Chybějící a odlehlé hodnoty; odstranění odlehlých hodnot pomocí algoritmu k-means

Návod ke druhému cvičení

Matěj Holec, holecmat@fel.cvut.cz

ZS 2011/2012

Úvod

Cílem cvičení je připomenout důležitost předzpracování dat. Konkrétně se seznámíte se základními metodami manipulace s odlehlými a chybějícími hodnotami.

1 Chybějící hodnoty

Výskyt chybějících hodnot mívá různé příčiny. Může to být například rozbitý senzor vedoucí k chybějícím hodnotám nebo lidé, kteří odmítli či opomněli vyplnit dotazník, nebo zkrátka nějaký atribut nemá smysl pro jisté druhy objektů (např. atribut *gravidita* nemá pro samce význam).

Je velice důležité identifikovat typ „skryté“ chybějící hodnoty, neboť analýza jinak může vést k nesmyslným závěrům. Nejčastěji se u chybějících hodnot předpokládá, že data chybějí zcela náhodně (*Missing Completely At Random*). Takovou chybu si lze představit jako náhodnou kaňku na papíře s vytištěnými daty. Naopak chybový model *Missing At Random* pracuje s předpokladem, že pravděpodobnost chybějící hodnoty závisí na nějaké jiné proměnné. Teploměr vrací chybové hodnoty s pravděpodobností $p > 0$ pouze při dešti, jinak funguje správně. Nejhorším případem jsou neignorovatelné chybějící hodnoty (*Nonignorable Missing Values*). Příkladem této situace může být teploměr, který kvůli poškození neměří teploty nižší než 0°C.

Imputace

Chybějící hodnoty nelze vždy vyloučit. Nejjednodušší způsob je nahradit chybějící hodnotu střední hodnotou u numerického atributu, nebo módem u atributu kategorického. Další možností je použít k nahrazení hodnoty záznam od nejbližšího souseda, tedy záznam, který je nejpodobnější tomu s chybějící hodnotou (např. podle Eukleidovské metriky na zbývajících atributech).

2 Normalizace dat

Normalizací pro účely zpracování dat se zpravidla rozumí postup, kterým se snažíme eliminovat nestejný vliv nebo důležitost proměnných. Typickými přístupy jsou min-max normalizace a z-skóre standardizace.

- **min-max normalizace:** Pro numerický atribut x kde min_x a max_x je minimální a maximální hodnota vzorku je min-max normalizace definována následovně:
$$x \mapsto \frac{x - min_x}{max_x - min_x}$$
- **Z-skóre standardizace:** Pro numerický atribut x kde $\hat{\mu}_x$ a $\hat{\sigma}_x$ jsou střední hodnota a odhad standardní odchylky vzorku je Z-skóre standardizace definována následovně:
$$x \mapsto \frac{x - \hat{\mu}_x}{\hat{\sigma}_x}$$

3 Odlehlé hodnoty

Odlehlé hodnoty bývají v zásadě způsobené chybným měřením nebo překlepem. Taková chybná data, nelze-li provést nápravu, je třeba z data setu vyloučit. Nicméně se může stát, že nějaká odlehlá hodnota je ve své podstatě hodnotou správnou a liší se výrazně od ostatních pouze náhodou, a nebo třeba náleží k nějaké výjimečné situaci. Přesto, že takové hodnoty jsou správné, může (ale samozřejmě nemusí) být vhodné je dočasně vyloučit z analýzy a zpracovávat je odděleně. Některé analytické metody jsou robustní proti takovým odlehlým hodnotám, zatímco jiné mohou být značně citlivé a i přítomnost jedné odlehlé hodnoty může vést k nesmyslným závěrům.

Odlehlé hodnoty silně závisí na distribuční funkci, ze které byla data navzorkovaná. Při detekci takových hodnot pomocí jediného atributu se u kategorických dat vychází z toho, že se odlehlé hodnoty vyskytují s extrémně malou frekvencí. U numerických atributů je situace složitější. Problém spočívá ve faktu, že u rozsáhlejších datových souborů vždy budou nějaké body označené jako odlehlé.

Univariátní odlehlé hodnoty

Nechť $\overline{x}$ je aritmetický průměr a s je standardní odchylka náhodného rozdělení dat. Pozorování je deklarováno jako odlehlé, pokud leží mimo interval

$$(\overline{x} - ks, \overline{x} + ks)$$

kde k je zpravidla 2 nebo 3. Za předpokladu normálního rozdělení tento interval odpovídá 95.45%, resp. 99.75% dat. Tento předpoklad bohužel často nebývá zcela splněn, navíc výpočty aritmetického průměru a standardní odchylky jsou vysoce citlivé právě na odlehlé hodnoty.

Multivariátní odlehlé hodnoty

U vícerozměrných dat se obvykle nevychází z žádných specifických předpokladů o distribuci. K detekci odlehlých hodnot lze užít techniky redukující dimenzi dat a pokusit se identifikovat odlehlé hodnoty v bodovém grafu, nebo přístupy založené na shlukování, kde jsou jako odlehlé hodnoty označeny takové hodnoty, které nemohou být smysluplně přiřazeny do žádného shluku. Příkladem tohoto přístupu je algoritmus COR [1].

Algoritmus COR

IN		**OUT**	
X	data	X	data without outliers
m	number of clusters		
th	threshold		
R	number of iterations		

$i \leftarrow 0$
repeat
  $i \leftarrow i + 1$
  $len \leftarrow numOfSamples(X)$
  $P_{new}, C_{new} \leftarrow k\_means(X, m)$
  {BEGIN: outlier removal}
  $X_{new} \leftarrow X$
  for cluster $c \subseteq X_{new}$ **do**
    if $numOfSamples(c) > 1$ **then**
      $s_{max} \leftarrow maxDistantSampleFromCentroid(c)$
      $s_{min} \leftarrow minDistantSampleFromCentroid(c)$
      $distortion \leftarrow distanceFromCentroid(s_{min}, c)/distanceFromCentroid(s_{max}, c)$
      if $distortion < th$ **then**
        $P_{new}, C_{new}, X_{new} \leftarrow removeSample(s_{max}, X_{new})$
      end if
    else
      $P_{new}, C_{new}, X_{new} \leftarrow removeCluster(c, X_{new})$
      $m \leftarrow m - 1$
    end if
  end for
  {END: outlier removal}
  $len_{new} \leftarrow numOfSamples(X_{new})$
  $X \leftarrow X_{new}$
until $i > R \wedge len_{new} == len$
$return(X)$

Zadání práce

Vypracujte níže uvedené úkoly do předpřipraveného souboru `missing_and_outlier_values.m`. Místa pro Vaše řešení jsou v kódu vyznačena.

Úkoly

1. Nahraďte chybějící hodnoty metodou nejbližšího souseda.

2. Data z předchozího bodu normalizujte pomocí Z-skóre standardizace.
   Na bodových grafech diskutujte vliv standardizace na PCA redukci.

3. V datech `outlier_data.mat` najděte a odstraňte odlehlé objekty. K tomuto kroku využijte metodu COR.

Očekávaný výstup

Ad 1. Nahrazení chybějících hodnot podle nejbližšího souseda by mělo na cvičných datech vykazovat mnohem více vizuálně konzistentní výsledky. Nevýhodou tohoto přístupu může být *obecně* nežádoucí vliv na umístění těžiště dat.

Ad 2. Některé algoritmy pro analýzu dat jsou citlivé na relativní měřítko vstupních proměnných. Pokud například máme dvě proměnné, které mají stejnou varianci a jsou pozitivně korelované, PCA nalezne transformaci, které bude natáčet původní prostor o 45°. Pokud ale například vynásobíme všechny hodnoty v první proměnné stokrát, pak tranformace bude respektovat téměř výhradně první proměnnou s malým příspěvkem druhé nepozměněné proměnné. Natočení proto bude minimální. Jako výsledek úlohy tedy můžeme očekávat vizuálně viditelnou eliminaci nestejného vlivu proměnných.

Ad 3. Cvičná data pro detekci odlehlých hodnot obsahují pět uměle zavedených odlehlých objektů. Tři objekty je možno dobře detekovat univariátními metodami (viz příklad 2). Další dva objekty jsou sice patrně viditelné, ale není možno je odstranit univariátním přístupem. U vícerozměrných dat důležitost multivariátních metod narůstá.

Reference

[1] Svetlana Cherednichenko. Outlier detection in clustering, 2005.