Informácie k projektom pre predmet „Analýza zhlukov a klasifikácia dát“

Radoslav Harman, KAMŠ FMFI UK

Vaše hodnotenie sa bude zakladať na prezentácii projektu a na jednej otázke z teórie. Tento text sa týka iba projektovej časti hodnotenia. Projekt je potrebné poslať vo formáte PDF na moju e-mailovú adresu harman@fmph.uniba.sk, najneskôr do 8:00 v deň Vašej skúšky. Projekt môže byť napísaný v slovenskom alebo anglickom jazyku.

V projekte sa snažte uviesť všetky dôležité informácie, najmä:

1) Vaše meno, dátum a informatívny názov projektu;
2) zdroj dát a opis objektov aj premenných;
3) cieľ analýzy zvoleného dátového súboru;
4) opis aplikácie aspoň dvoch rôznych metód. Za rôzne metódy považujem najmä:
   a) k-means alebo k-medoids,
   b) DBSCAN alebo príbuznú metódu OPTICS,
   c) zhlukovanie založené na normálnom modeli,
   d) hierarchickú analýzu zhlukov,
   e) lineárnu alebo kvadratickú diskriminačnú analýzu,
   f) metódu k najbližších susedov,
   g) klasifikačné stromy alebo lesy,
   h) kombináciu klasifikátorov pomocou boostingu,
   i) support vector machines,
   j) inú metódu schválenú vyučujúcim.

   Nie je potrebné podrobne písať všeobecnú teóriu zvolených metód; stačí niekoľkými vetami opísať ich základnú myšlienku.
5) opis výsledkov vo forme číselných údajov, grafov a plnovýznamových viet;
6) možnú interpretáciu výsledkov pre potreby aplikačnej oblasti, z ktorej dáta pochádzajú;
7) zhodnotenie získaných výsledkov s explicitným poukázaním na problematické aspekty zvoleného prístupu a interpretácie.

Na projekt nekladiem žiadne špeciálne formálne požiadavky. Snažte sa však písať štýlom, ktorý zodpovedá pravidlám a odporúčaniam uvedeným v texte `http://www.iam.fmph.uniba.sk/ospm/Harman/DIP.pdf`.

Rozsah projektu by mal byť 6–12 strán. Ak používate iné zdroje ako prednášky z predmetu AZKD, nezabudnite ich v projekte riadne citovať.

Pri príprave projektu je prípustné používať nástroje AI, avšak len na vyhľadávanie pôvodných zdrojov, doplnenie vedomostí a jazykové korekcie. Nie je dovolené používať AI na generovanie projektu po obsahovej alebo myšlienkovej stránke. Držte sa zásady, že AI možno používať iba takým spôsobom, ktorý v konečnom dôsledku prehlbuje alebo rozširuje Vaše pochopenie látky z prednášok, nie takým, ktorý Vám umožní toto pochopenie obísť.

Projekt budete prezentovať osobne, buď na vlastnom notebooku, alebo na mojom počítači; budete si môcť vybrať. Môžete byť tiež požiadaní, aby ste niektoré myšlienky vysvetlili na tabuli, ktorú mám v kancelárii. Pri hodnotení projektu budem brať do úvahy jeho písomnú podobu, ako aj kvalitu ústnej prezentácie, napríklad vecnú správnosť, zrozumiteľnosť a plynulosť vyjadrovania.

Dáta

Náročnou súčasťou práce na projekte je získanie vhodných dát. Dátový súbor by mal obsahovať aspoň 20 objektov, pri ktorých meriame minimálne tri premenné, teda príznaky alebo prediktory; viac premenných je spravidla vhodnejšie. Ak analyzujeme maticu nepodobností, nemusíme mať, samozrejme, k dispozícii žiadne explicitné premenné.

Nemôžete používať dáta, ktoré sme analyzovali na prednáške alebo cvičeniach. Na internete je však dostupné množstvo zaujímavých dátových súborov, ako aj celých dátových repozitárov, ktoré si môžete vyhľadať pomocou vyhľadávačov. Zároveň si prosím nevyberajte žiadny z nasledujúcich príliš často analyzovaných dátových súborov: Fisherov súbor kosatcov *Iris*, Titanic, Auto MPG, Old Faithful geyser, Mammalian milk composition.

Vyhľadanie vhodného dátového súboru sa Vám môže zdať zdĺhavé, ale aj táto časť práce má svoj pedagogický účel. Pri hľadaní dát budete uvažovať nad rôznymi typmi dát a dátových súborov a zároveň si urobíte prehľad o dátach dostupných na webe.

Dátový súbor si môžete vytvoriť aj sami, napríklad pomocou online dotazníka s vhodne zvolenými premennými, ktorý vyplnia Vaši známi. Keďže nejde o seriózny vedecký výskum, kvalitu ani dizajn zberu dát nebudem hodnotiť prísne; potenciálnych nedostatkov takto získaných dát by ste si však mali byť vedomí. Zaujímavé dáta sa dajú vygenerovať aj nástrojmi umelej inteligencie. Ak však chcete používať takéto alebo iné syntetické dáta, napíšte mi prosím v dostatočnom predstihu e-mail, aby sme to mohli prediskutovať.

Inou možnosťou je reprodukovať analýzu dát, prípadne jej časť, z Vami zvoleného vedeckého článku, ktorý používa niektorú z vyššie uvedených metód. V takomto prípade stačí použiť jednu metódu. Aj v tomto prípade mi však vopred napíšte, aby som sa mohol bližšie pozrieť na to, čo zamýšľate robiť.

Upozorňujem tiež, že niekedy si študenti zvolia problematiku a dáta, ktoré sú vzhľadom na dostupný čas a ich aktuálne vedomosti príliš náročné. Napríklad si študent vyberie veľmi komplexné dáta týkajúce sa génovej expresie, no pritom nevie, čo je gén, ako prebieha expresia génov, ani čo sa vlastne meralo. V takejto situácii je zmysluplná analýza takmer nemožná.

Svojim dátam musíte rozumieť. Vhodné je napríklad vybrať si dáta z oblasti Vášho užšieho záujmu, napríklad z témy diplomovej práce, obľúbenej vedeckej disciplíny, koníčka a podobne.

Prajem veľa úspechov,
RH