

Cvičení 7

Úvod do statistické analýzy dat

Jan Příkryl

ČVUT FD

5. března 2018

Problém 1

U každé části (a) až (d) uveďte, zda bychom obecně očekávali, že výkon neparametrické statistické metody učení bude lepší nebo horší, než u parametrické metody. Odůvodněte odpověď.

- (a) Velikost vzorku n je extrémně velká a počet prediktorů p je malý.
- (b) Počet prediktorů p je extrémně velký a počet pozorování n je malý.
- (c) Vztah mezi prediktory a odpovědí je velmi nelineární.
- (d) Rozptyl chybových členů, tj. $\sigma^2 = \text{var}(x)$, je extrémně vysoký.

Problém 2

Vysvětlete, zda daný scénář reprezentuje klasifikační nebo regresní problém a uveďte, zda nás více zajímá inference nebo predikce.

Uveďte n a p .

- (a) Shromáždili jsme soubor údajů o 500 nejlepších firmách v USA. Pro každou firmu jsme zaznamenali zisk, počet zaměstnanců, průmysl a mzdu generálního ředitele. Chceme pochopit, jaké faktory ovlivňují výši platu generálního ředitele.

Problém 2

Vysvětlete, zda daný scénář reprezentuje klasifikační nebo regresní problém a uveďte, zda nás více zajímá inference nebo predikce.

Uveďte n a p .

- (a) Shromáždili jsme soubor údajů o 500 nejlepších firmách v USA. Pro každou firmu jsme zaznamenali zisk, počet zaměstnanců, průmysl a mzdu generálního ředitele. Chceme pochopit, jaké faktory ovlivňují výši platu generálního ředitele.
- (b) Uvažujeme o uvedení nového produktu na trh a přejeme si vědět, zda bude úspěšný nebo ne. Shromáždíme údaje o průběhu uvedení na trh u 20 podobných produktů. Pro každý produkt jsme zaznamenáme, zda byl či nebyl úspěšný, cenu účtovanou za produkt, marketingový rozpočet, konkurenční cenu a deset dalších proměnných.

Problém 2

Vysvětlete, zda daný scénář reprezentuje klasifikační nebo regresní problém a uveďte, zda nás více zajímá inference nebo predikce.

Uveďte n a p .

- (a) Shromáždili jsme soubor údajů o 500 nejlepších firmách v USA. Pro každou firmu jsme zaznamenali zisk, počet zaměstnanců, průmysl a mzdu generálního ředitele. Chceme pochopit, jaké faktory ovlivňují výši platu generálního ředitele.
- (b) Uvažujeme o uvedení nového produktu na trh a přejeme si vědět, zda bude úspěšný nebo ne. Shromáždíme údaje o průběhu uvedení na trh u 20 podobných produktů. Pro každý produkt jsme zaznamenáme, zda byl či nebyl úspěšný, cenu účtovanou za produkt, marketingový rozpočet, konkurenční cenu a deset dalších proměnných.
- (c) Zajímá nás předpověď procentuální změny kurzu amerického dolaru ve vztahu k týdenním změnám na světových akciových trzích. Z tohoto důvodu v každém týdnu roku 2012 zaznamenáváme procentuální změnu kurzu dolaru, procentuální změnu na americkém trhu, procentuální změnu na britském trhu a procentuální změnu na

Problém 3

Vraťme se nyní k rozkladu chyby na vychýlení a rozptyl.

(a) Načrtněte závislost typického

- ▶ (kvadrátu) zkreslení,
- ▶ odchylky,
- ▶ trénovací chyby,
- ▶ testovací chyby a
- ▶ Bayesových (nebo neredukovatelných) chybových křivek

na jediném grafu, pokud postupujeme od méně flexibilních metod statistického učení k pružnějším přístupům. Osa x by měla představovat množství flexibility v metodě a osa y by měla představovat hodnoty pro každou křivku. Mělo by existovat pět křivek. Ujistěte se, že každou z nich označíte.

(b) Vysvětlete, proč každá z křivek má tvar, zobrazený v části (a).

Problém 4

Nyní se zamyslete nad některými reálnými aplikacemi pro statistické učení

- (a) Popište tři reálné aplikace, v nichž by mohla být užitečná klasifikace. Popište závislou proměnnou, stejně jako prediktory. Je cílem každé aplikace inference nebo predikce? Vysvětlete svoji odpověď.
- (b) Popište tři reálné aplikace, v nichž by mohla být užitečná regrese. Popište závislou proměnnou, stejně jako prediktory. Je cílem každé aplikace inference nebo predikce? Vysvětlete svoji odpověď.
- (c) Popište tři reálné aplikace, v nichž by mohlo být užitečné shlukování.

Problém 5 a 6

Jaké jsou výhody a nevýhody velmi flexibilního (oproti méně flexibilnímu) přístupu k regresi nebo klasifikaci? Za jakých okolností může být preferován flexibilnější přístup než méně flexibilní přístup? Kdy může být preferován méně flexibilní přístup?

Popište rozdíly mezi parametrickým a neparametrickým přístupem ke statistickému učení. Jaké jsou výhody parametrického přístupu k regresi nebo klasifikaci (na rozdíl od neparametrického přístupu)? Jaké jsou jeho nevýhody?

Problém 7

Níže uvedená tabulka obsahuje soubor trénovacích dat obsahující šest pozorování, tři prediktory a jednu kvalitativní cílovou proměnnou.

#	X_1	X_2	X_3	Y
1	0	3	0	červená
2	2	0	0	červená
3	0	1	3	červená
4	0	1	2	zelená
5	-1	0	1	zelená
6	1	1	1	červená

Problém 7 (pokračování)

Předpokládejme, že chceme použít tuto množinu dat k předpovědi Y pomocí metody k nejbližších sousedů (k -NN) v případě, když $X_1 = X_2 = X_3 = 0$.

- (a) Vypočítejte euklidovskou vzdálenost mezi každým pozorováním a zkušebním bodem $X_1 = X_2 = X_3 = 0$.
- (b) Jaká je naše předpověď s $k = 1$? Proč?
- (c) Jaká je naše předpověď s $k = 3$? Proč?
- (d) Je-li Bayesova rozhodovací hranice v tomto problému vysoce nelineární, budeme očekávat, že nejlepší hodnota pro k bude velká nebo malá? Proč?

Vysoké školy v USA

Prozkoumejme nyní datový soubor `islr-college.csv`, obsahující řadu proměnných pro 777 různých univerzit a vysokých škol v USA.

Proměnné jsou

- ▶ **Apps**: počet přihlášek
- ▶ **Accept**: počet přijatých žadatelů
- ▶ **Enroll**: počet imatrikulovaných studentů
- ▶ **Top10perc**: noví studenti z nejlepších 10% na střední škole
- ▶ **Top25perc**: noví studenti z nejlepších 25% na střední škole
- ▶ **F.Undergrad**: počet absolventů na plný úvazek
- ▶ **P.Undergrad**: počet studentů na částečný úvazek
- ▶ **Outstate**: poplatek za studium mimo stát
- ▶ **Room.Board**: náklady na ubytování a stravu
- ▶ **Books**: odhadované náklady na knihy
- ▶ **Personal**: odhadované osobní výdaje
- ▶ **PhD**: procento pedagogů s Ph.D.
- ▶ **Terminal**: procento fakulty s odborným vzděláním
- ▶ **S.F.Ratio**: počet studentů/pedagogů
- ▶ **perc.alumni**: procento absolventů, kteří darují
- ▶ **Expend**: náklady na výuku jednoho studenta
- ▶ **Grad.Rate**: míra absolvování

Vysoké školy v USA

pokračování

Před načtením dat do Matlabu si je můžete zobrazit v aplikaci Excel nebo v textovém editoru.

- (a) Použijte příkaz `readtable()` pro načtení dat do Matlabu. Zavolejte načtená data `college`. Zkontrolujte, zda máte adresář nastavený na správné umístění dat.
- (b) Podívejte se na data pomocí editoru. Měli byste si všimnout, že první sloupec nazvaný `Var1` je jen název každé univerzity. Nechceme, aby jej Matlab chápal jako údaje, může být ale užitečné mít tato jména později k dispozici. Vyzkoušejte následující příkaz:

```
college.Properties.RowNames = college.Var1;
```

Vysoké školy v USA

pokračování

- (c) Měli byste vidět, že nyní existuje nepojmenovaný očíslovaný sloupec s názvem každé univerzity. To znamená, že Matlab dal každému řádku název odpovídající vysoké škole a nepokusí provádět výpočty na názvech řádků. Musíme však ještě odstranit první sloupec v datech, kde je jména jsou uložena.

```
college.Var1 = [];
```

- (d) Nyní byste měli vidět, že první datový sloupec je **Private**. Všimněte si, že před sloupcem **Private** se objevil další sloupec označený řádkem. Nejedná se však o sloupec dat, ale o název, který Matlab dává každému řádku.

Vysoké školy v USA

pokračování

- (e) Použijte funkci `summary()` pro vytvoření číselného souhrnu proměnných v datové sadě.
- (f) Pomocí funkce `plotmatrix()` vytvořte scatterplotovou matici prvních deseti sloupců nebo proměnných dat. Připomeňme, že na prvních deset sloupců matice `A` se můžete odkázat pomocí `A(:,1:10)`.
- (g) Použijte funkci `boxplot()` pro vytvoření krabicových grafů (boxplotů) `Outstate` versus `Private` umístěných vedle sebe.
- (h) Vytvořte novou kvalitativní proměnnou nazvanou `Elite` tak, že rozdělíte proměnnou `Top10perc` na dva koše. Budeme dělit univerzity do dvou skupin založených na tom, zda podíl studentů pocházejících z top 10% tříd střední školy překročí 50%.

Vysoké školy v USA

pokračování

- (i) Použijte funkci `summary()`, abyste zjistili, kolik elitních univerzit existuje. Nyní použijte funkci `boxplot()` pro vytvoření krabicových grafů (boxplotů) `Outstate` versus `Elite` vedle sebe.
- (j) Použijte funkci `hist()` pro vytvoření některých histogramů s různým počtem košů pro několik kvantitativních proměnných. Příkaz `subplot(2,2,n)` může být užitečný: rozdělí okno tisku do čtyř oblastí tak, aby mohly být vykresleny současně čtyři grafy. Úprava argumentů této funkce rozdělí obrazovku jinými způsoby.
- (k) Pokračujte ve zkoumání dat a poskytněte stručné shrnutí toho, co zjistíte.

Automobily

Další datová sada je `islr_vehicles.csv`, obsahující neúplný soubor dat o vybraných typech osobních automobilů.

- (a) Před dalším zpracováním soubor vyčistěte a odstraňte záznamy s chybějícími hodnotami.
- (b) Který z prediktorů je kvantitativní a kvalitativní?
- (c) Jaký je rozsah každého kvantitativního prediktoru? Můžete odpovědět pomocí funkce `range()`.
- (d) Jaká je střední a standardní odchylka každého kvantitativního prediktoru?
- (e) Nyní odeberte desáté až 85. pozorování. Jaký je rozsah, průměr a standardní odchylka každého prediktoru v podmnožině dat, která zůstává?

Automobily

pokračování

Pokračujeme s grafy:

- (f) Pomocí úplné datové sady prozkoumejte prediktory graficky pomocí scatterplots nebo jiných nástrojů podle vašeho výběru. Vytvořte nějaké grafy, které zvýrazní vztahy mezi prediktory. Komentujte své zjištění.
- (g) Předpokládejme, že bychom chtěli předpovídat dojezd (mpg) na základě ostatních proměnných. Naznačují vaše grafy, že některá z dalších proměnných by mohla být užitečná při předvídání mpg? Odůvodněte odpověď.

Boston

Posledním souborem dat, jenž budeme zkoumat, jsou data o bydlení v Bostonu v 70. letech 20. století, uložená v `boston.csv`.

- (a) Načtěte sadu dat Boston. Kolik řádků je v tomto souboru dat? Kolik sloupců? Co znamenají řádky a sloupce?
- (b) V této množině dat proveďte několik párových scatterplot předpovědí (sloupců). Popište své zjištění.
- (c) Je některý z prediktorů spojený s kriminalitou na obyvatele? Pokud ano, vysvětlete vztah.
- (d) Zdá se, že některá z předměstí Bostonu mají obzvláště vysokou míru kriminality? Daňovou sazbu? Poměr žák-učitel? Komentujte rozsah každého prediktoru.
- (e) Kolik předměstí této datové sady sousedí s řekou Charles?
- (f) Jaký je medián poměru žáků-učitelů mezi městy v tomto datovém souboru?

Boston

pokračování

- (g) Které předměstí Bostonu má nejnižší mediánovou hodnotu domů v soukromém vlastnictví? Jaké jsou hodnoty ostatních prediktorů pro toto předměstí a jak se tyto hodnoty srovnávají s celkovým rozsahem těchto prediktorů? Komentujte své zjištění.
- (h) Kolik předměstí v tomto souboru údajů má průměrně více, než sedm pokojů na obydlí? Více než osm pokojů na jedno obydlí? Okomentujte předměstí, která vykazují průměrně více, než osm pokojů na obydlí.