Učení založené na instancích

Instance based learning

Charakteristika IBL – (nejbližších sousedů)

- Tyto metody nepředpokládají určitý model
- nejsou strukturované a typicky nejsou příliš užitečné pro porozumění vztahu mezi příznaky a cílovými třídami
- jsou užitečné a často velmi efektivní ”black box” nástroje pro klasifikaci
- k–NN lze použít i pro regresi, ale **bias–variance tradeof** není tak optimistický jako u klasifikace

IBL

- Příprava dat: Prostor atributů či příznaků nejdříve standardizuji tak, aby každý příznak/atribut měl střední hodnotu 0 a rozptyl 1 (přes všechny cílové třídy dohromady).
- Trénování v IBL probíhá tak, že pouze ukládáme příchozí data do paměti.
- Teprve, když přijde požadavek na klasifikaci (či predikci) nového případu, začneme tvořit model, k–nejbližších sousedů najde k nejbližších příkladů z trénovací databáze a klasifikuje nový příklad podle nejčastější klasifikace těchto k případů.

k –nejbližších sousedů

(k–NN, nearest neighbours)

Definuji:

- G je množina cílových tříd
- $\delta(g_1, g_2) = 1$ právě když $g_1 = g_2$, jinak $\delta(g_1, g_2) = 0$.

Klasifikace probíhá ve dvou krocích. Pro novou instanci x:

1. najdi k nejbližších instancí k x v $data$, označ je $x_1, \ldots, x_k$.
2. return $\hat{g}(x) = argmax_{g \in G} \sum_{i=1}^{k} \delta(g, g(x_i))$

Metrika pro nalezení k nejbližších sousedů

můžeme volit např. následující:

euklidovská	$d(x_i, x_j) = \sqrt{\sum_{r=1}^{n}(a_r(x_i) - a_r(x_j))^2)}$
Hammingova (Manhattan)	$d(x_i, x_j) = \sum_{r=1}^{n} \lvert a_r(x_i) - a_r(x_j) \rvert$
překrytí (overlap) pro kategorická data	$d(x_i, x_j) = \sum_{r=1}^{n}(1 - \delta(a_r(x_i), a_r(x_j)))$
kosínová	$d(x_i, x_j) = \frac{\sum_{r=1}^{n}(a_r(x_i) \cdot a_r(x_j))}{\sqrt{\sum_{r=1}^{n}(a_r(x_j) \cdot a_r(x_j)) \cdot \sum_{r=1}^{n}(a_r(x_i) \cdot a_r(x_i))}}$

Problémy k zlepšení

- příliš mnoho dat k ukládání
- 1–NN je citlivé na šum
- k–NN je pomalé na velké bázi dat
- mají–li všechny atributy stejnou váhu, pak je irelevantní atribut zmate
- **problém velké dimenzionality**

Idea komprese: ukládat jen špatně klasifikované vzhledem k předchozím

- to ale může vyhodit užitečné příklady
- ráda ukládá chybná (noisy) data, protože jsou špatně klasifikovaná
- ale idea je dobrá, jen potřebuje dopracovat

IB3

Idea: ukládat s každým exemplářem, kolikrát predikoval dobře a kolikrát ne.

- vymažu ty, co perdikují špatně
- nové instance predikuji na základě těch, co predikují výborně (acceptable)
- ty mezi držím a testuji – počítám, kolikrát by predikovaly dobře a kolikrát špatně
- vážený průměr $argmax_g \sum_{i=1}^{k} \delta(g(x_i), g) \cdot \frac{s_i}{N_i}$

IB3–algoritmus

$CD \leftarrow \{\}$ % Concept Description
pro každý příklad $x \in data$

- pro každé $u \in CD$: $dist(u) = distance(u, x)$
- pokud existuje výborně predikující $u \in CD$
  `pak` $u_{min} \leftarrow argmin_{u;u \text{ acceptable}} dist(u)$
  `jinak` $u_{min} \leftarrow$ náhodně vybraná instance z CD
- pokud $g(x) \neq g(u_{min})$ % nesprávná predikce
  `pak` $CD \leftarrow CD \cup \{x\}$
- pro každé $u \in CD$
  `pokud` $dist(u) \leq dist(u_{min})$
  - aktualizuj záznam předpovědí u
  - pokud u predikuje špatně, vyhoď $CD \leftarrow CD \setminus \{u\}$

Interval věrohodnosti – úvod

Na posousení kvality predikce potřebujeme intervaly věrohodnosti, pro ně potřebujeme znát:

p	pravdivá míra úspěšnosti
s	počet správných predikcí
N	celkový počet predikcí

Veličina s má binomické rozložení.

Odtud můžeme stanovit interval $\langle \frac{s}{N} - z\sigma, \frac{s}{N} + z\sigma \rangle$, ve kterém se skutečná hodnota p na 75% nachází (75%$= 1 - 2 \cdot \alpha$).

Interval věrohodnosti – návod

Pro velké N binomické rozložení konverguje k normálnímu rozložení o střední hodnotě $N \cdot p$ a rozptylu $\sigma^2 = N \cdot p \cdot (1-p)$, po normalizaci dostaneme ($f = \frac{s}{N}$) $\frac{f-p}{\sqrt{\frac{p(1-p)}{N}}} \approx N(0,1)$.

- Hledáme z, aby $P(-z < \frac{f-p}{\sqrt{\frac{p(1-p)}{N}}} < z) = 1 - 2\alpha$
- pro dané $f = \frac{s}{N}$ a α najdu v tabulce z
- puntičkáři přepočtou hranice pro p, tj. dosadí do

$$p = \frac{\left(f + \frac{z^2}{2N} \pm z\sqrt{\frac{f}{N} - \frac{f^2}{N} + \frac{z^2}{4N^2}}\right)}{\left(1 + \frac{z^2}{N}\right)}$$

- já odhadnu: $\frac{s}{N} - z \cdot \sqrt{\frac{\frac{s}{N}(1-\frac{s}{N})}{N}} \leq p \leq \frac{s}{N} + z \cdot \sqrt{\frac{\frac{s}{N}(1-\frac{s}{N})}{N}}$.

IB3–proměnné

- pravděpodobnost jednotlivých tříd označíme $\hat{p}_j$, počet dosud zpracovaných příkladů $\hat{N}$.
- pro každou třídu vždy spočteme $\langle l_{apri}, u_{apri} \rangle$,
  $l_{apri} = \hat{p}_j - 1,15 \cdot \sqrt{\frac{\hat{p}_j(1-\hat{p}_j)}{\hat{N}}}$
  $u_{apri} = \hat{p}_j + 1,645 \cdot \sqrt{\frac{\hat{p}_j(1-\hat{p}_j)}{\hat{N}}}$
- pravděpodobnost úspěšnosti instance *inst* označme $\hat{p}_{inst}$, počet jejích ″pokusů″ o predikci $\hat{N}_{inst}$.
- pro ″blízké″ instance spočteme $\langle l_{inst}, u_{inst} \rangle$,
  $l_{inst} = \hat{p}_{inst} - 1,645 \cdot \sqrt{\frac{\hat{p}_{inst}(1-\hat{p}_{inst})}{\hat{N}_{inst}}}$
  $u_{inst} = \hat{p}_{inst} + 1,15 \cdot \sqrt{\frac{\hat{p}_{inst}(1-\hat{p}_{inst})}{\hat{N}_{inst}}}$

IB3–hodnocení instancí

- **Odstranit příklad jako špatný?** Volíme $\alpha = 12,5\%$, tj. $z = 1,15$.
  - pro apriorní pravděpodobnost třídy, tj. kdybych neshromažďovala žádné příklady, jen četnost tříd; dostanu interval $\langle l_{apri}, u_{apri} \rangle$
  - pro daný příklad, tj. $p_{inst} = \frac{s_{inst}}{N_{inst}}$ a N_{inst}, dostanu interval $\langle l_{inst}, u_{inst} \rangle$
  - pokud $u_{inst} < l_{apri}$, tak instance predikuje špatně, vyhodit
- **Použít příklad pro predikci?** Volíme $\alpha = 5\%$, tj. $z = 1,645$.
  - pro apriorní pravděpodobnost třídy dostanu interval $\langle l_{apri}, u_{apri} \rangle$
  - pro daný příklad, tj. p_{inst} a N_{inst}, dostanu interval $\langle l_{inst}, u_{inst} \rangle$
  - pokud $u_{apri} < l_{inst}$, tak instance predikuje výborně, používat