Vilnius universitetas
Matematikos ir informatikos fakultetas
Duomenų mokslo ir skaitmeninių technologijų institutas

Duomenų klasifikavimas

prof. dr. Olga Kurasova
Olga.Kurasova@mif.vu.lt

Dar apie duomenų klasifikavimą

- Pagal turimus duomenis, kurių klasės yra žinomos, **reikia sukurti mechanizmą** (klasifikatorių), kuris gebėtų priskirti klases duomenims, kuriems jos nėra žinomos.
- Duomenims klasifikuoti taikomi **įvairūs klasifikavimo metodai**: Naive Bayes, k artimiausių kaimynų, atraminių vektorių, klasifikavimo medžių ir kt.
- Dirbtiniai neuroniniai tinklai taip pat yra plačiai **naudojami duomenims klasifikuoti**.
- Net **vienas neuronas geba** spręsti nesudėtingus klasifikavimo uždavinius.

Duomenys klasifikavimui

Sprendžiant **klasifikavimo** uždavinius išskiriami **trijų tipų duomenys**:

- **mokymo duomenys** naudojami klasifikatoriui sukurti,
- **testavimo duomenys** naudojami patikrinti (testuoto) klasifikatoriaus išmokymo klasifikuoti lygį,
- **nauji duomenys**, kurių klasės nėra žinomas, bet taikant sukurtą klasifikatorių jos yra nustatomos.

Klasifikavimo tikslumo matai

- Klasifikatorius **turi būti išmokytas** taip, kad gebėtų gerai klasifikuoti duomenys, kurių klasės nėra žinomos.
- Vadinasi reikia turėti to **išmokymo įvertinimo matus**.
- Klasifikavimo tikslumui nustatyti dažniausiai vertinami šie matai:
  - **jautrumas** (angl. *sensitivity*);
  - **specifiškumas** (angl. *specificity*);
  - **bendras klasifikavimo tikslumas** (angl. *accuracy*).

Klasifikavimo tikslumas

Apibrėžkime pagrindines sąvokas:

- **tikrai teigiamas** (TT) (angl. *true positive*) – objektas X_i priskirtas klasei C_j, ir iš tiesų jis jai priklauso,
- **tikrai neigiamas** (TN) (angl. *true negative*) – objektas X_i nepriskirtas klasei C_j, ir iš tiesų jis jai nepriklauso;
- **klaidingai teigiamas** (KT) (angl. *false positive*) – objektas X_i priskirtas klasei C_j, bet iš tiesų jis jai nepriklauso;
- **klaidingai neigiamas** (KN) (angl. *false negative*) – objektas X_i nepriskirtas klasei C_j, bet iš tiesų jis jai priklauso.

Klasifikavimo matrica

Apskaičiavus šiuos įverčius, sudaroma **klasifikavimo matrica** (angl. *classification* ar *confusion matrix*)

		gauta klasė	
		C_1	C_2
tikroji klasė	C_1	tikrai teigiamas (TT)	klaidingai teigiamas (KT)
	C_2	klaidingai neigiamas (KN)	tikrai neigiamas (TN)

Jautrumas ir specifiškumas

Klasifikavimo matų reikšmės yra apskaičiuojamos pagal šias formules:

$$\text{jautrumas} = \frac{\text{TT skaičius}}{\text{TT skaičius+KN skaičius}},$$

$$\text{specifiškumas} = \frac{\text{TN skaičius}}{\text{TN skaičius+KT skaičius}},$$

$$\text{bendras klasifikavimo tikslumas} = \frac{\text{TT skaičius+TN skaičius}}{\text{visų objektų skaičius}}.$$

Kryžminė patikra

- Klasifikavimo tikslumas gali priklausyti nuo to, kaip visa **duomenų aibė padalinta į mokymo ir testavimo** aibes.
- Todėl tikslinga **klasifikavimą atlikti keliems skirtingiems** tos pačios duomenų aibės mokymo ir testavimo rinkiniams ir **įvertinti vidutinį** klasifikavimo tikslumą.
- Tam tikslui dažnai naudojamas **kryžminės patikros metodas** (angl. *cross validation*).

Kryžminė patikra

- Kryžminės patikros metu duomenų aibė yra **suskaidoma** į *q* **nesusikertančių blokų** (angl. *folds*).
- Klasifikavimo algoritmas yra **apmokomas** naudojant *q* – 1 bloko duomenis, o likusi duomenų dalis yra panaudojama algoritmui **testuoti**.
- **Fiksuojamos** klasifikavimo matų reikšmės.
- Ši procedūra atliekama *q* **kartų**, mokymui imant vis kitus *q* – 1 blokus, pabaigoje randamos klasifikavimo matų **vidutinės reikšmės**. Pagal jas vertinamas **sukurto klasifikatoriaus tikslumas**.