MDL continued + Bayes Approach

AW

Lecture Overview

MDL Paradigm for Recursive DTree Algorithm

Minimum Description Length Paradigm

Given a hypothesis class $\mathscr{H}$ that is countable union of signleton classes each of which are agnostic PAC learnable and such that members of $\mathscr{H}$ are described by a prefix-free language L. Then for a training set $S \sim D^m$ and a confidence parameter $0 < \delta < 1$ the best classifier is

$$g \in \underset{h \in \mathscr{H}}{\arg\min} \left[L_S(h) + \sqrt{\frac{\log(2/\delta) + |h|}{2m}} \right]$$

where $|h|$ is the encoding length of classifier h.

The inductive tree-learning algorithm gives only approximate trees so we won't be able to find 'best' g. But if we compare two decision trees T_1 and T_2 w.r.t. MDL then the better tree is the one that has smaller value of

$$SE(T_i) = \left[L_S(h) + \sqrt{\frac{\log(2/\delta) + |h|}{2m}} \right]$$

.

When we expand a node of the tree T_1 to obtain T_2 compare the $SE(T_1)$ to $SE(T_2)$. If the latter is bigger DO NOT EXPAND!

MDL Paradigm for Recursive DTree Algorithm

Minimum Description Length Paradigm

Given a hypothesis class $\mathscr{H}$ that is countable union of signleton classes each of which are agnostic PAC learnable and such that members of $\mathscr{H}$ are described by a prefix-free language L. Then for a training set $S \sim D^m$ and a confidence parameter $0 < \delta < 1$ the best classifier is

$$g \in \arg\min_{h \in \mathscr{H}} \left[L_S(h) + \sqrt{\frac{\log(2/\delta) + |h|}{2m}} \right]$$

where $|h|$ is the encoding length of classifier h.

The inductive tree-learning algorithm gives only approximate trees so we won't be able to find 'best' g. But if we compare two decision trees T_1 and T_2 w.r.t. MDL then the better tree is the one that has smaller value of

$$SE(T_i) = \left[L_S(h) + \sqrt{\frac{\log(2/\delta) + |h|}{2m}} \right]$$

.

To implement the MDL paradigm we need prefix free encoding of DTrees

Prefix-free Representation of DTrees

Fix size-first description of DT for binary classification using γ-encoding that is known to be prefix free: all features are numbered from 2 to $k+1$ and 'leaf' designation is treated as a feature #1 that has class as 'domain values' (i.e. $\{1, 2\}$).

- number of features in the tree
- size of the sample
- maximum branching degree
- number of classes
- the following sequence of nodes is given in BFS order of walking the decision tree:
  - number of children of the node (since there are at least 2 childre 1 stands for no children),
  - the number of a feature used in the split,
  - domain value used in a split as the number of example in use.

Simple Gamma encoding

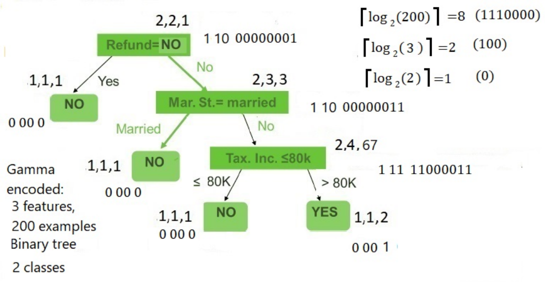
γ-encoding:

- Convert number into binary drop; leading 1. ex: $80 \rightarrow 1010000 \rightarrow 010000$
- Drop leading 1 (no number starts with 0) so we can always add 1 when decoding)
- Compute number of digits and put corresponding number of 1's as the beginning of the code, ex $6 \rightarrow 111111$
- add to the beginning of the code 0 and then the result of step 1: $80 \rightarrow 111111\,0\,010000$

Example of Encoding

- Features: leaf - #1 (1) refund - #2 $(= 10)$,Marital status - #3 $(= 11)$, taxable income - #4 $(= 100)$
- Domain values in use: no (1) $= 1$, yes (2) $= 10$, married (1) $(= 1)$, not married (2) $(= 11)$, 80 $(= 1010000)$
- number of nodes is 7 $(= 111)$

The encoding is:



encoding: 100 1110000 0 0 1 10 00000001 0000 1 10 00000011 0000 1 11 11000011 0000 0001
length: 11*3+4*4+7+3+1=60 bits
gamma length: 2*(60-1)+1=119 bits

MDL in C4.5/J48

```
library(RWeka); library(sets); library(mlbench)
data(BreastCancer)
BC<-BreastCancer
rm(BreastCancer)# no meaningful work
      #Breastcancer is too long to write each time
BC$Id <- NULL
     # remove id column that confuses learning
set.seed(2) #set random seed r
ind <- sample(2, nrow(BC),
               replace = TRUE, prob=c(2/3, 1/3))
#sample from values [1:2] with replacement with
#probbabilities 2/3 for Tr Set and 1/3 for Test Set
C45T <- J48(Class ~ ., data = BC[ind==1,],
      control = Weka_control(U =TRUE, M = 5)); C45T
      #unpprunned tree, min leaf size 5
plot(C45T)  # plot the tree
```

MDL in C4.5/J48

```
pred.C45T <- predict(C45T,newdata=BC[ind==2,-11])
      # classify TestSet
table(BC[ind==2,]$Class, pred.C45T,
       dnn = c("Actual class", "Predicted class"))
acc.C45T <- 100*sum(pred.C45T==BC[ind==2,]$Class)/
        dim(BC[ind==2,])[1]; acc.C45T
C45T1 <- J48(Class ~ ., data = BC[ind==1,],
   control = Weka_control(C=0.1,S =FALSE, M = 5));
 C45T1 #prunned tree: confidence 0.1, Tree Raising
plot(C45T1)
pred.C45T1 <- predict(C45T1,newdata=BC[ind==2,-11])
 # classify TC
table(BC[ind==2,]$Class, pred.C45T1,
      dnn = c("Actual class", "Predicted class"))
acc.C45T1 <- 100*sum(pred.C45T1==BC[ind==2,]$Class)/
         dim(BC[ind==2,])[1];acc.C45T1
```

MDL in C4.5/J48

```
C45T2 <- J48(Class ~ ., data = BC[ind==1,],
  control = Weka_control(S =TRUE,J=FALSE,M = 5))
  C45T2  #prunned tree no statistical tree
         #raising, but using MDL
plot(C45T2)
pred.C45T2 <- predict(C45T2,newdata=BC[ind==2,-11])
           # classify TC
table(BC[ind==2,]$Class, pred.C45T2,
           dnn = c("Actual class", "Predicted class"))
acc.C45T2 <- 100*sum(pred.C45T2==BC[ind==2,]$Class)/
dim(BC[ind==2,])[1];acc.C45T2
```

Lecture Overview

Conditional Probability

Given probability $\Pr(A), \Pr(B)$ of events A and B and probability $\Pr(A \wedge B)$ of a joint even A and B happening at the same time

$$\Pr(A|B) \stackrel{def}{=} \frac{\Pr(A \wedge B)}{\Pr(B)}$$

is a conditional probability of A given B.

Symmetrically, conditional probability of B given A is

$$\Pr(B|A) \stackrel{def}{=} \frac{\Pr(A \wedge B)}{\Pr(A)}$$

and

$$\Pr(A \wedge B) = \Pr(B|A) \cdot \Pr(A) = \Pr(A|B) \cdot \Pr(B)$$

so

$$\Pr(A|B) = \frac{\Pr(B|A) \times \Pr(A)}{\Pr(B)}$$

Example

Known facts:

- Meningitis causes stiff neck 50% of the time
- Only one in 50,000 people who seek help are diagnosed with meningitis
- It is known that on the average 1 in every 20 patients has stiff neck

If a patient walks into doctors office complaining of stiff neck, whats the probability he/she has meningitis?

Example

Known facts:

- Meningitis causes stiff neck 50% of the time
- Only one in 50,000 people who seek help are diagnosed with meningitis
- It is known that on the average 1 in every 20 patients has stiff neck

If a patient walks into doctors office complaining of stiff neck, whats the probability he/she has meningitis?

- $\Pr(stiff\ neck|meningitis) = 0.5$ - likelihood of stiff neck (evidence) given meningitis (hypothesis)
- $\Pr(meningitis) = 2 \cdot 10^{-6}$ - prior probability of meningitis (hypothesis)
- $\Pr(stiff\ neck) = 0.05$ - probability of evidence
- $\Pr(meningitis|stiff\ neck) =?$ - posterior probability of evidence

$$\Pr(meningitis|stiff\ neck) \stackrel{def}{=} \frac{\Pr(stiff\ neck \wedge meningitis)}{\Pr(stiff\ neck)} = \frac{\Pr(stiff\ neck|meningitis) \cdot \Pr(meningitis)}{\Pr(stiff\ neck)} = \frac{0.5 \cdot 2 \cdot 10^{-6}}{0.05}$$

One More Example

Suppose you wake up in the morning and you see that the grass is wet. Assuming that you know that

- Probability that grass is wet in the morning because there was rain at night is 0.7 (i.e. $\Pr(W|R) = 0.7$)
- Probability that the grass in wet in the morning, but there was no rain at night is 0.4 (i.e. $\Pr(W|\neg R) = 0.4$)

What is the probability that there was rain at night, given that probability of rain was forecasted last evening to be 0.8 (i.e. $\Pr(R) = 0.8$)?

One More Example

Suppose you wake up in the morning and you see that the grass is wet. Assuming that you know that

- Probability that grass is wet in the morning because there was rain at night is 0.7 (i.e. $\Pr(W|R) = 0.7$)
- Probability that the grass in wet in the morning, but there was no rain at night is 0.4 (i.e. $\Pr(W|\neg R) = 0.4$)

What is the probability that there was rain at night, given that probability of rain was forecasted last evening to be 0.8 (i.e. $\Pr(R) = 0.8$)? By law of total probability

$$\Pr(\neg W|\neg R) = 1 - \Pr(W|\neg R) = 1 - 0.4 = 0.6$$
$$\Pr(\neg R) = 1 - \Pr(R) = 1 - 0.8 = 0.2$$

Notice that

$$\Pr(R|W) = \frac{\Pr(W|R)\times\Pr(R)}{\Pr(W)} = \frac{\Pr(W|R)\times\Pr(R)}{\Pr(W|R)\Pr(R)+\Pr(W|\neg R)\Pr(\neg R)} = \frac{0.7\times0.8}{0.7\times0.8+0.6\times0.2} = 0.8$$

where second equality is by law of total probability in denominator

Bayes Rule

- From example: evidence can comes as a joint even $(W \wedge R) \cup (W \wedge \neg R)$.
- More generally, let probability space S be formed by a union of some k incompatible events $B_i,\ i \in \{1, \ldots, k\}$
  - In other words $B_i \cap B_j = \emptyset$ and $\bigcup_{i=1}^{k} B_i$ (partitioning).
- Then

$$\Pr(B_i|A) = \frac{\Pr(A|B_i)\times\Pr(B_i)}{\Pr(A)} = \frac{\Pr(A|B_i)\times\Pr(B_i)}{\sum_{j=1}^{k}\Pr(A\wedge B_j)}$$

- From example: probability of joint events $W \wedge R$ and $W \wedge \neg R$ were given with the help of conditional probabilities $\Pr(W|R$ and $Pr(W|\neg R)$ and absolute probabilit
- More generally, if probability of events $A \wedge B_j$ are given with the help of conditional probabilities $Pr(A \wedge B_j)$ and absolute probabilities $\Pr(B_)$ we have Bayes rule

$$\Pr(B_i|A) = \frac{\Pr(A|B_i)\times\Pr(B_i)}{\Pr(A)} = \frac{\Pr(A|B_i)\times\Pr(B_i)}{\sum_{j=1}^{k}\Pr(A\wedge B_j)} = \frac{\Pr(A|B_i)\times\Pr(B_i)}{\sum_{j=1}^{k}\Pr(A|B_j)\cdot\Pr(B_j)}$$

Lecture Overview

Bayes Predictor

- Approach so far was distribution free learning: no assumptions on the underlying distribution over the data
  - As a consequence - discriminative approach in which our goal is not to learn the underlying distribution but rather to learn an accurate predictor
- Change of strategy: generative approach, in which it is assumed that the underlying distribution over the data has a specific parametric form and our goal is to estimate the parameters of the model
  - This task is called parametric density estimation.
- If we succeed in learning the underlying distribution $\mathscr{D}$ over $X \times \{0, 1\}$ accurately, then we can predict by using the Bayes optimal classifier:

$$f_{\mathscr{D}}(x) = \begin{cases} 1 \text{ if } \Pr_{\mathscr{D}}(y = 1|x) \geq 1/2 \\ 0 \text{ otherwise} \end{cases} = \begin{cases} 1 \text{ if } \frac{\Pr_{\mathscr{D}}(y=1|x)}{\Pr_{\mathscr{D}}(y=0|x)} \geq 1 \\ 0 \text{ otherwise} \end{cases}$$

for proof of optimality see here

- Another way to describe Bayes predictor for a data point $\overline{x} \in X$ is $h_{Bayes} = \arg\max_{y \in \{0,1\}} \Pr_{\mathscr{D}}(Y = y|X = \overline{x})$

Reading

SSBD sections 7.1, 7.2, 7.3
You can skip proofs if you are not interested in technicalities

TSK (main texbook) section 3.5.2