

11MAMY – Cvičení 7

Úvod do statistické analýzy dat

Jan Příklad

ČVUT FD

14. března 2023

Obsah cvičení

Teorie

Vysoké školy v USA

Data o automobilech

Bydlení v Bostonu

Problém 1

U každé části (a) až (d) uveďte, zda bychom obecně očekávali, že výsledky neparametrické statistické metody učení budou lepší nebo horší, než výsledky 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 (závislou proměnnou) 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 (indukce) nebo predikce (dedukce). Uveďte n a p .

- (a) Shromáždili jsme soubor údajů o 500 nejlepších firmách v USA. Pro každou firmu jsme zaznamenali zisk, počet zaměstnanců, odvětví průmyslu 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 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 2019 sledujeme týdenní procentuální změnu kurzu dolaru a procentuální změny výkonnosti akcií na americkém, britském, německém a japonském trhu.

Problém 3

Vraťme se nyní k rozkladu chyby na zkreslení a rozptyl.

(a) Do jednoho grafu načrtněte typickou závislost

- ▶ kvadrátu zkreslení,
- ▶ rozptylu,
- ▶ kvadrátu trénovací chyby,
- ▶ kvadrátu testovací chyby a
- ▶ kvadrátu Bayesovy chyby (tedy rozptylu neredukovatelné chyby)

v závislosti na flexibilitě modelu. Postupujte 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. Každou z pěti výsledných křivek vhodně označte (barevně, typem čáry).

(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

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?

Problém 6

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, kdy $X_1 = X_2 = X_3 = 0$.

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

#	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

Vysoké školy v USA

Prozkoumejme nyní datový soubor `islr_college.csv`, obsahující řadu proměnných pro 777 různých univerzit a vysokých škol v USA. Jeho atributy 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 studenti 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. Nazvěte 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 vznikl 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 jsou jména uložena:

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

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

Vysoké školy v USA

pokračování

- (e) Použijte příkaz `summary()` pro vytvoření číselného souhrnu proměnných v datové sadě.
- (f) Pomocí příkazu `plotmatrix()` vytvořte scatterplotovou matici prvních deseti sloupců (atributů). Připomeňme, že na prvních deset sloupců matice `A` se můžete odkázat pomocí `A(:,1:10)`, pokud je `A` cell matice, potom `A{: ,1:10}`. Pokud tabulka obsahuje kategorické proměnné, je třeba tyto proměnné napřed převést na numerické hodnoty například pomocí funkce `grp2idx()`.
- (g) Použijte příkaz `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 podle toho, 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 několika histogramů s různým počtem košů pro několik kvantitativních proměnných. Použijte příkaz `subplot(2,2,n)`: rozdělí okno tisku do čtyř oblastí tak, aby mohly být vykresleny současně čtyři grafy. Úprava argumentů této funkce rozdělí obrazovku jinými způsoby.
- (k) Pokračujte v průzkumové analýze dat a stručně shrňte to, co zjistíte.

Automobily

Další datová sada je `islr_auto.csv`, obsahující neúplný soubor dat o vybraných typech osobních automobilů. Soubor má následující atributy:

- ▶ `mpg`: spotřeba vozidla [mi gal^{-1}]
- ▶ `cylinders`: počet válců motoru
- ▶ `displacement`: objem motoru [in^3]
- ▶ `horsepower`: výkon motoru v koních
- ▶ `weight`: hmotnost vozidla v librách
- ▶ `origin`: země původu
- ▶ `name`: název modelu
- ▶ `acceleration`: zrychlení z 0 na 60 mi h^{-1} [s]
- ▶ `year`: rok výroby (poslední dvojčíslí)

Před dalším zpracováním soubor vyčistěte a odstraňte záznamy s chybějícími hodnotami. Určete:

- Který z prediktorů je kvantitativní a který kvalitativní?
- Jaký je rozsah každého kvantitativního prediktoru? Můžete odpovědět pomocí funkce `range()`.

Automobily

pokračování

- (c) Jaký je průměr a standardní chyba každého kvantitativního prediktoru?
- (d) Nyní odeberte desáté až 85. pozorování. Jaký je rozsah, průměr a standardní chyba každého prediktoru v podmnožině dat, která zůstává?

Pokračujeme s grafy:

- (e) 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.
- (f) Předpokládejme, že bychom chtěli modelovat dojezd (**mpg**) na základě ostatních proměnných. Naznačují vaše grafy, že některá z dalších proměnných by mohla být užitečná pro predikci **mpg**? Odůvodněte odpověď.

Boston

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

- ▶ **crim**: míra kriminality na obyvatele podle města
- ▶ **zn**: podíl obytné půdy rozdělené na pozemky o rozloze více než 2300 m²
- ▶ **indus**: podíl maloobchodních ploch na jedno město
- ▶ **chas**: pomocná proměnná – 1, pokud oblast sousedí s řekou Charles, jinak 0
- ▶ **nox**: koncentrace NO_x v PPM/10
- ▶ **rm**: průměrný počet pokojů na obydlí
- ▶ **age**: podíl jednotek obývaných majiteli postavených před rokem 1940
- ▶ **dis**: vážený součet vzdálenosti k pěti zaměstnavatelům v Bostonu
- ▶ **rad**: index přístupnosti lokality z radiální dálnice
- ▶ **tax**: sazba daně z nemovitosti v plné hodnotě v 10000 \$
- ▶ **ptratio**: poměr žáků a učitelů podle města
- ▶ **black**: $1000(Bk - 0,63)^2$, kde Bk je podíl [lidí afrického amerického původu] podle města
- ▶ **lstat**: procento lidí s nižším vzděláním či nižšími příjmy
- ▶ **medv**: medián hodnoty domů obývaných majiteli v 1000 \$

Boston

pokračování

- (a) Načtěte sadu dat Boston do tabulky `bostondata`. Kolik řádků je v tomto souboru dat? Kolik sloupců? Jaké jsou typy jednotlivých atributů?
- (b) V této množině dat proveďte několik párových scatterplot předpovědí pro vhodně vybrané atributy. Popište svá zjištění.
- (c) Je patrná vazba některého z ostatních atributů na hodnotu `crim` (kriminalitu 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 medián hodnoty 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? Na základě ostatních parametrů okomentujte vlastnosti předměstí, která vykazují v průměru více, než osm pokojů na obydlí.