

# IBM SPSS NEURAL NETWORKS

## Nové nástroje pro tvorbu prediktivních modelů

Aby mohla vaše organizace zlepšit rozhodování ve všech procesních postupech, potřebuje odhalit vztahy a souvislosti v komplexním a dynamickém prostředí, které ji obklopuje. K tomu vám pomůže statistický software IBM SPSS Statistics včetně jeho přídatných modulů. Jako uživatel systému jistě oceníte jeho všestrannost a výkon. Statistický software IBM SPSS Statistics předčí vaše očekávání.

S modulem IBM SPSS Neural Networks vyhledáte v datech netriviální či skryté vztahy. Objevíte komplexní datové vazby a vybudujete přesnější predikční modely.

Procedury modulu IBM SPSS Neural Networks doplňují tradiční statistické nástroje systému IBM SPSS Statistics. Za pomoci modulu IBM SPSS Neural Networks hledejte v datech nové asociace a pak ověřte jejich statistickou významnost tradičními statistickými postupy.

Modul IBM SPSS Neural Networks můžete používat v rámci desktopové instalace statistického softwaru IBM SPSS Statistics. K dispozici je i jeho serverová verze, jenž se instaluje spolu s IBM SPSS Statistics Base Server. Serverová verze zajistí větší výkon a lepší škálovatelnost.

## Nejdůležitější:

Prozkoumejte skryté vzory ve vašich datech.

- Kvalitnější modely
- Bez nutnosti programování
- Výběr mezi MLP a RBF sítěmi

## Proč využívat neuronové sítě?

Neuronovou síť tvoří množina nelineárních převodníků (neuronů), které vytvářejí vstupní, skryté a výstupní vrstvy sítě. Spojení mezi neurony sousedních vrstev jsou reprezentována číselnými vahami. Váhy určí iterační algoritmus při trénování sítě tak, aby minimalizoval predikční chyby a bylo dosaženo co nejlepších výsledků pro budoucí předpovědi.

Můžete nastavit podmínky, za kterých učení sítě probíhá, a podrobná pravidla pro ukončení iteračního postupu. Také architekturu sítě (tj. uspořádání neuronů) si stanovíte sami nebo vám ji automaticky navrhne program.

Neuronové sítě budete v různých úlohách obvykle kombinovat s jinými statistickými postupy za účelem lepšího porozumění vašim datům. Při výzkumu trhu můžete vytvářet zákaznické profily a zjišťovat preference zákazníků. Při databázovém marketingu budete segmentovat zákaznickou bázi a optimalizovat marketingové kampaně.

Ve finanční sféře využijete neuronové sítě k analýze úvěrové bonity žadatelů o úvěr a vyfiltrujete pravděpodobné úvěrové podvody. Při operační analýze neuronové sítě použijete k řízení finančních toků a plánování logistiky. Aplikace ve vědě a zdravotnictví zahrnují například predikci nákladů na ošetření, audit léčebných postupů nebo predikci doby hospitalizace pacientů.

## Mějte své postupy plně pod kontrolou od začátku do konce

Modul IBM SPSS Neural Networks nabízí procedury Multilayer Perceptron (MLP) a Radial Basis Function (RBF).

Obě procedury se řadí mezi supervizované metody (učení s učitelem), protože mapují vztahy v tréninkových datech s cílem předpovědět vytyčený cíl. Využívají dopředné architektury, kdy se informace v síti šíří jednosměrně od vstupní vrstvy přes skryté vrstvy do vrstvy výstupní.

Výběr procedury závisí na typu dostupných dat a na složitosti vztahů, které chcete pro predikci využít. Zatímco procedura MLP dokáže zmapovat komplexnější vztahy, procedura RBF se učí rychleji.

Ať si zvolíte kteroukoli z procedur, je vhodné před učením sítě data rozdělit do tří skupin: tréninkové, testovací a validační.

Tréninková skupina slouží k nastavení parametrů sítě, pomocí testovací skupiny předcházíme naučení se dat z paměti. Validační skupina slouží k nezávislému odhadu kvality hotového modelu, jenž bude využit k predikci u stávajících i u nových dat.

Závislé (výstupní) proměnné pro neuronové sítě (může jich být i více) mohou být spojité i kategorizované. Je-li závislá proměnná číselná, neuronová síť odhadne hodnotu na základě aproximace určitou převodní spojitou funkcí vstupů. Je-li závislá proměnná kategorizovaná, síť provádí přiřazení do nejpravděpodobnější kategorie (klasifikaci) v závislosti na hodnotách vstupních proměnných.

V procedurách modulu IBM SPSS Neural Networks volíte způsob rozdělení dat, architekturu sítě a postupy, kterými bude síť analyzována. Výsledky se zobrazují v tabulkách a grafech, ukládají se do aktivní datové matice a exportují se do formátu XML pro budoucí využití při skórování nových dat.

## SPECIFIKACE

### Vícevrstvý perceptron

Procedura MLP vytváří specifickou neuronovou síť nazývanou vícevrstvý perceptron. Vícevrstvý perceptron se řadí mezi supervizované sítě s dopřednou architekturou. Může obsahovat více skrytých vrstev. Závislé proměnné mohou být číselné, kategorizované nebo obou typů. Pokud je závislá proměnná číselná, pak neuronová síť realizuje odhad na základě aproximace funkcí mezi vstupy a výstupy. Je-li závislá proměnná kategorizovaná, síť provádí přiřazení do nejpravděpodobnější kategorie (klasifikaci) v závislosti na hodnotách vstupních proměnných.

- *prediktory*
  - *kategorizované*
  - *spojité*
- *Specifikace EXCEPT umožňuje zadat seznam kategorizovaných i spojitých vstupních proměnných, které budou z analýzy vyloučeny. Lze například výhodně použít v situacích, kdy je seznam vstupních proměnných rozsáhlý a je zadán pomocí konstrukcí TO nebo ALL.*
- *Specifikace RESCALE slouží ke změně škály vstupních a výstupních spojitých proměnných.*
  - *transformace spojitých vstupů: standardizace (průměr=0, směrodatná odchylka=1), normalizace (interval 0 až 1), adjustovaná normalizace (interval -1 až 1), bez transformace*
  - *transformace spojitých výstupů: standardizace (průměr=0, směrodatná odchylka=1), normalizace (interval 0 až 1), adjustovaná normalizace (interval -1 až 1), bez transformace*

- **Specifikace PARTITION** určuje metodu rozdělení aktivní datové matice na tréninkovou, testovací a validační množinu. Tréninková množina slouží k nastavení parametrů sítě, testovací množina je určena ke sledování predikční chyby během učení, aby nedošlo k naučení se dat zpaměti, a validační množina slouží k nezávislému odhadu kvality hotového modelu. Volby:
  - poměr (zadáno relativními četnostmi), v jakém budou případy z aktivní datové matice náhodně rozděleny do tréninkové, testovací a validační množiny
  - přiřazení případů do tréninkové, testovací a validační množiny na základě vybrané proměnné
- **Specifikace ARCHITECTURE** určuje architekturu neuronové sítě. Volby:
  - automatický nebo manuální výběr architektury. Další volby platí pro manuální nastavení:
  - počet skrytých vrstev sítě a počty neuronů v nich
  - aktivační funkce všech neuronů ve skrytých vrstvách (hyperbolický tangens nebo logistická funkce)
  - aktivační funkce všech neuronů výstupní vrstvy (identická funkce, hyperbolický tangens, logistická funkce, pravděpodobnostní transformace softmax)
- **Specifikace CRITERIA** určuje způsob využití zdrojových dat při nastavování sítě. Lze specifikovat způsob trénování, který určí po průchodu kolika případů dojde ke změně nastavení. Možnosti jsou batch (změna nastavení po průchodu všech případů), online (změna nastavení po každém případě) a minibatch (rozdělení datové matice na části a změna nastavení po průchodu částí dat). Volby:
  - počet případů v rozdělených částech pro tréninkovou metodu minibatch
  - maximální počet záznamů uložených v paměti při výběru automatického stanovení architektury a/nebo při trénování metodou mini-batch
  - optimalizační algoritmus pro nastavování synaptických vah: gradientní metoda, sdružený gradient
  - výchozí rychlost učení v gradientní metodě
  - dolní mez rychlosti učení v gradientní metodě při volbách online a mini-batch
  - moment změny vah v gradientní metodě
  - výchozí hodnota parametru lambda pro sdružený gradient
  - gradient
  - výchozí hodnota parametru sigma pro sdružený gradient
  - rozsah hodnot náhodného nastavování vah  $[a0-a, a0+a]$  při simulovaném žihání
- **Specifikace STOPPINGRULES** vymezuje pravidla pro zastavení iteračního algoritmu učení sítě. Volby:
  - maximální počet kroků, během nichž nedojde ke snížení predikční chyby
  - zapnutí nebo vypnutí měření času učení a maximální doba učení
    - maximální počet průchodů daty
    - minimální relativní změna predikční chyby vzhledem k předchozí epoše/výchozímu řešení
- **Specifikace MISSING** určuje způsob zacházení s uživatelem definovanými vynechanými hodnotami kategorizovaných vstupních a výstupních proměnných. Případy lze vyřadit ze zpracování nebo považovat kódy vynechaných hodnot za platné údaje.
- **Specifikace PRINT** udává, jaké tabulky budou zobrazeny ve výstupovém okně. Lze také požádat o analýzu senzitivity. Volby:
  - sumární tabulka s počty platných případů a rozdělení do tréninkové, testovací a validační množiny
  - informace o neuronové síti: vstupy a výstupy, počty neuronů ve vrstvách, aktivační funkce
  - sumární tabulka učení sítě: průměrná chyba, pravidlo, které zastavilo učení, doba učení
  - matice záměn (klasifikační tabulka) pro každou kategorizovanou výstupní proměnnou
  - synaptické váhy, které představují odhady koeficientů přiřazených spojení neuronů
  - analýza senzitivity vyčísující důležitost každé predikční proměnné v neuronové síti
- **Specifikace PLOT** udává, které grafy budou zobrazeny ve výstupovém okně. Volby:
  - diagram neuronové sítě
  - graf závislosti predikce a skutečnosti pro každou výstupní proměnnou
  - graf závislosti reziduí na predikované hodnotě pro spojitě výstupní proměnné
  - ROC (Receiver Operating Characteristics) křivky pro kategorizované výstupní proměnné včetně tabulky s plochou pod křivkou
  - kumulativní grafy Gain pro kategorizované výstupní proměnné
  - grafy Lift pro kategorizované výstupní proměnné
- **Specifikace SAVE** určuje nové proměnné, které se připojí do aktivní datové matice. Uložte:
  - predikované hodnoty nebo kategorie
  - predikované pseudopravděpodobnosti pro jednotlivé kategorie
- **Specifikace OUTFILE** uloží model spolu s nastavením synaptických vah ve formátu XML.

### Radial Basis Function (RBF)

Procedura RBF vhodně nastaví neuronové sítě s radiálními bazickými funkcemi. Jedná se o sítě s dopřednou architekturou skládající se ze vstupní vrstvy, skryté vrstvy s radiálními bazickými funkcemi a výstupní vrstvy. Skrytá vrstva transformuje vstupní signály na hodnoty radiálních bazických funkcí. Podobně jako u MLP mohou RBF sítě provádět predikci číselných veličin i klasifikovat do skupin.

Procedura RBF trénuje síť ve dvou krocích:

1. Určení počtu radiálních bazických funkcí seskupovacím algoritmem. Stanovení těžiště a šířky každé radiální bazické funkce.
2. Určení synaptických vah mezi skrytou a výstupní vrstvou. Neurony výstupní vrstvy nepoužívají aktivační funkci (identita), proto nastavení vah může být provedeno tradiční metodou nejmenších čtverců.

*Díky popsanému dvoustupňovému algoritmu se RBF síť zpravidla učí rychleji než síť MLP.*

- *Specifikace pro proceduru RBF jsou podobné jako specifikace procedury MLP s těmito výjimkami:*
  - *specifikace ARCHITECTURE umožňuje zvolit typ radiálních bazických funkcí skryté vrstvy: normalizované RBF, běžné Gaussovské RBF*
  - *specifikace CRITERIA dovoluje nastavit parametry algoritmu učení procedury RBF, tj. míru překrytí radiálních bazických funkcí skryté*