Strojové učení: klasifikace (6. přednáška)

April 6, 2016

Machine Learning — Naučit stroje se učit

O co jde?

Máme **model výpočtu** (t.j. výpočetní postup jednoznačně daný vstupy a nějakými parametry), chceme najít vhodné nastavení parametrů, aby postup (model) dával řešení našeho problému.

Při strojovém **učení s učitelem** zvolíme dostatečně silný model a nastavení parametrů chceme nalézt automaticky na základě vzorových řešení daného problému.

Klasifikace — příklad

Rozpoznávání věcí na fotkách

(zdroj: G. Hinton, Neural Networks for ML)

40 vydra
15 křepelka
7 tetřev
6 koroptev

85 sněžný pluh
6 vrtná plošina
6 záchranný člun
2 popelářské auto

15 žížala
12 gilotina
7 orangutan
6 koště

Klasifikace — klasický příklad

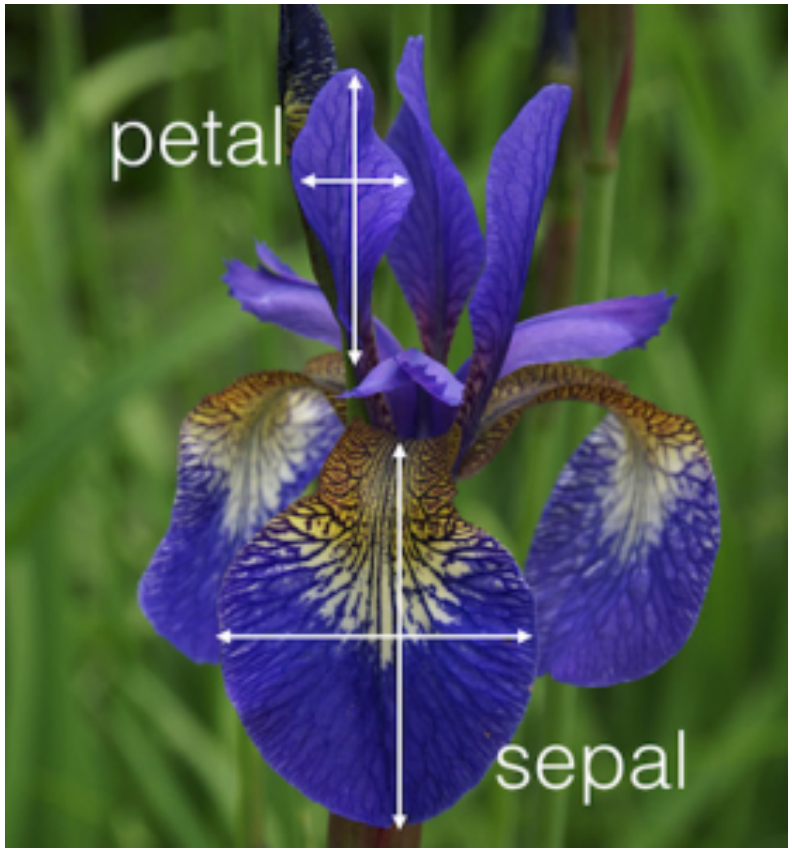
iris setosa

iris versicolor

iris virginica

- 50 vzorků od každého z těchto tří druhů
- měření je od Adgara Andersona z roku 1935, vzorky dvou ze tří zahrnutých druhů pochází "z téže pastviny, byly sebrány týž den, měřeny ve stejnou dobu touž osobou a stejnými přístroji"
- sadu proslavil Ronald Fischer tím, že ji využil jako příklad ve svém článku *The use of multiple measurements in taxonomic problems* (1936)

Klasifikace — Klasický příklad (pokračování)



měřily se 4 vlastnosti květů:

- sepal length = délka okvětních plátků (v cm)
- sepal width = šířka okvětních plátků (v cm)
- petal length = délka kališních lístků (v cm)
- petal width = šířka kališních lístků (v cm)

Hodnocení klasifikátoru

matice záměn (confusion matrix)

	předvídaný pozitivní	předvídaný negativní
skutečně pozitivní	a	b
skutečně negativní	c	d

správnost, přesnost (accuracy) $\frac{a+d}{a+b+c+d}$ kolik toho klasifikujeme správně

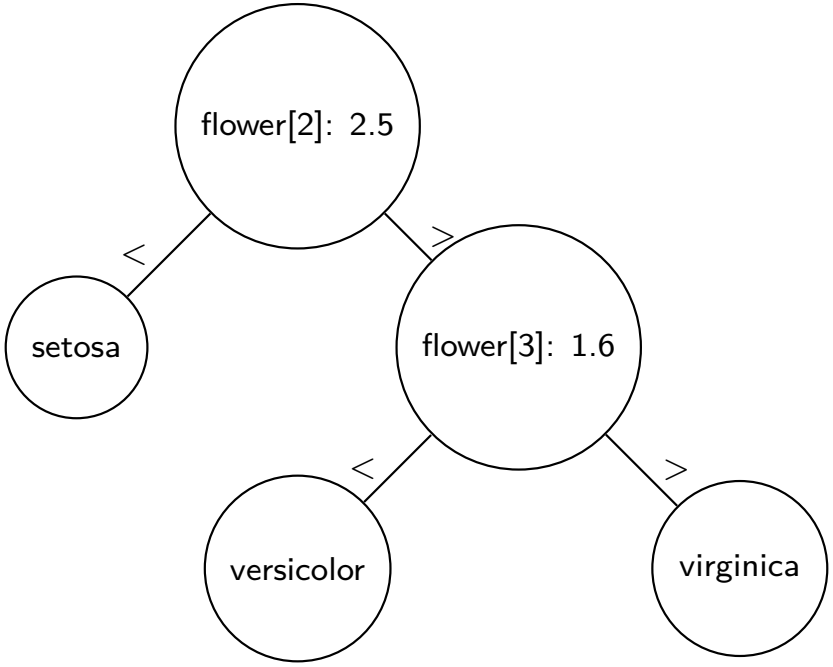
precisnost, přesnost (precision) $\frac{a}{a+c}$ jak moc se dá věřit našemu ANO

úplnost, senzitivita (recall, sensitivity) $\frac{a}{a+b}$ jaký podíl všech ANO odhalíme

specificita (specificity) $\frac{d}{c+d}$ jaký podíl všech NE odhalíme

F-míra (F-measure) $\frac{2a}{2a+b+c} = \frac{2*\text{přesnost}*\text{úplnost}}{\text{precisnost}+\text{úplnost}}$

Rozhodovací strom



Každý **uzel (node)** odpovídá podmnožině trénovacích vzorků.
V každém uzlu je zvolen jeden příznak; jsou-li jeho hodnoty číselné, tak též dělicí hodnota. (Jinak se strom rozvětví na všechny hodnoty, které příznak může nabývat.)

Rozhodovací stromy — budování

Nečistota uzlu (node impurity) je míra toho, jak moc jsou v daném uzlu pomíchané různé třídy.
Při budování stromu v každém uzlu zvolíme takový příznak a dělicí hodnotu, aby byl vážený součet nečistot v následnících tohoto uzlu co nejnižší. (V praxi se ukazuje, že na konkrétní volbě míry nečistoty až tolik nezáleží.)
Když vybudujeme velký strom, některé větve zase uřízneme, čímž zabráníme přeučení.

Rozhodovací stromy — nečistota uzlu

Nechť rozhodujeme mezi množinou tříd $\{c_1, \ldots, c_m\}$, a nechť f_i je podíl vzorků v daném uzlu, které patří do třídy c_i.

Giniho nečistota (Gini impurity) kdybych si náhodně vybrala jeden vzorek z množiny odpovídající danému uzlu, a tomuto vzorku náhodně přiřadila třídu na základě podílu, jaký mají na uzlu jednotlivé třídy, jaká je pravděpodobnost, že jsem ho označila špatně?

$$I_G(f) = \sum_{i=1}^{m} f_i(1 - f_i) = \sum_{i \neq k} f_i f_k$$

Kullback-Leiblerova divergence (information gain) tj. pokles nečistoty = vzájemná informace (mutual information), tj. pokles entropie: entropie děleného uzlu - vážený součet entropií jeho následníků, kde entropii definujeme takto:

$$I_E(f) = -\sum_{i=1}^{m} f_i \log_2 f_i$$

Přeučení

- přeučení (overfitting)
  - příliš silný model sice má velmi malou chybu na trénovacích datech, ale velkou chybu na datech, která předtím neviděl

prořezávání stromu na základě snížení chyby: začnu v listech, v každém uzlu zkusím všem vzorkům přiřadit nejčastější třídu; pokud se tím nesníží přesnost klasifikace (na testovacích datech), tak toto zjednodušení zachovám