsets bottom up, size 1,2,...

Apriori

Apriori observation

If Z is not a frequent itemset, no superset $Y \supseteq Z$ can be frequent

- For any frequent pair $\{x, y\}$, both $\{x\}$ and $\{y\}$ must be frequent
- Build frequent itemsets bottom up, size 1,2,. . .
- F_i : frequent itemsets of size i — Level i
- F_1: Scan T, maintain a counter for each $x \in I$
- C_k = subsets of size k, every $(k-1)$-subset is in F_{k-1}
- F_k: Scan T, maintain a counter for each $X \in C_k$

Z.count

$\geq \sigma \cdot M$

$X \rightarrow Y$

Association rules

- Given sets of items I and transactions T, with confidence χ and support σ, find all valid association rules $X \rightarrow Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - $\dfrac{(X \cup Y).count}{M} \geq \sigma$

Association rules

- Given sets of items I and transactions T, with confidence χ and support σ, find all valid association rules $X \to Y$
  - $X, Y \subseteq I$, $X \cap Y = \emptyset$
  - $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - $\dfrac{(X \cup Y).count}{M} \geq \sigma$
- For a rule $X \to Y$ to be valid, $X \cup Y$ should be a frequent itemset
- Apriori algorithm finds all $Z \subseteq I$ such that $Z.\text{count} \geq \sigma \cdot M$

Association rules

Naïve strategy

- For every frequent itemset Z
  - Enumerate all pairs $X, Y \subseteq Z$, $X \cap Y = \emptyset$
  - Check $\dfrac{(X \cup Y).count}{X.count} \geq \chi$

Association rules

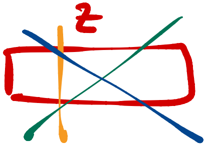
Naïve strategy

- For every frequent itemset Z
  - Enumerate all pairs $X, Y \subseteq Z$, $X \cap Y = \emptyset$
  - Check $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
- Can we do better?

Association rules

Naïve strategy

- For every frequent itemset Z
  - Enumerate all pairs $X, Y \subseteq Z$, $X \cap Y = \emptyset$
  - Check $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
- Can we do better?
- Sufficient to check all partitions of Z
  - If $X, Y \subseteq Z$, $X \cup Y$ is also a frequent itemset



Association rules

- Sufficient to check all partitions of Z
- Suppose $Z = X \uplus Y$, $X \to Y$ is a valid rule and $y \in Y$
- What about $(X \cup \{y\}) \to Y \setminus \{y\}$?

$$\begin{tikzcd} X \arrow[r] & Y \arrow[l, bend right, "y"'] \end{tikzcd}$$

$??\; X \cup \{y\} \to Y \setminus \{y\}$

Association rules

- Sufficient to check all partitions of Z

disjoint union i.e $X \cap Y = \phi$

$X \cup Y = Z$

- Suppose $Z = X \uplus Y$, $X \to Y$ is a valid rule and $y \in Y$
- What about $(X \cup \{y\}) \to Y \setminus \{y\}$?
  - Know $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - Check $\dfrac{(X \cup Y).count}{\underline{\underline{(X \cup \{y\}).count}}} \geq \chi$

at most as frequent as X

Association rules

- Sufficient to check all partitions of Z
- Suppose $Z = X \uplus Y$, $X \to Y$ is a valid rule and $y \in Y$
- What about $(X \cup \{y\}) \to Y \setminus \{y\}$?
  - Know $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - Check $\dfrac{(X \cup Y).count}{(X \cup \{y\}).count} \geq \chi$
  - X.count $\geq (X \cup \{y\})$.count, always
  - Second fraction has smaller denominator, so $(X \cup \{y\}) \to Y \setminus \{y\}$ is also a valid rule

$X \subseteq X \cup \{y\}$

Every time I see $X \cup \{y\}$ I also see X

Association rules

- Sufficient to check all partitions of Z
- Suppose $Z = X \uplus Y$, $X \rightarrow Y$ is a valid rule and $y \in Y$
- What about $(X \cup \{y\}) \rightarrow Y \setminus \{y\}$?
  - Know $\dfrac{(X \cup Y).count}{X.count} \geq \chi$
  - Check $\dfrac{(X \cup Y).count}{(X \cup \{y\}).count} \geq \chi$
  - $X.\text{count} \geq (X \cup \{y\}).\text{count}$, always
  - Second fraction has smaller denominator, so $(X \cup \{y\}) \rightarrow Y \setminus \{y\}$ is also a valid rule

Observation: Can use apriori principle again!

Apriori for association rules

- If $X \to Y$ is a valid rule, and $y \in Y$,
  $(X \cup \{y\}) \to Y \setminus \{y\}$ must also be a valid rule
- If $X \to Y$ is not a valid rule, and $x \in X$,
  $(X \setminus \{x\}) \to Y \cup \{x\}$ cannot be a valid rule

Apriori for association rules

- If $X \to Y$ is a valid rule, and $y \in Y$, $(X \cup \{y\}) \to Y \setminus \{y\}$ must also be a valid rule
- If $X \to Y$ is not a valid rule, and $x \in X$, $(X \setminus \{x\}) \to Y \cup \{x\}$ cannot be a valid rule
- Start by checking rules with single element on the right
  - $Z \setminus z \to \{z\}$
- For $X \to \{x, y\}$ to be a valid rule, both $(X \cup \{x\}) \to \{y\}$ and $(X \cup \{y\}) \to \{x\}$ must be valid
- Explore partitions of each frequent itemset "level by level"

Other questions

1 Items are not of same status

Variable thresholds

2 Time

Association rules for classification

- Classify documents by topic
- Consider the table on the right

Words in document	**Topic**
student, teach, school	Education
student, school	Education
teach, school, city, game	Education
cricket, football	Sports
football, player, spectator	Sports
cricket, coach, game, team	Sports
football, team, city, game	Sports

Association rules for classification

- Classify documents by topic
- Consider the table on the right
- Items are regular words and topics
- Documents are transactions — set of words and one topic

X → Y

Words in document	**Topic**
student, teach, school	Education
student, school	Education
teach, school, city, game	Education
cricket, football	Sports
football, player, spectator	Sports
cricket, coach, game, team	Sports
football, team, city, game	Sports

Association rules for classification

- Classify documents by topic
- Consider the table on the right
- Items are regular words and topics
- Documents are transactions — set of words and one topic
- Look for association rules of a special form
  - $\{\text{student, school}\} \to \{\text{Education}\}$
  - $\{\text{game, team}\} \to \{\text{Sports}\}$

Words in document	**Topic**
student, teach, school	Education
student, school	Education
teach, school, city, game	Education
cricket, football	Sports
football, player, spectator	Sports
cricket, coach, game, team	Sports
football, team, city, game	Sports

Association rules for classification

- Classify documents by topic
- Consider the table on the right
- Items are regular words and topics
- Documents are transactions — set of words and one topic
- Look for association rules of a special form
  - $\{$student, school$\} \rightarrow \{$Education$\}$
  - $\{$game, team$\} \rightarrow \{$Sports$\}$
- Right hand side always a single topic
- Class Association Rules

Words in document	Topic
student, teach, school	Education
student, school	Education
teach, school, city, game	Education
cricket, football	Sports
football, player, spectator	Sports
cricket, coach, game, team	Sports
football, team, city, game	Sports

Summary

- Market-basket analysis searches for correlated items across transactions
- Formalized as association rules
- Apriori principle helps us to efficiently
  - identify frequent itemsets, and
  - split these itemsets into valid rules
- Class association rules — simple supervised learning model

Supervised learning

- A set of items
  - Each item is characterized by attributes $(a_1, a_2, \ldots, a_k)$
  - Each item is assigned a class or category c

a1	a2	a3 --	ak	c

- Given a set of examples, predict c for a new item with attributes $(a'_1, a'_2, \ldots, a'_k)$

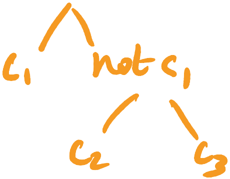
Supervised learning

- A set of items
  - Each item is characterized by attributes $(a_1, a_2, \ldots, a_k)$
  - Each item is assigned a class or category c
- Given a set of examples, predict c for a new item with attributes $(a'_1, a'_2, \ldots, a'_k)$
- Examples provided are called training data
- Aim is to learn a mathematical model that generalizes the training data
  - Model built from training data should extend to previously unseen inputs

Supervised learning

- A set of items
  - Each item is characterized by attributes $(a_1, a_2, \ldots, a_k)$
  - Each item is assigned a class or category c
- Given a set of examples, predict c for a new item with attributes $(a'_1, a'_2, \ldots, a'_k)$
- Examples provided are called training data
- Aim is to learn a mathematical model that generalizes the training data
  - Model built from training data should extend to previously unseen inputs
- Classification problem
  - Usually assumed to binary — two classes

c_1
c_2
c_3



Example: Loan application data set

ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	No
2	young	false	false	good	No
3	young	true	false	good	Yes
4	young	true	true	fair	Yes
5	young	false	false	fair	No
6	middle	false	false	fair	No
7	middle	false	false	good	No
8	middle	true	true	good	Yes
9	middle	false	true	excellent	Yes
10	middle	false	true	excellent	Yes
11	old	false	true	excellent	Yes
12	old	false	true	good	Yes
13	old	true	false	good	Yes
14	old	true	false	excellent	Yes
15	old	false	false	fair	No

9/15

Yes

Basic assumptions

Fundamental assumption of machine learning

- Distribution of training examples is identical to distribution of unseen data

Basic assumptions

Fundamental assumption of machine learning

- Distribution of training examples is identical to distribution of unseen data

What does it mean to learn from the data?

- Build a model that does better than random guessing
  - In the loan data set, always saying Yes would be correct about $9/15$ of the time
- Performance should ideally improve with more training data

Basic assumptions

Fundamental assumption of machine learning

- Distribution of training examples is identical to distribution of unseen data

What does it mean to learn from the data?

- Build a model that does better than random guessing
  - In the loan data set, always saying Yes would be correct about $9/15$ of the time
- Performance should ideally improve with more training data

How do we evaluate the performance of a model?

- Model is optimized for the training data. How well does it work for unseen data?
- Don't know the correct answers in advance to compare — different from normal software verification

The road ahead

Many different models

- Decision trees
- Probabilistic models — naïve Bayes classifiers
- Models based on geometric separators
  - Support vector machines (SVM)
  - Neural networks

The road ahead

Many different models

- Decision trees
- Probabilistic models — naïve Bayes classifiers
- Models based on geometric separators
  - Support vector machines (SVM)
  - Neural networks

With 6 parameters I can fit an elephant.

J. von Neumann

Important issues related to supervised learning

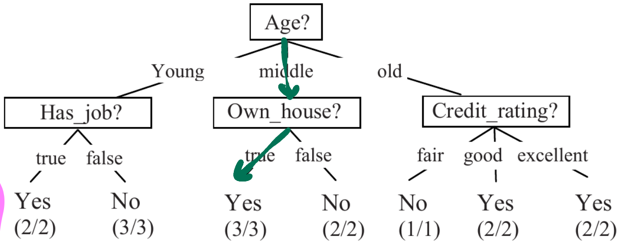
- Evaluating models
- Ensuring that models generalize well to unseen data
  - A theoretical framework to provide some guarantees
- Strategies to deal with the training data bottleneck

Decision trees

20 questions

Questions are adaptive

Which question to ask, when?

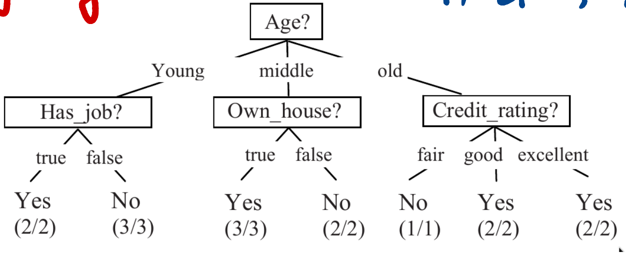


ID	Age	Has_job	Own_house	Credit_rating	Class
1	young	false	false	fair	**No**
2	young	false	false	good	**No**
3	young	true	false	good	**Yes**
4	young	true	true	fair	**Yes**
5	young	false	false	fair	**No**
6	middle	false	false	fair	**No**
7	middle	false	false	good	**No**
8	middle	true	true	good	**Yes**
9	middle	false	true	excellent	**Yes**
10	middle	false	true	excellent	**Yes**
11	old	false	true	excellent	**Yes**
12	old	false	true	good	**Yes**
13	old	true	false	good	**Yes**
14	old	true	false	excellent	**Yes**
15	old	false	false	fair	**No**

Decision tree algorithm

Choose an attribute to query · A_j

$A_1, A_2, \ldots, A_k$



A_j?

$A_1 \ldots A_{j-1}$, $A_{j+1} \ldots A_k$

Pick A_ℓ

Pick A_m

Pick A_i

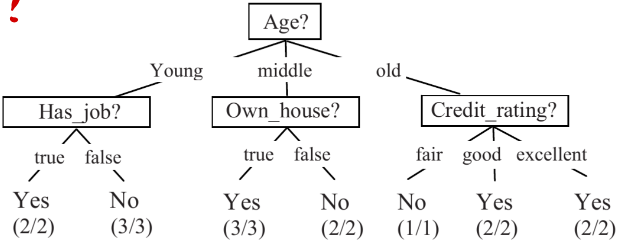
- → Each step selects a subset of rows
- → Stop when subset has uniform ans.
- → Run out of questions (k) → Majority

Comparing decision trees

Which is the better tree?

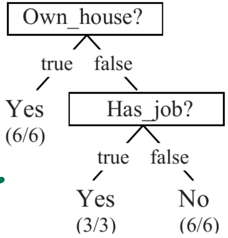
Fewer parameters is better

Occam's Razor



Smaller how?

- No of questions overall size
- No of questions in each path height

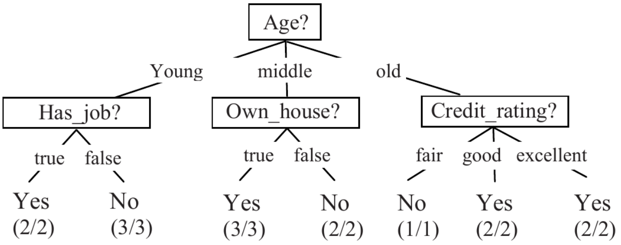


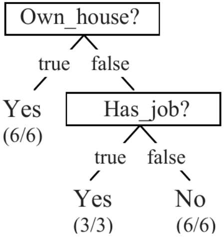
Comparing decision trees

Bad news

For any defn of smallest,
computing the smallest
tree is infeasible

NP-complete





Greedy heuristic — impurity

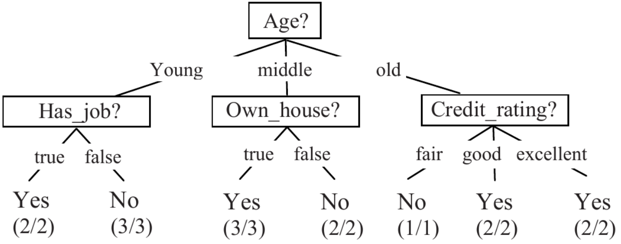
Goal is a subset of rows with same category

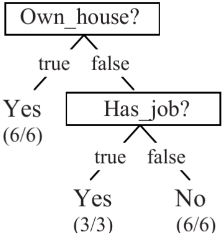
Pure subset

Reduce "impurity"

Max impurity – 50% each way

Define impurity as fraction of minority answer   0 – 0.5

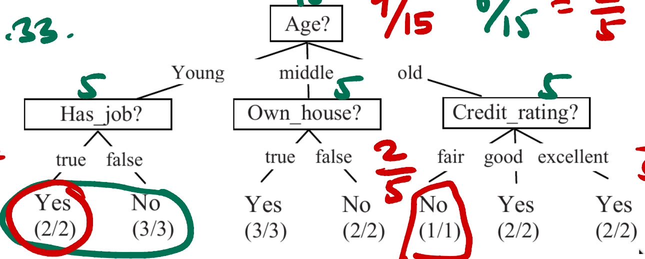




Greedy heuristic — impurity

$$\frac{5}{15} = \frac{1}{3} \quad 0.33.$$

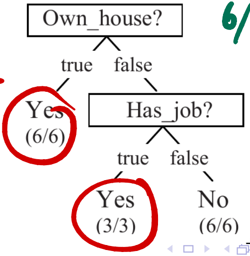
$$\frac{1}{3}\cdot\frac{2}{5} + \frac{1}{3}\cdot\frac{2}{5}\cdot\frac{1}{3}\,\frac{1}{5}$$



15 5 5 5 $\frac{2}{5}$ $\frac{2}{5}$ $\frac{1}{5}$

$9/15 \quad 6/15 = \frac{2}{5} \quad 0.4$

$$\frac{2}{5}\cdot 0 + \frac{3}{5}\cdot\frac{1}{3} = \frac{1}{5} = 0.2$$



0

$6/15 = 0.4$

$1/3$

Greedy heuristic — impurity

Impurity = "Misclassification ratio"