1 Generování pseudo-náhodných čísel

Generování pseudo-náhodných čísel (PRNG, *Pseudo-Random Number Generation*) je proces vytváření sekvencí čísel, které se zdají být náhodné, ale ve skutečnosti jsou deterministicky vytvořené na základě počáteční hodnoty (tzv. semínka; anglicky seed).

Klíčové vlastnosti PRNG

- **Deterministický proces**: pokud známe semínko a algoritmus, můžeme přesně replikovat stejnou sekvenci čísel.
- **Periodičnost**: po určitém počtu generovaných čísel se sekvence začne opakovat.
- **Rychlost a efektivita**: PRNG je optimalizované pro rychlé generování čísel.
- **Přibližná náhodnost**: výstupy by měly být dostatečně náhodné pro mnoho aplikací, ale nejsou vhodné pro kryptografické účely.

Jak PRNG funguje

Většina PRNG je založena na matematických algoritmech, které postupně transformují semínko pomocí aritmetických operací.

1. **Inicializace**: algoritmus začíná semínkem (počáteční hodnotou).
   - Semínko může být pevně dané, nebo dynamicky generované (například podle systémového času).
2. **Rekurzivní výpočet**: každé nové číslo je vypočítáno na základě předchozího čísla pomocí vzorce:
   $$X_{n+1} = f(X_n)$$
   kde f je deterministická funkce.
3. **Výstupní transformace**: vygenerované číslo je obvykle upraveno tak, aby spadalo do požadovaného intervalu (například mezi 0 a 1).

Příklady algoritmů

- **Lineární kongruenční generátor (LCG)**:
  $$X_{n+1} = (aX_n + c) \mod m,$$
  kde m je modul určující maximální periodu, a je tzv. multiplikátor a c je tzv. přírůstek. Tento jednoduchý algoritmus je historicky významný, ale pro moderní použití už není příliš vhodný.
- **Mersenne Twister**:
  - Velmi populární algoritmus s dlouhou periodou ($2^{19937} - 1$).
  - Používá se v mnoha programovacích jazycích (např. Python, MATLAB).
- **XORShift**:
  - Efektivní algoritmus založený na bitových operacích XOR a posunech.

Použití PRNG

- **Simulace**: např. Monte Carlo metody (viz dále).
- **Generování dat**: náhodná čísla pro hry, testovací data.
- **Statistické vzorkování**: náhodné výběry z datových sad.
- **Kryptografie**: jen v případě, že se použije kryptograficky bezpečný RNG.

2 Metoda Monte Carlo

Metoda Monte Carlo je numerická metoda, která využívá náhodné vzorkování k řešení problémů, které mohou být příliš složité nebo nepraktické řešit analytickými nebo deterministickými postupy. Je pojmenována po kasinu v Monte Carlu kvůli své závislosti na náhodnosti. Metoda Monte Carlo je široce používaná v různých oblastech vědy a inženýrství díky své flexibilitě a jednoduchosti, přestože může být výpočetně náročná.

Metoda Monte Carlo funguje na základě simulace velkého počtu náhodných vzorků k odhadu výsledků. Tato metoda je vhodná například pro numerickou integraci, optimalizaci, statistickou analýzu a simulace.

Postup metody Monte Carlo

1. **Definování problému:**
   - Formulujeme problém tak, aby řešení bylo možné vyjádřit jako matematickou funkci, kterou lze vyhodnocovat na základě náhodných vstupů.
2. **Generování náhodných čísel:**
   - Vygenerujeme náhodné vstupy (vzorky) z definované distribuční funkce nebo náhodného prostoru. Například pro jednotkový interval $[0, 1]$ nebo rovnoměrně v nějaké oblasti.
3. **Vyhodnocení funkce:**
   - Pro každý náhodný vzorek vypočítáme hodnotu funkce, kterou zkoumáme.
4. **Agregace výsledků:**
   - Získané výsledky zprůměrujeme, což vede k odhadu řešení. Pro větší přesnost je potřeba velký počet vzorků.
5. **Vyhodnocení přesnosti:**
   - Odhadujeme chybu nebo nejistotu na základě rozptylu výsledků.

Příklady použití

- **Numerická integrace:**
  - Metoda Monte Carlo je často používána pro výpočet integrálů (třeba i vícerozměrných). Pokud máme například odhadnout integrál:
    $$I = \int_a^b f(x)\,dx,$$
    vytvoříme N náhodných bodů x_i v intervalu $[a, b]$, vypočítáme hodnoty $f(x_i)$ a integrál odhadneme jako:
    $$I \approx \frac{b-a}{N} \sum_{i=1}^{N} f(x_i).$$
- **Simulace fyzikálních jevů:**
  - Používá se například v termodynamice, kde je potřeba modelovat chování systému velkého počtu částic, nebo v analýze radiace.
- **Finanční modelování:**
  - Odhaduje se například budoucí hodnota portfolia na základě náhodných scénářů vývoje trhu.
- **Optimalizace:**
  - Například hledání minim nebo maxim funkcí v rozsáhlých nebo složitých prostorech.

- **Simulace náhodných procesů:**
  - Používá se pro modelování systémů zahrnujících pravděpodobnost, jako je pohyb částic nebo šíření epidemií.

Výhody MMC:

- Vhodné pro problémy, které jsou analyticky neřešitelné.
- Snadno se aplikuje na vícedimenzionální problémy.
- Relativně snadná implementace.

Nevýhody MMC:

- Požaduje velký počet iterací (vzorků) pro dosažení přesného výsledku.
- Výpočetně náročné pro složité problémy, typicky pro vícedimenzionální problémy.

Přesnost metody

Přesnost odhadu roste s počtem vzorků N a je úměrná $\frac{1}{\sqrt{N}}$. To znamená, že zdvojnásobení přesnosti vyžaduje čtyřnásobný počet iterací.

3 Explorativní analýza dat

Explorativní analýza dat (**Exploratory Data Analysis**, EDA) je přístup k analýze dat, který se zaměřuje na jejich prozkoumání, pochopení a sumarizaci pomocí vizualizačních a statistických metod. Cílem je získat přehled o struktuře dat, identifikovat vzorce, vztahy a potenciální problémy, které by mohly ovlivnit další analýzy nebo modelování.

Klíčové aspekty EDA

1. **Popisná analýza**:
   - Sumarizace hlavních charakteristik dat.
   - Výpočet základních statistik, jako jsou:
     - Průměr, medián, rozptyl, směrodatná odchylka.
     - Minimální, maximální hodnoty a kvartily.
2. **Identifikace anomálií**:
   - Hledání chyb v datech (např. chybějící hodnoty, odlehlé hodnoty, nesprávné formáty).
3. **Vizualizace dat**: využití grafů a diagramů pro identifikaci vztahů a rozložení:
   - Histogramy (pro zobrazení distribuce).
     Histogram je grafické znázornění distribuce dat pomocí sloupcového grafu se sloupci stejné šířky, vyjadřující šířku intervalů (tříd), přičemž výška sloupců vyjadřuje četnost sledované veličiny v daném intervalu.
   - Box ploty (pro identifikaci odlehlých hodnot).
     Krabicový graf či krabicový diagram je jeden ze způsobů grafické vizualizace numerických dat pomocí jejich kvartilů. Střední „krabicová“ část diagramu je shora ohraničena 3. kvartilem, zespodu 1. kvartilem a mezi nimi se nachází linie vymezující medián. Boxploty mohou obsahovat také linie vycházející ze střední části diagramu kolmo nahoru a dolů, tzv. vousy, vyjadřující variabilitu dat pod prvním a nad třetím kvartilem. Odlehlé hodnoty, tzv. outliery, pak mohou být vykresleny jako jednotlivé body.
   - Scatter ploty (pro vizualizaci vztahů mezi proměnnými).
     Korelační diagram nebo též bodový graf je matematický graf, který zobrazuje v kartézských souřadnicích hodnoty dvou proměnných. Data jsou znázorněna jako množina bodů, jejichž umístění na vodorovné ose udává hodnota první proměnné a umístění na svislé ose hodnota druhé proměnné.
   - Heatmapy (pro zobrazení korelace mezi více proměnnými).
     Teplotní mapa je grafické zobrazení dat, ve kterém je každá hodnota reprezentována barvou určitého spojitého barevného spektra.
4. **Zkoumání vztahů**:
   - Identifikace korelací nebo jiných vzorců mezi proměnnými.
   - Analýza závislostí pomocí statistik, jako je Pearsonův korelační koeficient.
5. **Porozumění struktuře dat**:
   - Zkoumání typů proměnných (kategorické, číselné).
   - Hledání latentních dimenzí nebo skupin v datech (např. pomocí analýzy hlavních komponent nebo s využitím clusteringu). Analýza hlavních komponent se často používá ke snížení dimenze dat s co nejmenší ztrátou informace.

Fáze explorativní analýzy

1. **Předzpracování dat**:
   - Načtení dat, kontrola konzistence a kvality.
   - Ošetření chybějících hodnot, přejmenování atributů, transformace formátů.
2. **Základní přehled**:
   - Prvotní prozkoumání velikosti, typů a distribuce dat.
3. **Identifikace problémů**:
   - Detekce potenciálních problémů, jako jsou odlehlé hodnoty nebo neobvyklé rozložení.
4. **Detailní analýza**:
   - Prohloubení porozumění vybraných vztahů a vlastností.

Výhody EDA:

- Pomáhá pochopit kontext dat, což je klíčové pro další analýzu nebo modelování.
- Umožňuje objevit problémy v datech, které by mohly ovlivnit výsledky.
- Pomáhá objevit neočekávané vztahy nebo trendy.
- Umožňuje identifikovat relevantní proměnné a vztahy pro prediktivní modely.

Nevýhody EDA:

- Není automatická – vyžaduje zkušenosti a intuici analytika.
- Může být časově náročná u rozsáhlých datasetů.

Nástroje a techniky

- **Statistické nástroje:**
  - Python (knihovny jako pandas, NumPy, SciPy, matplotlib, seaborn).
  - R (balíčky jako ggplot2, dplyr).
  - MATLAB, SAS.
- **Specializované platformy:**
  - Tableau, Power BI (pro vizualizace).
  - Jupyter Notebook (pro kombinaci kódu a výsledků).

Příklad

Pokud máme dataset obsahující údaje o prodejích (např. produkt, cena, region, počet prodaných kusů), EDA by mohla zahrnovat:

- Výpočet průměrné ceny a počtu prodaných kusů.
- Histogram cen pro zobrazení jejich rozložení.
- Scatter plot pro zobrazení vztahu mezi cenou a počtem prodaných kusů.
- Heatmapu pro analýzu korelací mezi proměnnými, jako je cena, region a počet prodejů.

4 Regresní analýza

Regresní analýza je statistická metoda sloužící k modelování a analýze vztahů mezi nezávislou proměnnou (prediktorem) a závislou proměnnou (odezvou). Cílem je najít rovnice, které nejlépe popisují tento vztah, a umožňují předpovídat hodnoty závislé proměnné na základě hodnot nezávislých proměnných.

1. Metoda nejmenších čtverců

Metoda nejmenších čtverců je základní technika používaná při regresní analýze k odhadu parametrů regresního modelu. Funguje na principu minimalizace součtu čtverců reziduí, což jsou rozdíly mezi skutečnými hodnotami a hodnotami předpovězenými modelem.

Předpokládejme, že máme data (x_i, y_i) pro $i = 1, 2, \ldots, n$. Hledáme regresní model ve tvaru $y = f(x)$, kde $f(x)$ je funkce vztahu mezi x a y (např. lineární nebo kvadratická).

Podle metody nejmenších čtverců je cílem minimalizovat:

$$S = \sum_{i=1}^{n} \left(y_i - f(x_i)\right)^2.$$

2. Lineární regrese

Lineární regrese modeluje vztah mezi dvěma proměnnými pomocí přímky. Rovnice má tvar:

$$y = \beta_0 + \beta_1 x,$$

kde y je závislá proměnná, x je nezávislá proměnná, β_0 je průsečík s osou y, β_1 je směrnice (sklon přímky).

Postup odhadu:

1. Sestavíme funkci ztráty (součet čtverců reziduí):

$$S = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i)\right)^2.$$

2. Najdeme hodnoty β_0 a β_1, které minimalizují S.

Řešení:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}, \quad \beta_0 = \bar{y} - \beta_1 \bar{x},$$

kde $\bar{x}$ a $\bar{y}$ jsou průměry hodnot x a y.

3. Kvadratická regrese

Kvadratická regrese je rozšířením lineární regrese a modeluje vztah mezi proměnnými pomocí paraboly. Rovnice má tvar:

$$y = \beta_0 + \beta_1 x + \beta_2 x^2,$$

kde β_2 je koeficient kvadratického členu.

Postup odhadu:

1. Sestavíme funkci ztráty:

$$S = \sum_{i=1}^{n} \left(y_i - (\beta_0 + \beta_1 x_i + \beta_2 x_i^2)\right)^2.$$

2. Pomocí metody nejmenších čtverců odhadneme koeficienty $\beta_0, \beta_1, \beta_2$.

Kvadratická regrese je vhodná, pokud vztah mezi proměnnými vykazuje nelineární chování, například když má závislost konvexní nebo konkávní tvar.

Příklady využití regresní analýzy

- **Ekonomie**: předpověď prodejů na základě ceny.
- **Biologie**: vztah mezi dávkou léku a účinkem.
- **Strojírenství**: závislost pevnosti materiálu na teplotě.
- **Finance**: analýza vztahu mezi rizikem a návratností investic.

Regresní analýza je důležitý nástroj pro pochopení a kvantifikaci vztahů mezi proměnnými. Lineární regrese je vhodná pro jednoduché lineární vztahy, zatímco kvadratická (případně kubická atd.) regrese umožňuje modelovat složitější nelineární vztahy.

Porovnání lineární a kvadratické regrese

Vlastnost	**Lineární regrese**	**Kvadratická regrese**
Tvar modelu	Přímka ($y = \beta_0 + \beta_1 x$)	Parabola ($y = \beta_0 + \beta_1 x + \beta_2 x^2$)
Použití	Lineární vztah x a y	Nelineární vztah x a y
Počet parametrů	2 (β_0, β_1)	3 ($\beta_0, \beta_1, \beta_2$)
Výpočetní náročnost	Nižší	Vyšší

Tabulka 1: Porovnání lineární a kvadratické regrese.

Příklad: lineární regrese

Máme následující data, která popisují vztah mezi proměnnými x (nezávislá proměnná) a y (závislá proměnná):

x	y
1	2
2	3
3	5
4	7
5	11

Chceme najít lineární model $y = \beta_0 + \beta_1 x$, který co nejlépe popisuje závislost y na x.

Postup

1. Vzorce

Metoda nejmenších čtverců minimalizuje součet čtverců reziduí:

$$S = \sum_{i=1}^{n} (y_i - (\beta_0 + \beta_1 x_i))^2.$$

Koeficienty β_1 a β_0 se vypočítají následovně:

$$\beta_1 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n} (x_i - \bar{x})^2},$$

$$\beta_0 = \bar{y} - \beta_1 \bar{x},$$

kde $\bar{x}$ a $\bar{y}$ jsou aritmetické průměry hodnot x a y.

2. Výpočty

Vypočítáme střední hodnoty x a y:

$$\bar{x} = \frac{1+2+3+4+5}{5} = 3, \quad \bar{y} = \frac{2+3+5+7+11}{5} = 5,6.$$

Spočítáme jednotlivé složky potřebné pro výpočet β_1:

$$\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}) = (1-3)(2-5,6) + (2-3)(3-5,6) + (3-3)(5-5,6) + (4-3)(7-5,6) + (5-3)(11-5,6)$$

$$= (-2)\cdot(-3,6) + (-1)\cdot(-2,6) + 0\cdot(-0,6) + 1\cdot(1,4) + 2\cdot(5,4) = 7,2 + 2,6 + 0 + 1,4 + 10,8 = 22,0.$$

$$\sum_{i=1}^{n}(x_i - \bar{x})^2 = (1-3)^2 + (2-3)^2 + (3-3)^2 + (4-3)^2 + (5-3)^2 = 4 + 1 + 0 + 1 + 4 = 10.$$

Koeficient β_1 je tedy:

$$\beta_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} = \frac{22,0}{10} = 2,2.$$

Vypočítáme β_0:

$$\beta_0 = \bar{y} - \beta_1\bar{x} = 5,6 - 2,2\cdot 3 = -1,0.$$

3. Výsledný model

Rovnice přímky je $y = -1,0 + 2,2x$.

4. Interpretace

Model předpovídá, že hodnota y roste o $2,2$ jednotky při každém zvýšení x o 1. Hodnota y, pokud $x = 0$, je $-1,0$.

5. Predikce

Například pro $x = 6$ předpovídáme: $y = -1,0 + 2,2\cdot 6 = 12,2$.

Příklad: kvadratická regrese – odhad rychlosti tělesa

Orbitální stanice naměřila v pěti po sobě jdoucích dnech, ve stejnou hodinu následující rychlosti neznámého vesmírného tělesa (v km/s): 10, 11,4, 13,1, 15,8 a 18,7. Odhadněte rychlost tělesa desátého dne.

Řešení: Máme data z měření rychlosti vesmírného tělesa (y) v pěti po sobě jdoucích dnech (x):

x	y (km/s)
1	10,0
2	11,4
3	13,1
4	15,8
5	18,7

Chceme nalézt kvadratický model ve tvaru: $y = \beta_0 + \beta_1 x + \beta_2 x^2$, a pomocí něj odhadnout rychlost y, když $x = 10$.

Vyřešením příslušné soustavy rovnic dostaneme odhady parametrů:

$$\beta_0 \doteq 9,26, \quad \beta_1 \doteq 0,4657, \quad \beta_2 \doteq 0,2857.$$

Výsledný kvadratický model je: $y \doteq 9,26 + 0,4657x + 0,2857x^2$.

Predikci rychlosti pro $x = 10$ získáme snadno přímým dosazením. Rychlost tělesa desátého dne se odhaduje na 42,487 km/s.