Roy Chan
301202770
chanroyc@sfu.ca

CMPT 459 Fall 2017
Data Mining
Martin Ester
TA: Zhilin Zhang

Assignment 4

Assignment 4.1

a) Let A_i be the base algorithm using nearest neighbours. The nearest neighbour identifier identifies the *i* nearest neighbours, where *i* is the number of nearest neighbours considered, with a decision set of *i*-nearest neighbours considered for classification. The decision rule will be to choose the majority rule within the decision set.

b) Let the training datasets be chosen via the method used in the stacking method. Two levels of classification will be used with the training dataset *j-th* training dataset D_J .

Choose a training dataset D_J from *D*, then split dataset D_J into two subsets D_A and D_B . Subsequent datasets chosen from D are sampled independently of each other.

Train the (first level) ensemble algorithm (nearest neighbour) on training dataset D_A to learn models $M_1, ..., M_k$, then apply the models to subset D_B to create *k* new features, where *i-th feature* is the class label predicted by model M_I .

Train the (second level) classifier on subset D_B to train model M that combines the first level classifiers.

c) Random variations in the choice of training data means that different models will be learnt, even when using the same classification algorithms. Different models may disagree on some classifications. In order to reduce bias and variance, the base algorithm used and the way the dataset is created and approached can affect the results significantly. The datasets chosen are sampled independently of each other, so that each training dataset is unique, as mentioned above. By using the stacking method and splitting the dataset into even smaller subsets, we can reduce bias, variation and promote independence between classifiers.

Roy Chan
301202770
chanroyc@sfu.ca

Assignment 4.2

a) Minimum Support: 50%

M:100%	MO:60%	MOT:60%
O:60%	MP:60%	MPQ:60%
P:60%	MQ:80%	MQT:60%
Q:80%	MT:80%	
T:80%	OT:60%	
	PQ:60%	
	QT:60%	

b)

Closed: M, MQ, MT, MOT, MPQ, MQT

Maximal: MOT, MPQ, MQT

c)

MOT

MO	->	T	100%
MT	->	O	75%
OT	->	M	100%
O	->	MT	100%
T	->	MO	75%

MPQ

MP	->	Q	100%
MQ	->	P	75%
PQ	->	M	100%
P	->	MQ	100%
Q	->	MP	75%

MQT

MQ	->	T	75%
MT	->	Q	75%
QT	->	M	100%
Q	->	MT	75%
T	->	MQ	75%

Roy Chan
301202770
chanroyc@sfu.ca

d)

MOT

MO	->	T	100%	lift=1.0/0.8=1.25
MT	->	O	75%	lift=0.75/0.6=1.25
OT	->	M	100%	lift=1.0/1.0=1.0
O	->	MT	100%	lift=1.0/0.8=1.25
T	->	MO	75%	lift=0.75/0.6=1.25

MPQ

MP	->	Q	100%	lift=1.0/0.8=1.25
MQ	->	P	75%	lift=0.75/0.6=1.25
PQ	->	M	100%	lift=1.0/1.0=1
P	->	MQ	100%	lift=1.0/0.8=1.25
Q	->	MP	75%	lift=0.75/0.6=1.25

MQT

MQ	->	T	75%	lift=0.75/0.8=0.9375
MT	->	Q	75%	lift=0.75/0.8=0.9375
QT	->	M	100%	lift=1.0/1.0=1.0
Q	->	MT	75%	lift=0.75/0.8=0.9375
T	->	MQ	75%	lift=0.75/0.8=0.9375

Assignment 4.3

a) FDBi is a set of frequent itemsets FDBi in DBi
FDB is a global set of frequent itemsets

If an itemset is globally frequent, then it is frequent in at least one of the local databases

$FDB \subseteq \cup FDB_i$

Assume: If an itemset is globally frequent, then it is not frequent in any of the local databases.

Let T_1 be transaction that contains frequent itemset FDB of items *I*
Let T_2 be transaction that contains frequent itemset FDB_i of items *I*

$\forall\, T_1 \subseteq I,\ T_2 \subseteq I\text{: } T_1 \subseteq T_2 \wedge freq(T_2, DB) \nRightarrow freq(T_1, DB)$

But the relationship $FDB \subseteq \cup FDB_i$ holds then

$\forall\, T_1 \subseteq I,\ T_2 \subseteq I\text{: } T_1 \subseteq T_2 \wedge sup(T_2, DB) \geq sup(T_1, DB)$ *therefore*
$\forall\, T_1 \subseteq I,\ T_2 \subseteq I\text{: } T_1 \subseteq T_2 \wedge freq(T_2, DB) \Rightarrow freq(T_1, DB)$

Roy Chan
301202770
chanroyc@sfu.ca

Which contradicts the assumption, therefore an an itemset can only be globally frequent if it is frequent in at least one of the local databases.

b)

c)