

Jerzy Stefanowski

Instytut Informatyki Politechniki Poznańskiej

Wykład z przedmiotu Uczenie maszynowe i sieci neuronowe

## INDUKCJA DRZEW DECYZYJNYCH

---

1. Pojęcia podstawowe.
2. Idea algorytmów **TDIT**.
3. Kryteria oceny atrybutów – entropia.
4. "Klasyczna" postać algorytmu **ID3**.
5. Przykład ilustracyjny.
6. Transformacja drzewa do zbioru reguł.
7. Ograniczenia w stosowaniu drzew decyzyjnych i rozszerzenia algorytmów.
8. Uwzględnianie atrybutów wielowartościowych oraz binaryzacja drzew.
9. Uwzględnianie atrybutów liczbowych.
10. Zjawisko przeuczenia i wwzględnianie danych „zazumionych”.
11. Techniki ograniczania rozmiarów drzew (pre-pruning oraz post-pruning).
12. Przykłady zastosowań praktycznych

# 1. Pojęcia podstawowe

---

Przykłady uczące - dane w reprezentacji atrybut-wartość

Obiekt = (wektor atrybut-wartość, klasa decyzyjna)

Obiekt	Atr1	Atr2	...	Klasa
1	v11	v12	...	C1
2	v21	v22	...	C2
3	...	...	...	...
...	...	...	...	...

**CEL:** znaleźć drzewo decyzyjne, które "wyjaśnia" przykłady uczące oraz poprawnie klasyfikuje nowe obiekty.

## Uwagi:

- Poszukiwana funkcja ma dyskretne wartości wyjściowe.
- Opisy przykładów uczących oraz poszukiwana wiedza klasyfikacyjna reprezentowane są w postaci symbolicznej.
- Pojęcia mogą być opisywane w sposób dysjunkcyjny.
- Dane treningowe mogą zawierać błędy.

## Drzewa decyzyjne

### **Ogólna idea:**

- Rekurencyjny podział zbioru uczącego na podzbiory
- Predykcja: przejście gałęzi, wybór najbardziej prawdopodobnej klasy decyzyjnej

### **Określenia:**

**TDIDT = Top Down Induction of Decision Trees**

**Kryterium wyboru:** używane do podziału zbioru uczącego

**Kryterium zatrzymania:** definiuje warunki zatrzymania procesu podziału

**Algorytm redukcji** (ang. tree pruning): próba zapobiegania przespecjalizowania tworzonego drzewa.

### **Zalety:**

Szybkość klasyfikowania, zrozumiałość, "dojrzała" metodologia, liczne zastosowania.

### **Ukierunkowanie:**

Małe drzewa decyzyjne minimalizujące błędne zaklasyfikowania

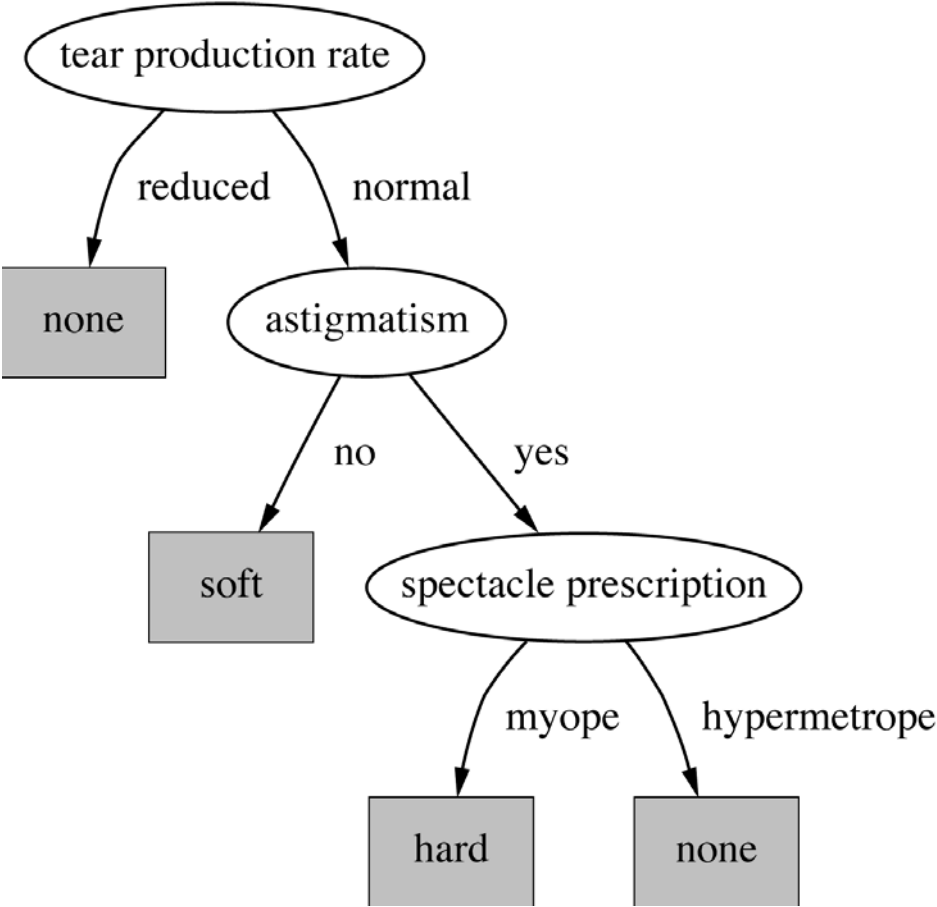
Znalezienie optymalnego drzewa - NP-trudne !!

**Strategia przeszukiwania:** Top-down

## Przykład budowy drzewa – contact lenses

Age	Spectacle prescription	Astigmatism	Tear production rate	Recommended lenses
Young	Myope	No	Reduced	None
Young	Myope	No	Normal	Soft
Young	Myope	Yes	Reduced	None
Young	Myope	Yes	Normal	Hard
Young	Hypermetrope	No	Reduced	None
Young	Hypermetrope	No	Normal	Soft
Young	Hypermetrope	Yes	Reduced	None
Young	Hypermetrope	Yes	Normal	hard
Pre-presbyopic	Myope	No	Reduced	None
Pre-presbyopic	Myope	No	Normal	Soft
Pre-presbyopic	Myope	Yes	Reduced	None
Pre-presbyopic	Myope	Yes	Normal	Hard
Pre-presbyopic	Hypermetrope	No	Reduced	None
Pre-presbyopic	Hypermetrope	No	Normal	Soft
Pre-presbyopic	Hypermetrope	Yes	Reduced	None
Pre-presbyopic	Hypermetrope	Yes	Normal	None
Presbyopic	Myope	No	Reduced	None
Presbyopic	Myope	No	Normal	None
Presbyopic	Myope	Yes	Reduced	None
Presbyopic	Myope	Yes	Normal	Hard
Presbyopic	Hypermetrope	No	Reduced	None
Presbyopic	Hypermetrope	No	Normal	Soft
Presbyopic	Hypermetrope	Yes	Reduced	None
Presbyopic	Hypermetrope	Yes	Normal	None

# Budowa drzewa ID3 – contact lenses



# Top Down Induction of Decision Trees

---

za "klasyczną" postacią ID3 [Quinlan 1983]

Podstawowy zapis techniki uczenia się drzewa  $T$  ze zbioru uczącego  $S$ :

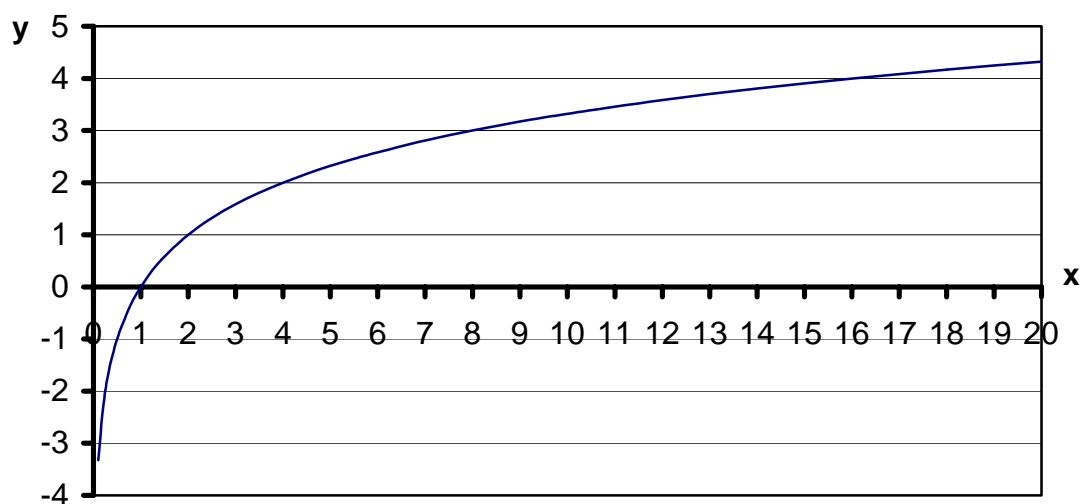
1. Wybierz "najlepszy" atrybut  $A$
2. Rozbuduj drzewo poprzez dodanie do węzła nowych gałęzi odpowiadającym poszczególnym wartościom  $v_1, v_2, \dots, v_n$  atrybutu  $A$
3. Podziel  $S$  na podzbiory  $S_1, S_2, \dots, S_n$  odpowiadające wartościom  $v_1, v_2, \dots, v_n$  atrybutu  $A$  i przydziel te podzbiory do odpowiednich gałęzi rozbudowywanego drzewa
4. Jeśli wszystkie przykłady w  $S_i$  należą do tej samej klasy  $C$  zakończ gałąź liściem wskazującym  $C$ ; w przeciwnym razie rekurencyjnie buduj drzewo  $T_i$  dla  $S_i$  (tzn. powtórz kroki 1-4)

## Funkcja logarytmiczna

$$y = \log_a x$$

$a$  – podstawa logarytmu  $x = a^y$

Rozważmy funkcję logarytmiczną dla  $a = 2$  (tj.  $y = \log_2 x$ )



$x$	1/8	1/4	1/2	1	2	4	8
$y$	-3	-2	-1	0	1	2	3

## Miary wyboru testów/atributów

---

$S$  - zbiór przykładów jednej z  $k$  klas decyzyjnych, ozn.  $K_1, \dots, K_k$ .

$n$  - liczba przykładów z  $S$ , oraz  $n_i$  - liczebność klasy  $K_i$ .

**Entropia** związana z klasyfikacją zbioru przykładów  $S$ :

$$Ent(S) = - \sum_{i=1}^k p_i \lg_2 p_i ,$$

gdzie  $p_i$  jest prawdopodobieństwem, że losowo wybrany przykład z  $S$  należy do klasy  $K_i$  ( $p_i$  estymowane jako  $n_i / n$ ).

---

Atrybut  $a$  użyty do zbudowania testu w węźle drzewa -  
**entropia warunkowa.**

**Zał.** atrybut  $a$  przyjmuje  $p$  różnych wartości  $\{v_1, v_2, \dots, v_p\}$ .  
 $S = \{S_1, S_2, \dots, S_p\}$ , gdzie  $S_j$  zawiera przykłady z  $v_j$  wartością atrybutu ( $j=1, \dots, p$ );  $n_{S_j}$  - liczebność zbioru  $S_j$ .

**Entropia** podziału zbioru przykładów  $S$  ze względu na atrybut  $a$ :

$$Ent(S | a) = \sum_{j=1}^p \frac{n_{S_j}}{n} \cdot Ent(S_j)$$

**Przyrost informacji** wynikający z zastosowania atrybutu  $a$  do zbudowania testu dzielącego zbiór przykładów  $S$ :

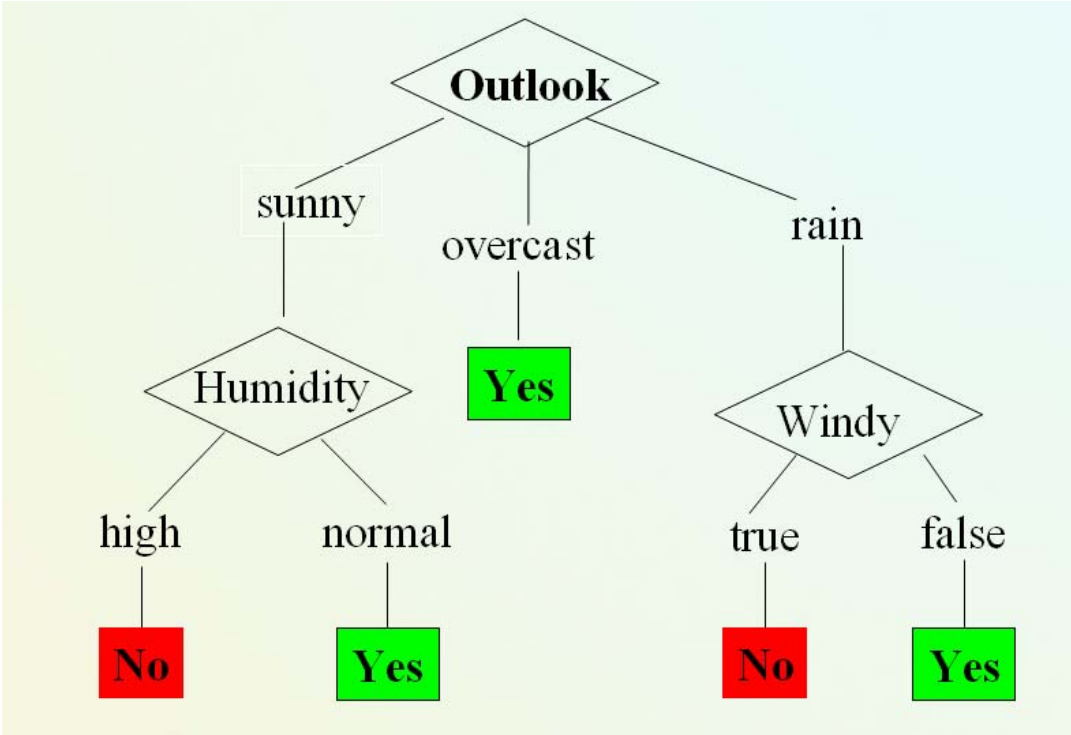
$$Gain(S, a) = Ent(S) - Ent(S | a)$$



## Przykład Quinlana – „Play Sport”

---

Outlook	Temperature	Humidity	Windy	Play?
sunny	hot	high	false	No
sunny	hot	high	true	No
overcast	hot	high	false	Yes
rain	mild	high	false	Yes
rain	cool	normal	false	Yes
rain	cool	normal	true	No
overcast	cool	normal	true	Yes
sunny	mild	high	false	No
sunny	cool	normal	false	Yes
rain	mild	normal	false	Yes
sunny	mild	normal	true	Yes
overcast	mild	high	true	Yes
overcast	hot	normal	false	Yes
rain	mild	high	true	No



# Rozwiązanie przykładu – Play Sport

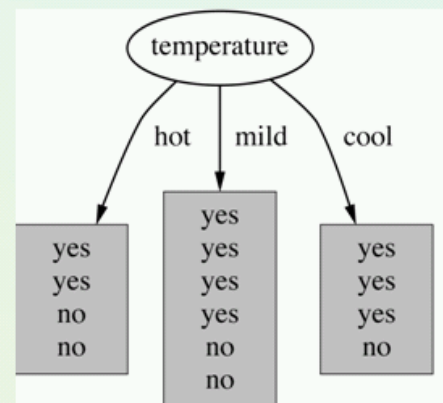
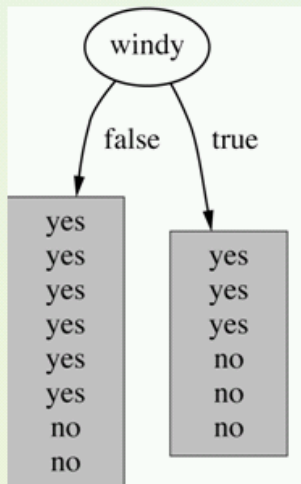
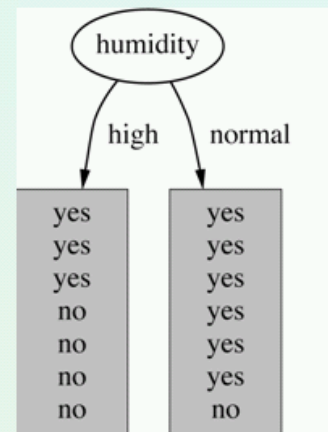
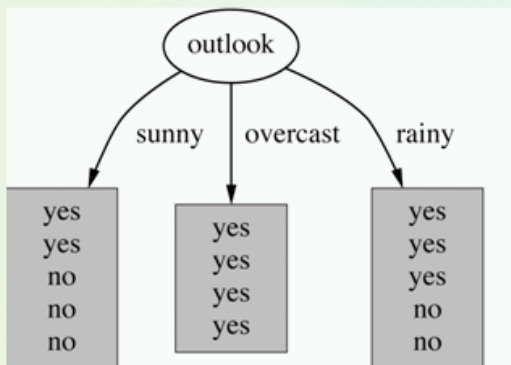
$$P_{yes} = -\left(\frac{9}{14}\right) \log_2\left(\frac{9}{14}\right) = 0.41$$

$$P_{no} = -\left(\frac{5}{14}\right) \log_2\left(\frac{5}{14}\right) = 0.53$$

$$E(S) = P_{yes} + P_{no} = 0.94$$

Outlook	Temp.	Humidity	Windy	Play
Sunny	Hot	High	False	No
Sunny	Hot	High	True	No
Overcast	Hot	High	False	Yes
Rainy	Mild	High	False	Yes
Rainy	Cool	Normal	False	Yes
Rainy	Cool	Normal	True	No
Overcast	Cool	Normal	True	Yes

Outlook	Temp.	Humidity	Windy	play
Sunny	Mild	High	False	No
Sunny	Cool	Normal	False	Yes
Rainy	Mild	Normal	False	Yes
Sunny	Mild	Normal	True	Yes
Overcast	Mild	High	True	Yes
Overcast	Hot	Normal	False	Yes
Rainy	Mild	High	True	No



## Example: attribute "Outlook"

- "Outlook" = "Sunny":

$$\text{info}([2,3]) = \text{entropy}(2/5,3/5) = -2/5 \log(2/5) - 3/5 \log(3/5) = 0.971$$

- "Outlook" = "Overcast":

$$\text{info}([4,0]) = \text{entropy}(1,0) = -1 \log(1) - 0 \log(0) = 0$$

*Note:  $\log(0)$  is not defined, but we evaluate  $0 \cdot \log(0)$  as zero*

- "Outlook" = "Rainy":

$$\text{info}([3,2]) = \text{entropy}(3/5,2/5) = -3/5 \log(3/5) - 2/5 \log(2/5) = 0.971$$

- Expected information for attribute:

$$\begin{aligned} \text{info}([3,2],[4,0],[3,2]) &= (5/14) \times 0.971 + (4/14) \times 0 + (5/14) \times 0.971 \\ &= 0.693 \end{aligned}$$

- Information gain:

(information before split) – (information after split)

$$\begin{aligned} \text{gain}(\text{"Outlook"}) &= \text{info}([9,5]) - \text{info}([2,3],[4,0],[3,2]) = 0.940 - 0.693 \\ &= 0.247 \end{aligned}$$

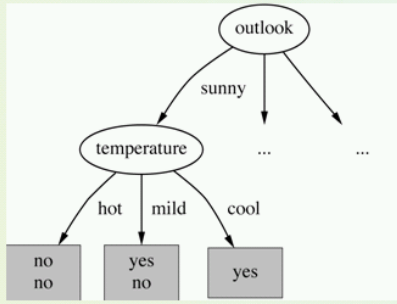
- Information gain for attributes from weather data:

$$\text{gain}(\text{"Outlook"}) = 0.247$$

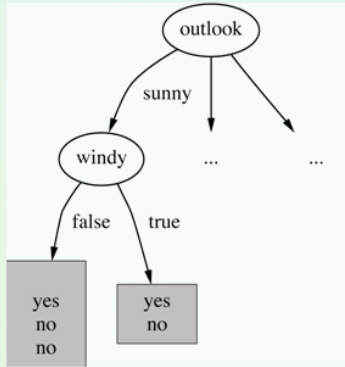
$$\text{gain}(\text{"Temperature"}) = 0.029$$

$$\text{gain}(\text{"Humidity"}) = 0.152$$

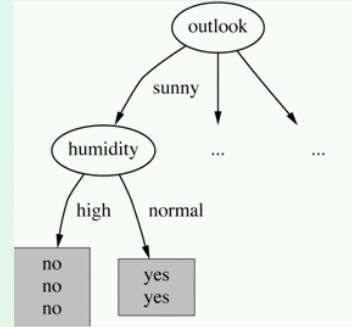
$$\text{gain}(\text{"Windy"}) = 0.048$$



gain("Temperature") = 0.571



gain("Windy") = 0.020



gain("Humidity") = 0.971

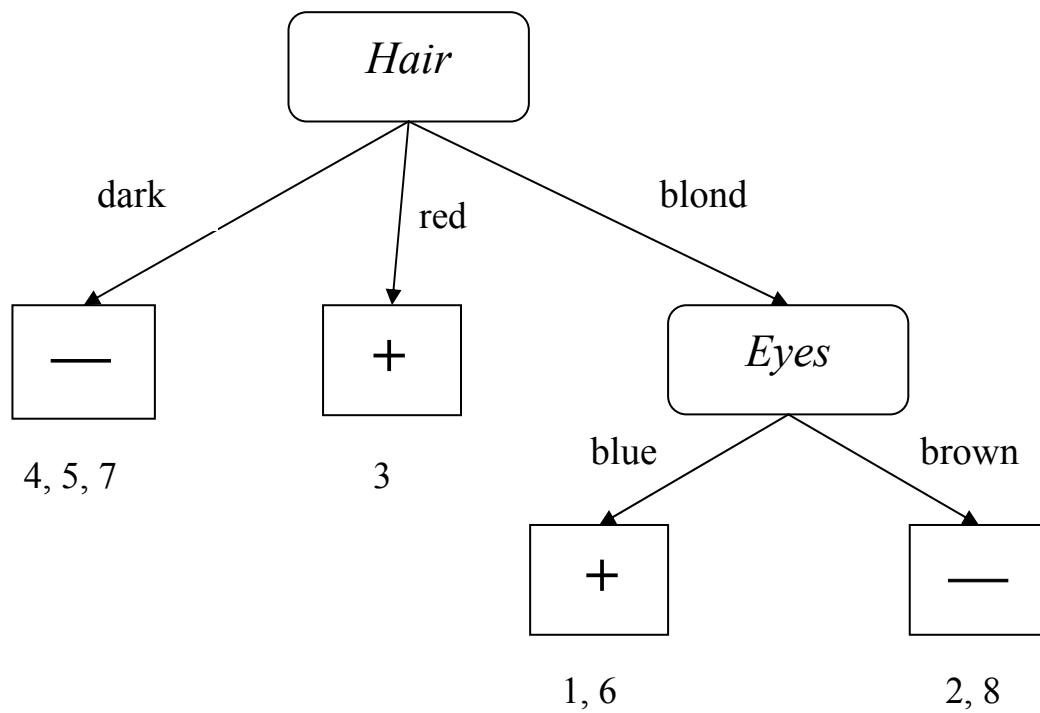
## Zbiór przykładów uczących - J.R. Quinlan

---

no.	Height	Hair	Eyes	Attractiveness
1	short	blond	blue	+
2	tall	blond	brown	-
3	tall	red	blue	+
4	short	dark	blue	-
5	tall	dark	blue	-
6	tall	blond	blue	+
7	tall	dark	brown	-
8	short	blond	brown	-

## Drzewo decyzyjne dla przykładów *Attractiveness*

---



## Zbiór reguł decyzyjnych:

---

r1: (Hair=Dark)  $\rightarrow$  (Attractiveness,-)

Pokryte przykłady {4,5,7}

r2: (Hair=Red)  $\rightarrow$  (Attractiveness,+)

Pokryte przykłady {3}

r3: (Hair=Blond) $\wedge$ (Eyes=Blue)  $\rightarrow$  (Attractiveness,+)

Pokryte przykłady {1,6}

r4: (Hair=blond) $\wedge$ (Eyes=Brown)  $\rightarrow$  (Attractiveness,-)

Pokryte przykłady {2,8}

Techniki poszukiwania nienadmiarowych reguł decyzyjnych: tzw. "dropping condition" (ang.):

- liniowa (ang. linear),
- wykładnicza (ang. exponential).

r4': (Eyes=Brown)  $\rightarrow$  (Attractiveness,-)

Pokryte przykłady {2,7,8}



## Przeszukiwanie przestrzeni hipotez w ID3

---

- Wybór języka reprezentacji silnie ogranicza przestrzeń hipotez
- Preferencja poszukiwania porządkuje wybór hipotez
- Strategia poszukiwania bez nawrotu
- Statystyczne kryterium wyboru atrybutu dość odporne na "szumy" danych wejściowych
- Ukierunkowanie do wyboru prostych drzew - *Occam's Razor* (jest mniej prostych hipotez niż rozbudowanych; prosta hipoteza zgodna z przykładami uczącymi nie jest przypadkowa, proste hipotezy lepiej sprawdzają się w praktycznych zastosowaniach)
- **Lecz !** nadal możemy otrzymać wiele równoważnych prostych drzew decyzyjnych; istnieją dalsze ograniczenia dla "źle zdefiniowanych" zbiorów danych.
- Nadal wiele dyskusji nad kryteriami wyboru atrybutów i sposobów redukcji rozmiarów drzew

## **Drzewa decyzyjne: Kryteria wyboru (cd.)**

### **Information gain (Quinlan 1983)**

$$\text{Gain}(A) = E(S) - I(A)$$

$I(A)$  = Oczekiwana informacyjność drzewa z atrybutem  $A$  jako korzeniem =  $\sum_{j=1}^n q_{v_j} E(S_j)$

$$E(S) = -\sum_{i=1}^m p_i \log p_i$$

### **Wiele innych propozycji kryteriów wyboru: np.**

GINI Index (CART, 84)

Gain ration (Quinalan, 86)

Binarization (Bratko & Kononenko, 86)

Chi-square statistic (Mingres, 89)

Normalized information gain (Mantara, 91)

$$\text{GINI Index: } 1 - \sum_i p_i^2$$

# Ograniczenia w uczeniu się drzew decyzyjnych

---

## Pytania i problemy, np.:

- Kiedy należy zaprzestać rozbudowywać drzewa?
  - aby zapobiec przespecjalizowaniu opisu
  - duże drzewa są trudne do analizy i zrozumienia
- Jak uwzględniać atrybuty ilościowe?
- Jak uwzględniać atrybuty ze zbyt dużą liczbą wartości w stosunku do dziedzin pozostałych atrybutów?
- Jak uwzględniać atrybuty z nieznanymi wartościami?
- Jak uwzględniać dane "zazsumione"?

## Uwzględnianie atrybutów o zbyt dużej liczbie wartości

---

**Trudność:** miara *Information Gain* faworyzuje atrybuty o dużej liczbie przyjmowanych wartości

**Przykład:** atrybut - data urodzenia

**Podejścia:**

- binaryzacja atrybutów
- zmiana miary *Inf Gain* na inną

Propozycje: ***Gain Ratio*** [Quinlan 93]

**Intuicja:** karać zbyt szerokie dziedziny atrybutów; dążyć do znormalizowanych rozkładów wartości.

$$Gain\_Ratio(S,A) = Gain(S,A) / Info(A)$$

gdzie

$$Info(A) = - \sum_{j=1}^n \frac{|S_j|}{|S|} \log_2 \frac{|S_j|}{|S|}$$

## Binaryzacja drzew decyzyjnych - przykład:

---

no.	Height	HairBlond	HairRed	HairDark	Eyes	Attractiveness
1	short	+	-	-	blue	+
2	tall	+	-	-	brown	-
3	tall	-	+	-	blue	+
4	short	-	-	+	blue	-
5	tall	-	-	-	blue	-
6	tall	+	-	-	blue	+
7	tall	-	-	+	brown	-
8	short	+	-	-	brown	-

$$H(A|Height)=0.951$$

$$H(A|HairBlond)=0.781$$

$$H(A|HairRed)=0.754$$

$$H(A|HairDark)=0.607$$

$$H(A|Eyes)=0.607$$

## Binaryzacja drzew decyzyjnych - cd:

no.	Height	Hair <sub>Blond</sub>	Hair <sub>Red</sub>	Eyes	Attractiveness
1	short	+	-	blue	+
2	tall	+	-	brown	-
3	tall	-	+	blue	+
6	tall	+	-	blue	+
8	short	+	-	brown	-

Reguły decyzyjne:

r1: (Hair=Dark)  $\rightarrow$  (Attractiveness,-)

Pokryte przykłady {4,5,7}

r2: (Hair=not Dark) $\wedge$ (Eyes=Blue)  $\rightarrow$  (Attractiveness,+)

Pokryte przykłady {1,3,6}

r3: (Hair=not Dark) $\wedge$ (Eyes=Brown)  $\rightarrow$  (Attractiveness,-)

Pokryte przykłady {2,8}

## Uwzględnianie atrybutów liczbowych

---

**Problem:** Poszukiwanie prostych i łatwych do zrozumienia warunków.

Pre-processing: Dyskretyzacja przed uruchomieniem algorytmu indukcji drzewa.

**Propozycja:** tworzyć testy wartości atrybutów „dynamicznie” w procesie budowy drzewa (*atrybut rel próg*); gdzie *rel* jest operatorem typu  $<$ ,  $\leq$ ,  $>$ ,  $\geq$ , *próg* jest pewną charakterystyczną wartością atrybutu.

Np. Temperatura  $\rightarrow$  (*Temperatura*  $>$  *c*)

Temperatura	5	8	17	19	22	29
Play-Tennis?	Nie	Nie	Tak	Tak	Tak	Nie

Jakie wartości *c* powinny być rozważane?

Wiele istniejących propozycji: [Quinalan 86,93], [Wiess 90], [Shapiro 91],...

## Przykład drzewa binarnego (c4.5 dane crx)

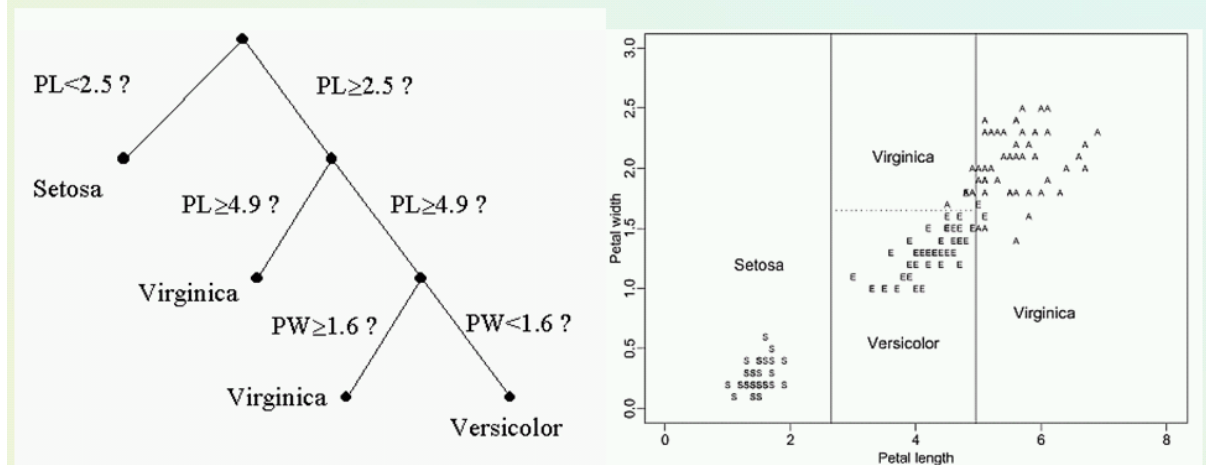
Pruned decision tree:

```
A9 - t:
|
|   A15 > 228 : + (106.0/3.8)
|   A15 <= 228 :
|   |
|   |   A14 <= 102 :
|   |   |
|   |   |   A4 in {l,t}: + (0.0)
|   |   |   A4 - u:
|   |   |   |
|   |   |   |   A6 in {c,d,cc,i,k,m,q,w,x,e,aa}: + (46.4/3.1)
|   |   |   |   A6 in {j,ff}: - (2.0/1.0)
|   |   |   |   A6 - r: + (0.0)
|   |   |   |
|   |   |   |   A4 - y:
|   |   |   |   |
|   |   |   |   |   A6 in {c,i,aa,ff}: - (7.0/3.4)
|   |   |   |   |   A6 in {d,j,w,x}: + (4.0/1.2)
|   |   |   |   |   A6 in {cc,k,m,r,q,e}: + (0.0)
|   |   |   |
|   |   |   |   A14 > 102 :
|   |   |   |   |
|   |   |   |   |   A6 in {j,r}: + (0.0)
|   |   |   |   |   A6 in {c,d,k,m,e,aa,ff}:
|   |   |   |   |   |
|   |   |   |   |   |   A14 <= 132 : - (4.1/1.2)
|   |   |   |   |   |   A14 > 132 :
|   |   |   |   |   |   |
|   |   |   |   |   |   |   A3 <= 1.625 :
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   A14 <= 292 : - (13.0/1.3)
|   |   |   |   |   |   |   |   A14 > 292 :
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   A13 = g: + (2.0/1.0)
|   |   |   |   |   |   |   |   |   A13 = s: - (6.0/2.3)
|   |   |   |   |   |   |   |   |   A13 = p: - (0.0)
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   A3 > 1.625 :
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   A6 in {k,m}: + (5.0/1.2)
|   |   |   |   |   |   |   |   |   A6 = ff: + (0.0)
|   |   |   |   |   |   |   |   |   A6 in {c,d,e,aa}:
|   |   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   |   A2 <= 32.08 : + (9.5/4.1)
|   |   |   |   |   |   |   |   |   |   A2 > 32.08 : - (8.0/3.5)
|   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   A6 in {cc,i,q,w,x}:
|   |   |   |   |   |   |   |   |
|   |   |   |   |   |   |   |   |   A8 <= 10.75 : + (36.0/9.3)
|   |   |   |   |   |   |   |   |   A8 > 10.75 : - (2.0/1.0)
|
|   A9 - f:
|   |
|   |   A4 in {u,y}: - (237.0/17.3)
|   |   A4 = l: + (2.0/1.0)
|   |   A4 = t: - (0.0)
```



# Interpretacja graficzna drzewa dla danych liczbowych

Hierarchical partitioning of feature space into hyper-rectangles.  
Example: Iris flowers data, with 4 features; displayed in 2-D.



## Problemy z przetwarzaniem danych niespójnych i zaszumionych

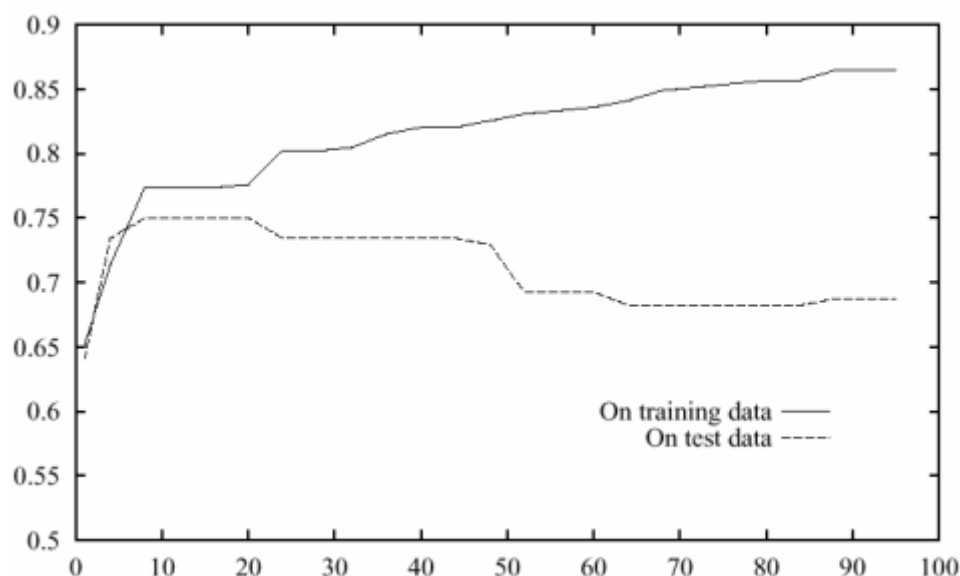
---

**Problem:** Klasyczne algorytmy (ID3) nie radzi sobie z tego przykładami:

- Posiadającymi niespójne (sprzeczne – inconsistent) opisy
- Mające charakter szumu (noise) z uwagi na położenie w zbiorze przykładów

Indukowane drzewa

- Nadmieranie rozbudowane – przespecjalizowane, przeuczenie (overfitting)
- Słabsze zdolności generalizacyjne
- Trudności w interpretacji struktury zbyt dużego drzewa



Zbiór przykładów uczących - wersja zawierająca tzw. sprzeczne przykłady (ang. inconsistent examples).

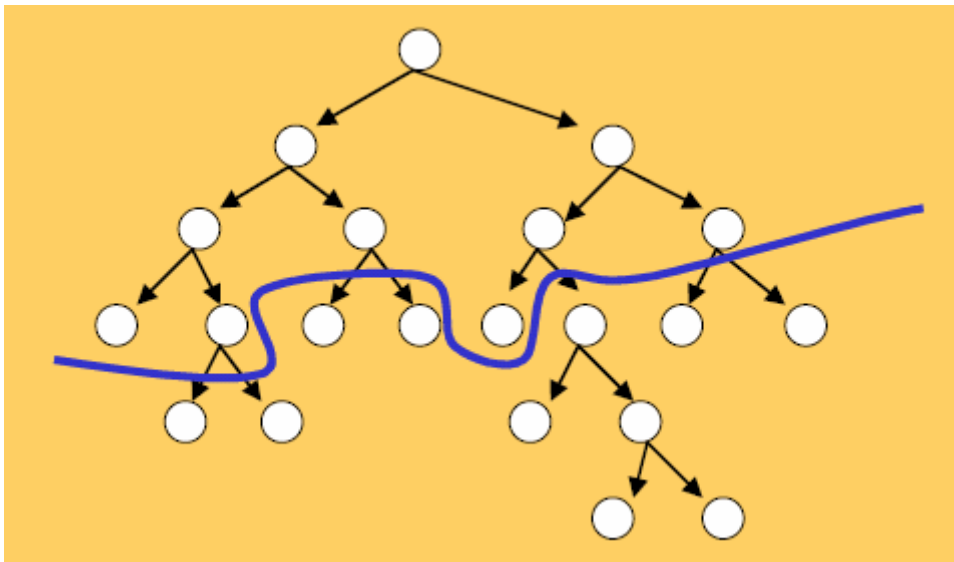
---

No	Height	Hair	Eyes	Attractiveness
1	short	blond	blue	+
2	tall	blond	brown	-
3	tall	red	blue	+
4	short	dark	blue	-
5	tall	dark	blue	-
6	tall	blond	blue	+
7	tall	dark	brown	-
8	short	blond	brown	-
9	tall	dark	brown	+

# Przeuczenie drzewa

---

- Zaszumione („noise”) lub zbyt nietypowe dane mogą powodować nadmierne dopasowanie do przykładów uczących.
- Dwa sposoby unikania przeuczenia:
  - (Pre-pruning): Zatrzymuj budowę drzewa wcześniej.
  - (Post-pruning): zbuduj przespecjalizowane pełne drzewo, później je zredukuj.
- Ocena wielkości redukcji:
  - Różne sposoby.
- Rule post-pruning (C4.5): zamień drzewo na reguły.



# Techniki unikania przespecjalizowania opisów oraz redukcji drzew decyzyjnych

---

## Propozycje:

- Zmienić kryterium zatrzymania, aby kończyć rozbudowę drzewa wcześniej (*pre-pruning/ forward pruning*)
- Zezwolić na budowę pełnego drzewa (przespecjalizowanego), następnie zredukować / obcinać zbędne gałęzie (*post-pruning*)

## Jak tworzyć i używać kryteria:

- optymalizować działanie na oddzielnym zbiorze testowym
- uwzględniać tylko statystycznie znaczące podziały drzew
- Minimalizuj( $Rozmiar[drzewo] + Liczba[źle sklasyfikowanych przykładów]$ )

## Techniki redukcji w trakcie budowy drzew decyzyjnych - *forward pruning, pre-pruning*

---

Zatrzymaj rozbudowę drzewa w danym węźle jeśli:

- *information gain* jest zbyt niski,
- liczba rozważanych przykładów uczących jest niewiarygodna (mała),
- zbiór przykładów jest statystycznie nieznaczący (np. Chi-square test),
- prównanie estymat wybranych błędów jest niekorzystne (static and back-up error Bratko 86).

## Podejścia "*Train and Test*"

---

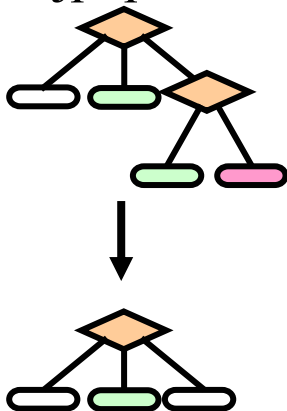
Uczyć się na części danych, testować na pozostałych

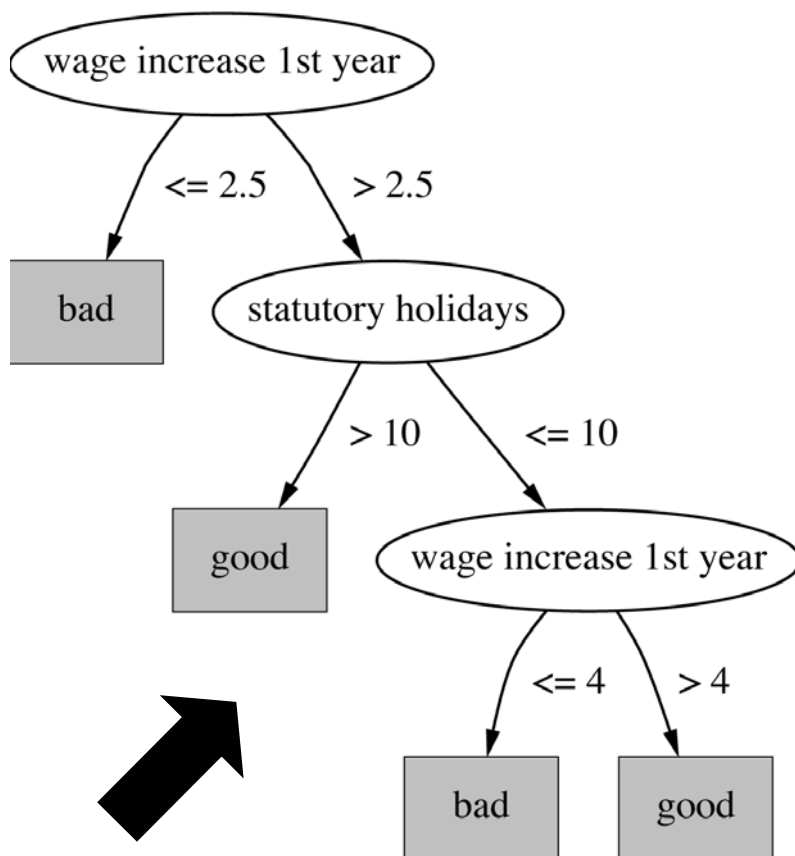
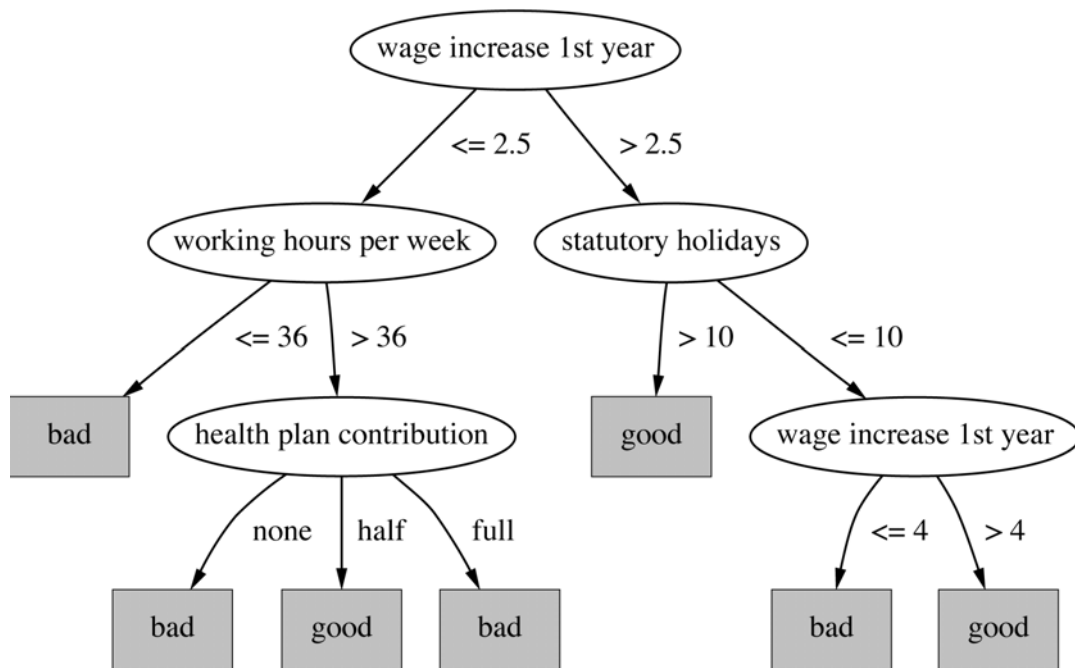
(Nie uwzględnia się założeń o rozkładach gęstości prawdopodobieństw !)

"*Reduced error pruning*":

1. Zbuduj drzewo decyzyjne w oparciu o zbiór uczący
2. Rozważaj każdy węzeł (poddrzewo) jako kandydata do usunięcia
3. Heurystycznie usuwaj węzły, które pozwalają uzyskać zysk na kryterium dla przykładów testowych

Znajdź najmniejsze drzewo spośród tych, które są najpoprawniejsze dla danych testowych







## Modyfikacje w podstawowym algorytmie

---

- Jeśli  $S$  jest niepustym zbiorem o niejednoznacznej klasyfikacji przykładów uczących i nie dysponuje się większą liczbą atrybutów do konstrukcji testów zwróć jako rezultat klasę większościową w rozkładzie gęstości klas decyzyjnych.
- "Najbardziej informatywne" atrybuty - nowe bardziej wyrafinowane miary.
- "Windowing" idea uwzględniania podzbiorów przykładów w analizie dużych zbiorów danych.