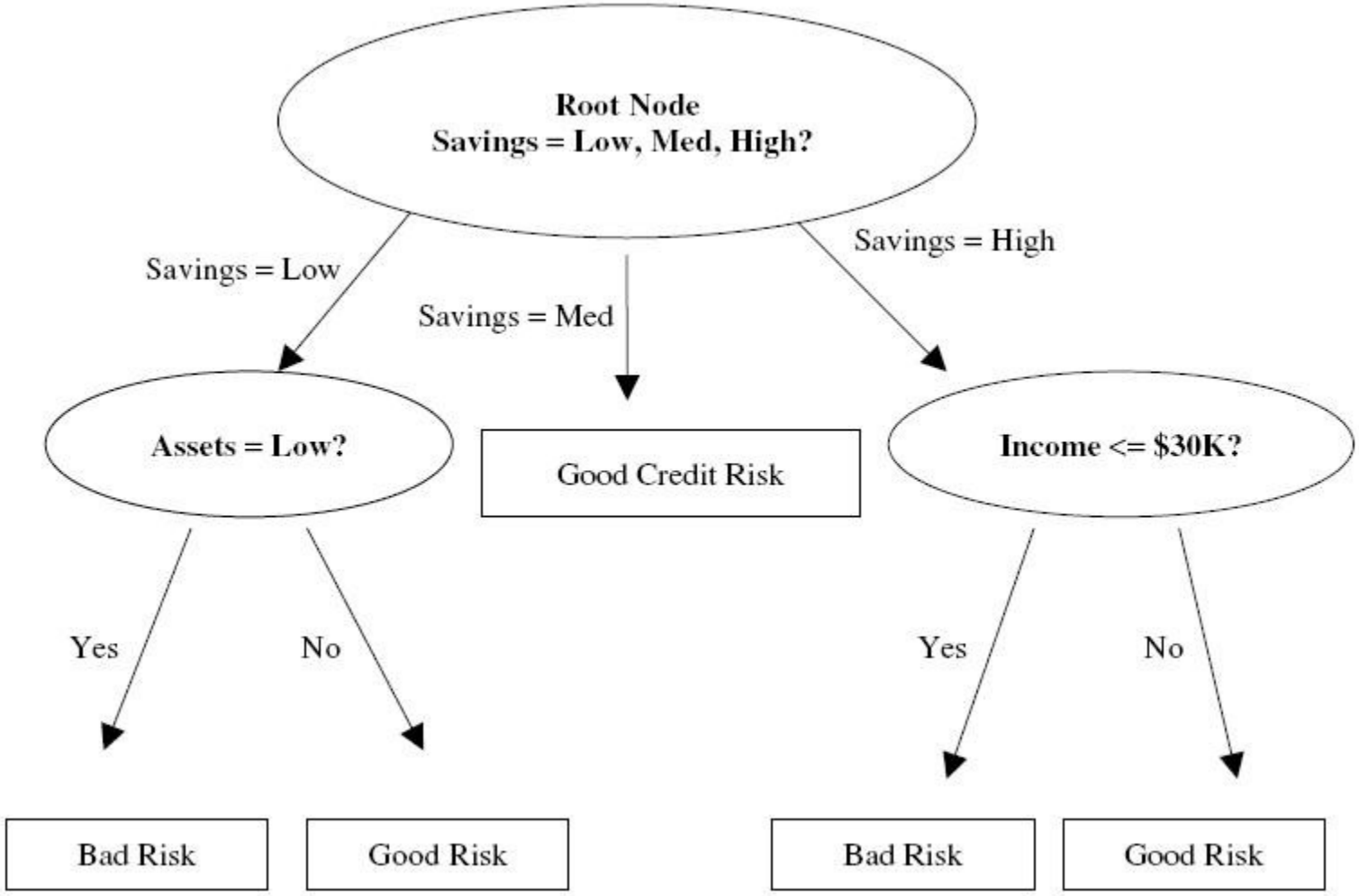
Stabla odlučivanja

Bojan Furlan

УНИВЕРЗИТЕТ У БЕОГРАДУ

ЕЛЕКТРОТЕХНИЧКИ ФАКУЛТЕТ

Stabla odlučivanja - cilj



Stabla odlučivanja – svojstva

- Kreiranje stabla odlučivanja (*Decision Trees* – DT) :
  - Ručno – na osnovu ekspertskog znanja
  - Automatsko – na osnovu trening skupa (ML)
- Zahtevi:
  - Trening skup sa vrednostima ciljnog atributa (nadgledano učenje)
  - Ciljni atribut mora biti diskretan (ili diskretizovan)

Stabla odlučivanja – svojstva

- Pitanje#1: Koji atribut uzeti za grananje?
  - Onaj koji deli početni čvor na što “čistije” potomke
    - U listovima treba da bude zastupljeno što više instanci iste klase
  - Ovo obezbeđuje klasifikaciju sa najvećom pouzdanošću.
  - Npr. za prethodni primer atribut ušteđevina (*savings*) je uzet jer najbolje deli početni trening skup.

Stabla odlučivanja – svojstva

- Pitanje#2: Kada stati sa grananjem?
  I. Kada sve instance u čvoru pripadaju istoj klasi.
  II. Kada su sva grananja iscrpljena.
- Algoritmi:
  - Classification and regression trees (CART) algoritam
  - C4.5 algoritam

CART algoritam

- CART stabla su binarna stabla i svaki čvor odlučivanja ima tačno dve grane.
- CART rekurzivno deli početni skup u podskupove sa istim vrednostima ciljnog atributa (iste klase).

$$\Phi(s|t) = 2P_L P_R \sum_{j=1}^{\#\text{ classes}} |P(j|t_L) - P(j|t_R)|$$

- *Φ(s|t)* je mera "povoljnosti" grananja za kandidat grananje *s* čvora *t*
- t_L i t_R su levi odnosno desni potomak čvora *t*
- Optimalno grananje je ono sa maksimalnom vrednosti *Φ(s|t)* za sva moguća grananja za čvor *t.*

CART algoritam – svojstva

$$\Phi(s|t) = 2P_L P_R \sum_{j=1}^{\#\text{ classes}} |P(j|t_L) - P(j|t_R)|$$

- Zastupljenost levog i desnog potomka: $P_L = |t_L|/|t|$ i $P_R = |t_R|/|t|$
- Zastupljenost klase *j* u levom i desnom potomku:
  - $P(j|t_L) = |t_L == j|/|t_L|$
  - $P(j|t_R) = |t_R == j|/|t_R|$

CART algoritam – svojstva

- Φ(s|t) raste kada obe komponente proizvoda rastu
  $2P_L P_R$ i $\sum_{j=1}^{\#\text{ classes}} |P(j|t_L) - P(j|t_R)|$

1. $2P_L P_R$ ima maksimalnu vrednost kada su oba potomka iste veličine (ista zastupljenost): 0.5*0.5 = 0.25 ili 0.9*0.1= 0.09

2. Q (s|t) = $\sum_{j=1}^{\#\text{ classes}} |P(j|t_L) - P(j|t_R)|$

   - Maksimalna vrednost je kada sve instance čvora potomka su potpuno uniformne (čiste).
   - Teoretski maksimum za Q (s|t) je k,
     gde je k broj klasa koje može uzimati ciljna promenljiva.

CART Primer

Customer	Savings	Assets	Income ($1000s)	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

Trening skup

Candidate Split	Left Child Node, t_L	Right Child Node, t_R
1	$Savings = low$	$Savings \in \{medium, high\}$
2	$Savings = medium$	$Savings \in \{low, high\}$
3	$Savings = high$	$Savings \in \{low, medium\}$
4	$Assets = low$	$Assets \in \{medium, high\}$
5	$Assets = medium$	$Assets \in \{low, high\}$
6	$Assets = high$	$Assets \in \{low, medium\}$
7	$Income \leq \$25{,}000$	$Income > \$25{,}000$
8	$Income \leq \$50{,}000$	$Income > \$50{,}000$
9	$Income \leq \$75{,}000$	$Income > \$75{,}000$

Kandidat grananja za koreni čvor *t*

CART Primer

- Grananje 1. -> Savings == low (L-true, R-false)
  - Right:1,3,4,6,8
  - Left:2,5,7

- P_R=5/8 = 0.625 P_L=3/8=0.375 -> 2*$P_L P_R$=15/64=0.46875

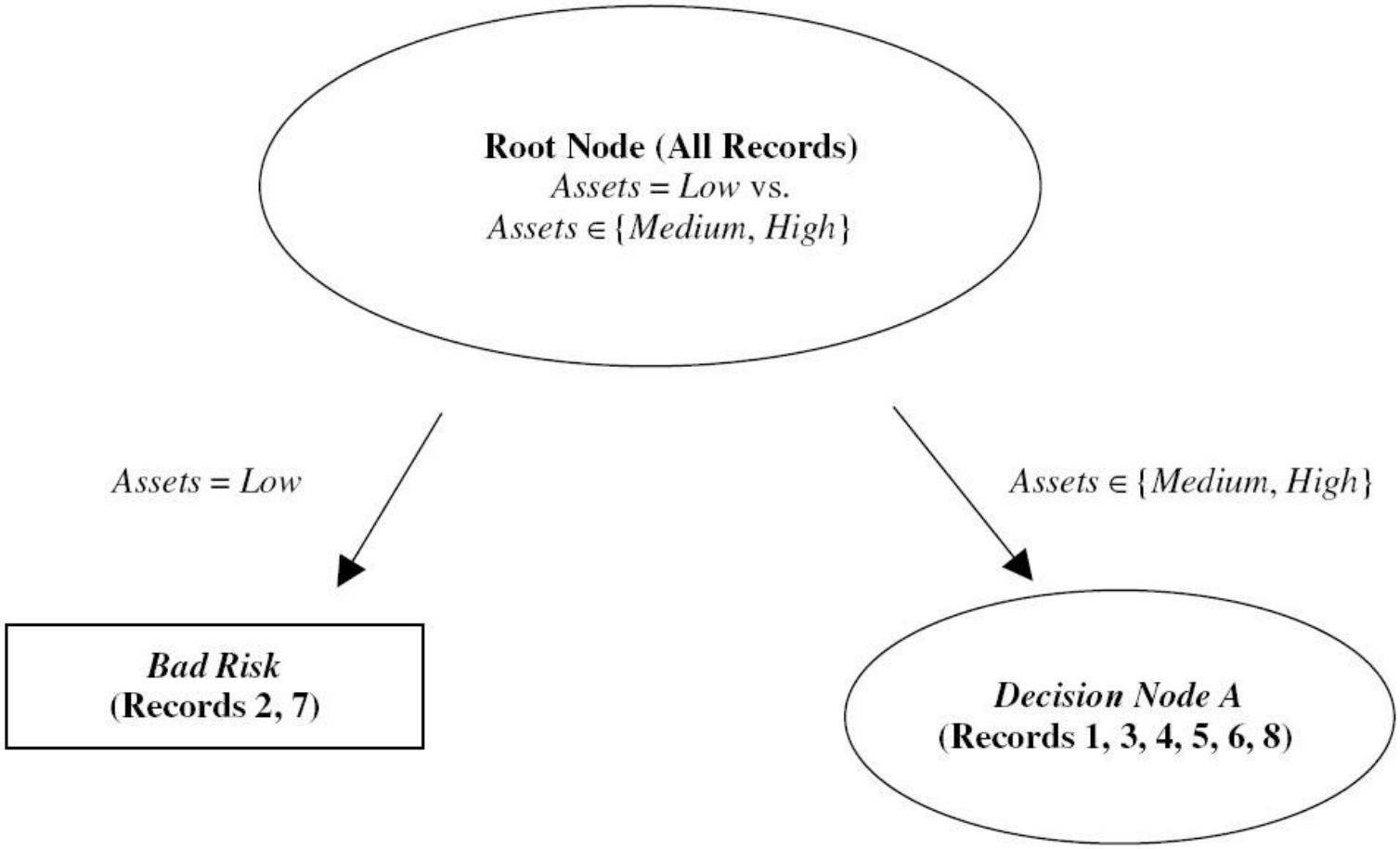
- za j(klasu) = Bad
  - P(Bad|t_R)= 1/5=0.2
  - P(Bad|t_L)= 2/3=0.67
- za j(klasu) = Good
  - P(Good|t_R)= 4/5 = 0.8
  - P(Good|t_L)= 1/3 = 0.33
- Q(s|t)= |0.67-0.2|+|0.8-0.33| = 0.934

CART Primer

Split	P_L	P_R	$P(j\|t_L)$	$P(j\|t_R)$	$2P_L P_R$	$Q(s\|t)$	$\Phi(s\|t)$
1	0.375	0.625	G: .333 B: .667	G: .8 B: .2	0.46875	0.934	0.4378
2	0.375	0.625	G: 1 B: 0	G: 0.4 B: 0.6	0.46875	1.2	0.5625
3	0.25	0.75	G: 0.5 B: 0.5	G: 0.667 B: 0.333	0.375	0.334	0.1253
4	0.25	0.75	G: 0 B: 1	G: 0.833 B: 0.167	0.375	1.667	0.6248
5	0.5	0.5	G: 0.75 B: 0.25	G: 0.5 B: 0.5	0.5	0.5	0.25
6	0.25	0.75	G: 1 B: 0	G: 0.5 B: 0.5	0.375	1	0.375
7	0.375	0.625	G: 0.333 B: 0.667	G: 0.8 B: 0.2	0.46875	0.934	0.4378
8	0.625	0.375	G: 0.4 B: 0.6	G: 1 B: 0	0.46875	1.2	0.5625
9	0.875	0.125	G: 0.571 B: 0.429	G: 1 B: 0	0.21875	0.858	0.1877

Vrednosti $\Phi(s|t)$ za svako kandidat grananje za koreni čvor

CART Primer

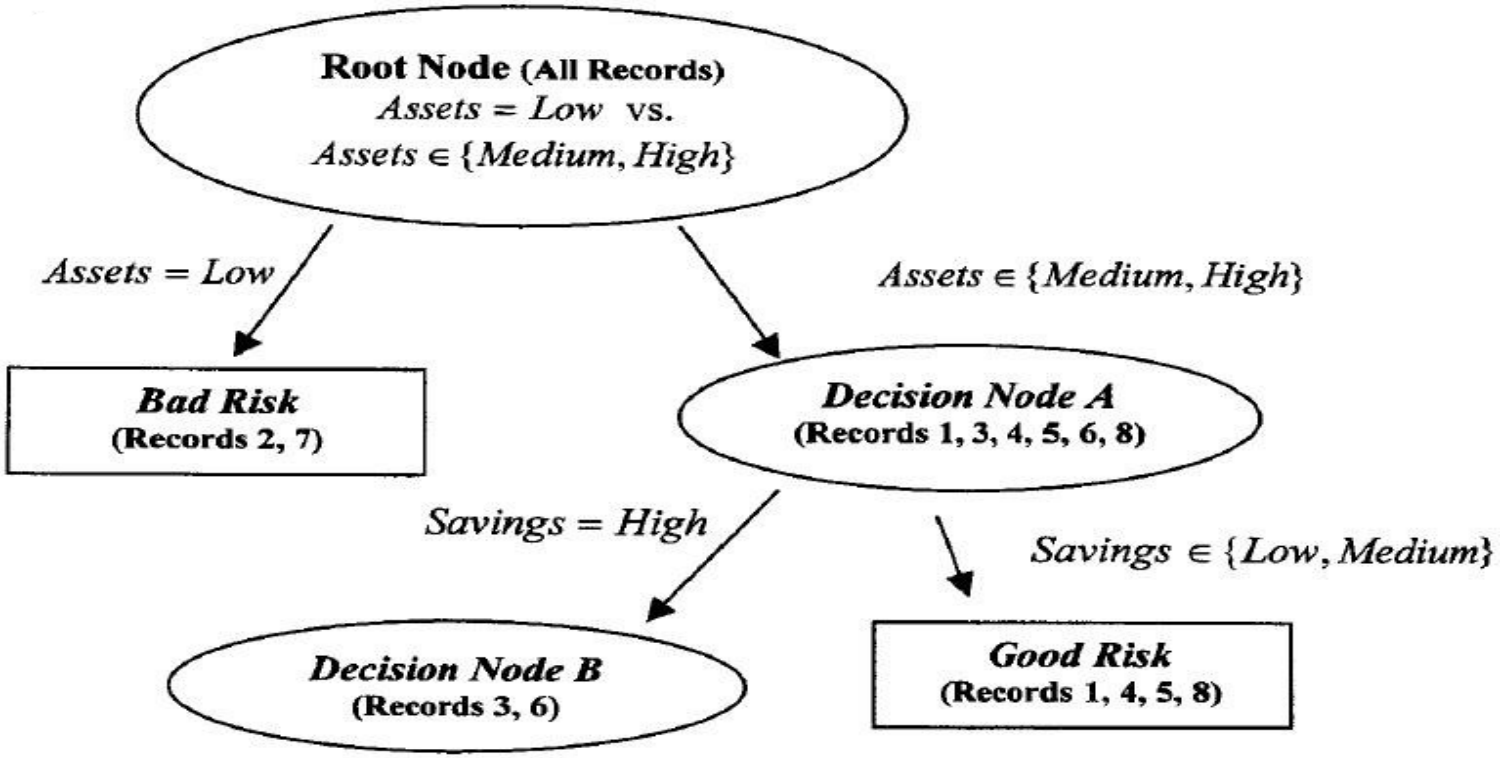

Root Node (All Records)
Assets = Low vs.
Assets ∈ {Medium, High}
Assets = Low
Assets ∈ {Medium, High}
Bad Risk
(Records 2, 7)
Decision Node A
(Records 1, 3, 4, 5, 6, 8)

CART Primer

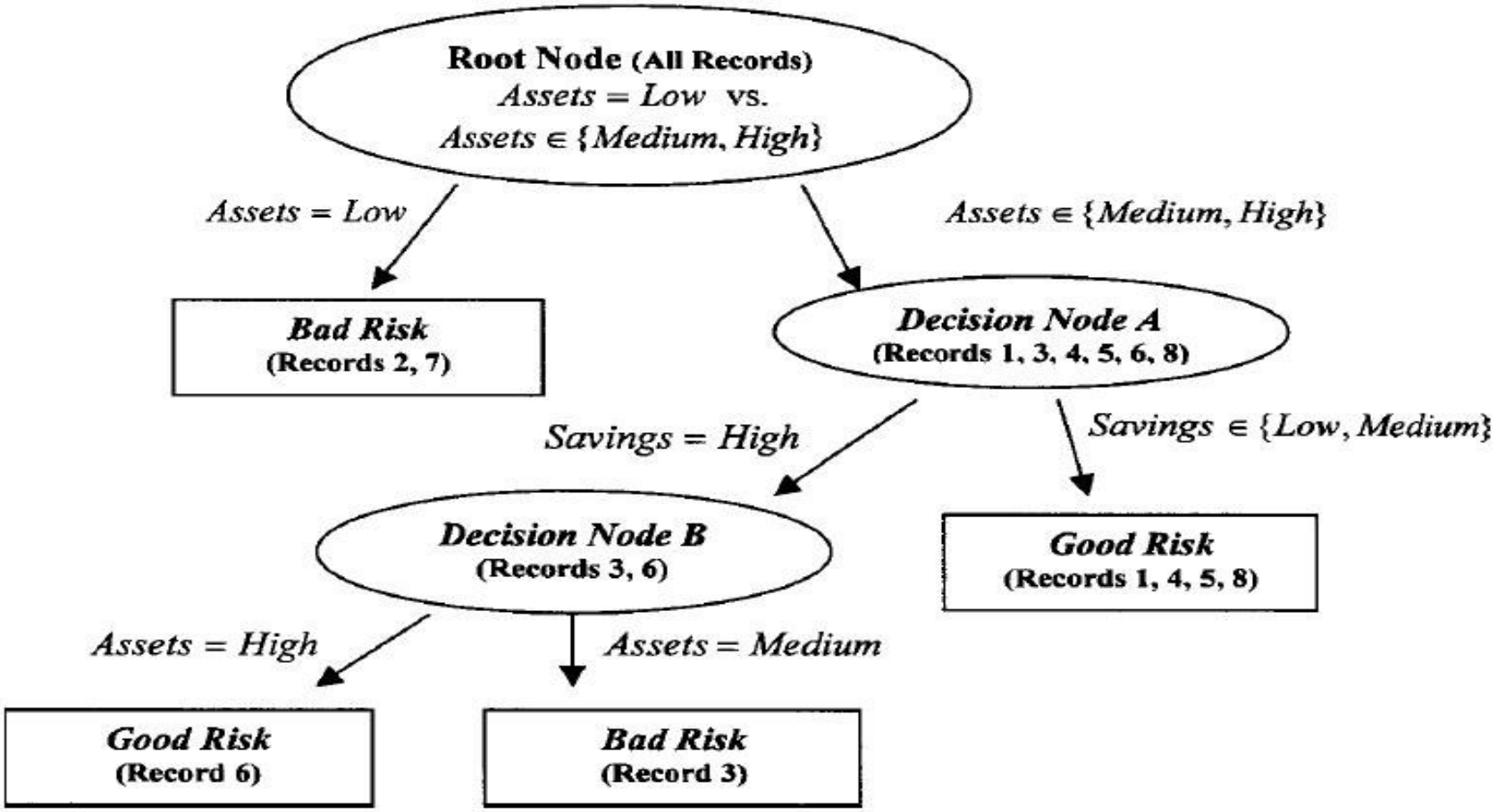
Split	P_L	P_R	$P(j\|t_L)$	$P(j\|t_R)$	$2P_LP_R$	$Q(s\|t)$	$\Phi(s\|t)$
1	0.167	0.833	G: 1 B: 0	G: .8 B: .2	0.2782	0.4	0.1112
2	0.5	0.5	G: 1 B: 0	G: 0.667 B: 0.333	0.5	0.6666	0.3333
3	0.333	0.667	G: 0.5 B: 0.5	G: 1 B: 0	0.4444	1	0.4444
5	0.667	0.333	G: 0.75 B: 0.25	G: 1 B: 0	0.4444	0.5	0.2222
6	0.333	0.667	G: 1 B: 0	G: 0.75 B: 0.25	0.4444	0.5	0.2222
7	0.333	0.667	G: 0.5 B: 0.5	G: 1 B: 0	0.4444	1	0.4444
8	0.5	0.5	G: 0.667 B: 0.333	G: 1 B: 0	0.5	0.6666	0.3333
9	0.167	0.833	G: 0.8 B: 0.2	G: 1 B: 0	0.2782	0.4	0.1112

Vrednosti $\Phi(s|t)$ za svako kandidat grananje za čvor A

CART Primer


Root Node (All Records)
Assets = Low vs.
Assets ∈ {Medium, High}
Assets = Low
Assets ∈ {Medium, High}
Bad Risk
(Records 2, 7)
Decision Node A
(Records 1, 3, 4, 5, 6, 8)
Savings = High
Savings ∈ {Low, Medium}
Decision Node B
(Records 3, 6)
Good Risk
(Records 1, 4, 5, 8)

CART Primer


Root Node (All Records)
Assets = Low vs.
Assets ∈ {Medium, High}
Assets = Low
Assets ∈ {Medium, High}
Bad Risk
(Records 2, 7)
Decision Node A
(Records 1, 3, 4, 5, 6, 8)
Savings = High
Savings ∈ {Low, Medium}
Decision Node B
(Records 3, 6)
Good Risk
(Records 1, 4, 5, 8)
Assets = High
Assets = Medium
Good Risk
(Record 6)
Bad Risk
(Record 3)

Završetak

- Kada se iscrpe sva moguća grananja generisano je “puno” stablo.
- Ovakvo stablo ima najmanji stepen greške na trening skupu, ali često može dovesti do prepodešenosti modela.
- Stoga se stablo “odseca” kako bi se postigla generalnost modela.

Prečišćenost stabla

- Često svi listovi stabla nisu homogeni, što ostavlja određeni stepen greške prilikom klasifikacije.

Customer	Savings	Assets	Income	Credit Risk
004	High	Low	≤$30,000	Good
009	High	Low	≤$30,000	Good
027	High	Low	≤$30,000	Bad
031	High	Low	≤$30,000	Bad
104	High	Low	≤$30,000	Bad

Primer lista koji nije potpuno “čist”

- Dalja grananja nisu moguća,
  pa krajnje pravilo (odluka) ima pouzdanost od 60%.

C4.5 algoritam

- C4.5 u odnosu na CART nije ograničen na binarna grananja.
  - Generiše zasebnu granu za svaku vrednost kategoričkog atributa.
- C4.5 metod se zasniva na meri homogenosti.
- Svako kandidat grananje deli polazni skup T na nekoliko podskupova T_1, T_2, . . . , T_k.
- *entropy_reduction*(S) = H(T)−H_S(T), gde je

$$H(X) = -\sum_j p_j \log_2(p_j)$$

C4.5 algoritam

- $H_S(T)$ je težinska suma entropija podskupova $T_1, T_2, \ldots, T_k$ i računa se kao:

$$H_S(T) = \sum_{i=1}^{k} P_i H_S(T_i)$$

- P_i predstavlja zastupljenost instanci u skupu *i*
- Za optimalno grananje uzima se ono sa najvećom vrednosti *entropy_reduction* (redukcijom entropije)

C4.5 primer

Customer	Savings	Assets	Income ($1000s)	Credit Risk
1	Medium	High	75	Good
2	Low	Low	50	Bad
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
6	High	High	25	Good
7	Low	Low	25	Bad
8	Medium	Medium	75	Good

Trening skup

Candidate Split	Child Nodes		
1	*Savings* = *low*	*Savings* = *medium*	*Savings* = *high*
2	*Assets* = *low*	*Assets* = *medium*	*Assets* = *high*
3	*Income* $\leq$ \$25,000		*Income* > \$25,000
4	*Income* $\leq$ \$50,000		*Income* > \$50,000
5	*Income* $\leq$ \$75,000		*Income* > \$75,000

Kandidat grananja za koreni čvor *t*

C4.5 primer

- 5/8 instanci je klasifikovano u klasu “Good”, a 3/8 u klasu “Bad”
- Početna entropija (pre grananja):

$$H(T) = -\sum_j p_j \log_2(p_j) = -\tfrac{5}{8}\log_2\left(\tfrac{5}{8}\right) - \tfrac{3}{8}\log_2\left(\tfrac{3}{8}\right) = 0.9544$$

- Za svako grananje se dobijena entropija poredi sa ovom vrednosti kako bi se odredilo koje grananje ima najveću redukciju u entropiji.

C4.5 primer

- Za kandidat grananje 1 (savings):
  $P_{high}=\frac{2}{8}$, $P_{medium}=\frac{3}{8}$, $P_{low}=\frac{3}{8}$.
- Entropija za čvor <u>savings = high</u>

$$-\frac{1}{2}\log_2\left(\frac{1}{2}\right)-\frac{1}{2}\log_2\left(\frac{1}{2}\right)=1$$

- Entropija za čvor <u>savings =medium</u>

$$-\frac{3}{3}\log_2\left(\frac{3}{3}\right)-\frac{0}{3}\log_2\left(\frac{0}{3}\right)=0$$

- Entropija za čvor <u>savings = low</u>

$$-\frac{1}{3}\log_2\left(\frac{1}{3}\right)-\frac{2}{3}\log_2\left(\frac{2}{3}\right)=0.9183$$

C4.5 primer

- Težinska kombinacija ovih entropija:

$$H_{\text{savings}}(T) = \frac{2}{8}(1) + \frac{3}{8}(0) + \frac{3}{8}(0.9183) = 0.5944$$

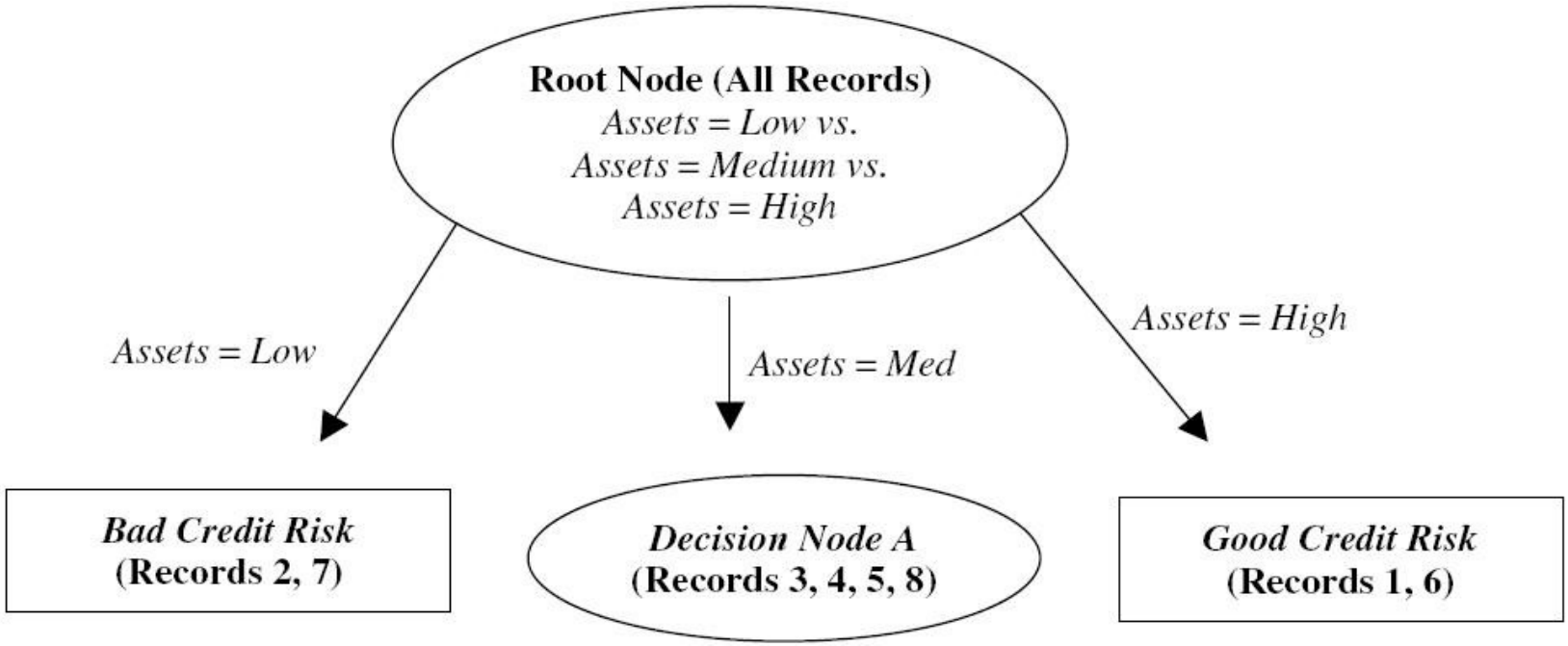
- Redukcija entropije za ovaj atribut je:

$$H(T) - H_{\text{savings}}(T) = 0.9544 - 0.5944 = 0.36$$

C4.5 primer

Candidate Split	Child Nodes	Information Gain (Entropy Reduction)
1	*Savings* = *low* *Savings* = *medium* *Savings* = *high*	0.36 bits
2	*Assets* = *low* *Assets* = *medium* *Assets* = *high*	0.5487 bits
3	*Income* ≤ $25,000 *Income* > $25,000	0.1588 bits
4	*Income* ≤ $50,000 *Income* > $50,000	0.3475 bits
5	*Income* ≤ $75,000 *Income* > $75,000	0.0923 bits

C4.5 primer


Root Node (All Records)
Assets = Low vs.
Assets = Medium vs.
Assets = High
Assets = Low
Assets = Med
Assets = High
Bad Credit Risk
(Records 2, 7)
Decision Node A
(Records 3, 4, 5, 8)
Good Credit Risk
(Records 1, 6)

C4.5 primer

Customer	Savings	Assets	Income ($1000s)	Credit Risk
3	High	Medium	25	Bad
4	Medium	Medium	50	Good
5	Low	Medium	100	Good
8	Medium	Medium	75	Good

Instance čvora A

- Početna (pre grananja) entropija:

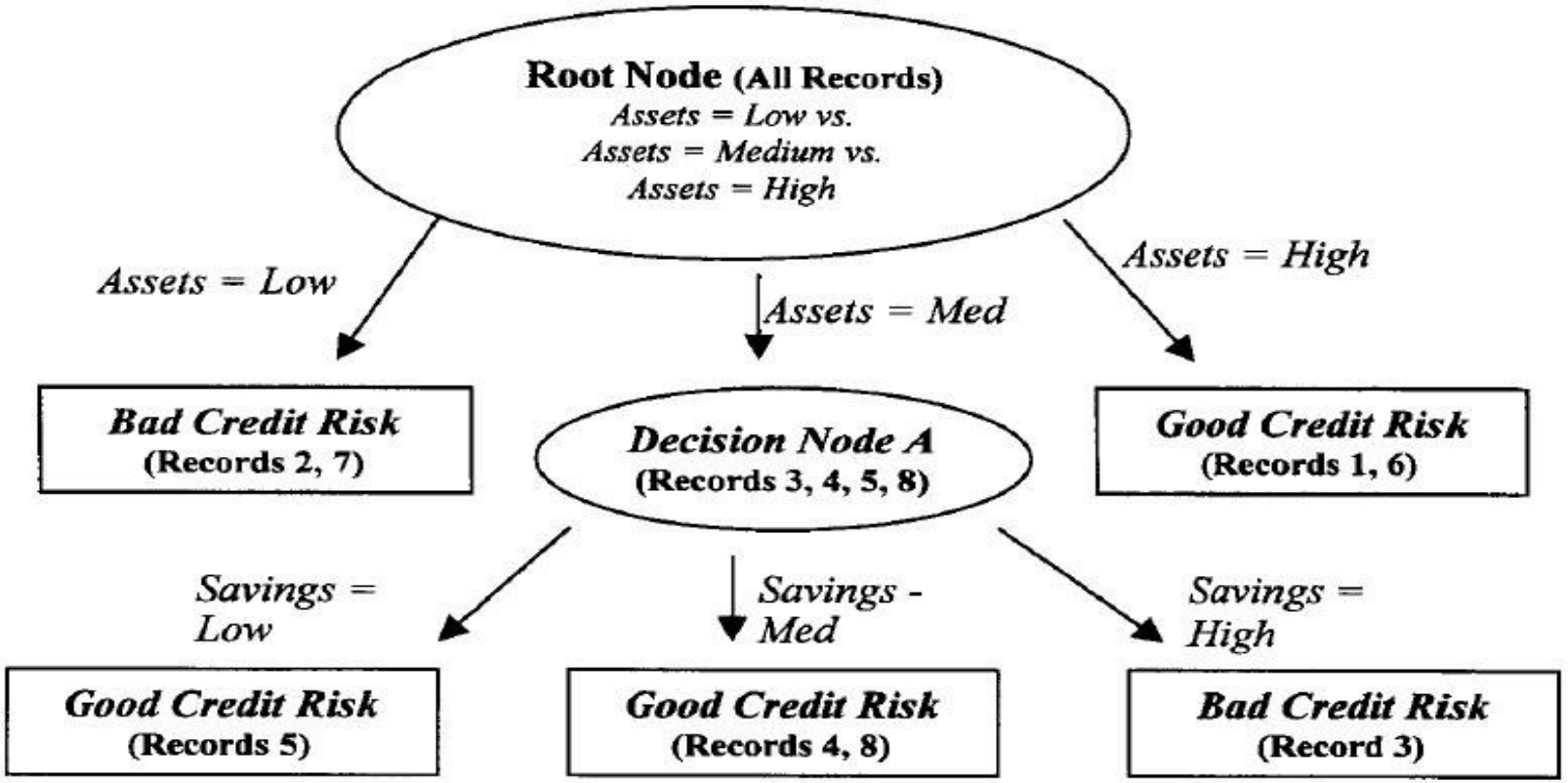
$$H(A) = -\sum_j p_j \log_2(p_j) = -\tfrac{3}{4}\log_2\left(\tfrac{3}{4}\right) - \tfrac{1}{4}\log_2\left(\tfrac{1}{4}\right) = 0.8113$$

C4.5 primer

Candidate Split	Child Nodes		
1	*Savings* = *low*	*Savings* = *medium*	*Savings* = *high*
3	*Income* $\leq$ \$25,000		*Income* > \$25,000
4	*Income* $\leq$ \$50,000		*Income* > \$50,000
5	*Income* $\leq$ \$75,000		*Income* > \$75,000

Kandidat grananja za čvor A

C4.5 primer



C4.5 stablo odlučivanja

Pravila

Antecedent	Consequent	Support	Confidence
If *assets* = *low*	then *bad credit risk.*	$\frac{2}{8}$	1.00
If *assets* = *high*	then *good credit risk.*	$\frac{2}{8}$	1.00
If *assets* = *medium* and *savings* = *low*	then *good credit risk.*	$\frac{1}{8}$	1.00
If *assets* = *medium* and *savings* = *medium*	then *good credit risk.*	$\frac{2}{8}$	1.00
If *assets* = *medium* and *savings* = *high*	then *bad credit risk.*	$\frac{1}{8}$	1.00

Pravila generisana na osnovu stabla odlučivanja

Reference

- *Discovering Knowledge in Data: An Introduction to Data Mining* (Wiley, 2005) Larose D.
- SQL Server Data mining tutorial (Lessons 1-6)

  http://technet.microsoft.com/en-us/library/ms167167.aspx