

Testování a validace modelů

Michael Anděl



Evropský sociální fond
Praha & EU: Investujeme do vaší budoucnosti

Fakulta elektrotechnická, ČVUT

- ▶ V RapidMineru vytvořte co nejpřesnější prediktivní model pro data *horse colic*, který bude predikovat `surgical lesion`!
- ▶ Bodována bude přesnost modelu na neznámých datech.

Jak na to:

1. Založte proces v RM. Importujte data *horse colic*. Data obsahují různé typy příznaků (spojité, diskrétní/ordinární, nominální). Věnujte tedy pozornost správnému otagování atributů při importu.
2. Implementujte bloky vám známých klasifikátorů (z přednášek, cvičení).
3. Pomocí operáru `X-Validation` evaluujte přesnost vašich klasifikátorů a vyberte nejlepší. Věnujte pozornost volbě evaluační metriky.
4. Připravte testovací podproces, který:
 - 4.1 stejným způsobem jako v 1) načte „budoucí“ testovací data a otaguje atributy
 - 4.2 použije vybraný (nejlepší) naučený model z 2) a použije je na načtená data a vyčíslí klasifikační přesnost, tj. `accuracy`

- ▶ Protokol = 2 body
- ▶ V protokolu bude:
 - ▶ Krátký popis klasifikační úlohy, tj. co klasifikujeme, jaké typy atributů máte, resp. jste otagovali.
 - ▶ Popis validačního protokolu
 - ▶ Nejlepší model, odhad jeho přesnosti a interpretace.
- ▶ Odevzdejte váš RM proces
- ▶ Bude otestován na neznámých datech:
 - ▶ 80 – 100 % plus 6 b.
 - ▶ 70 – 80 % plus 4 b.
 - ▶ 55 – 70 % plus 2 b.
- ▶ Deadline až za 2 týdny od zadání!!!