I-SUNS: Zadanie č.1

Neurónové siete

Vo vybranom programovacom jazyku implementujte program, ktorý bude predpovedať pitnosť vody alebo kvalitu vína pomocou umelej neurónovej siete. V tomto zadaní budete pracovať s dátami z AIS. Pre každý dataset budú dostupné dva súbory - testovací a trénovací. Čas odovzdania je určený časom vloženia do AIS. Deadline pre získanie 15 bodov je **21.10.2021 o 13:00/15:00**. Každý týždeň omeškania je penalizovaný stratou dvoch bodov.

- Načítajte dáta z oboch množín (trénovacej a testovacej) a správne ich normalizujte/škálujte. Výsledok si overte:
  - zobrazením priemernej hodnoty a št. odchýlky pre stĺpce pred a po normalizácii; **(1b)**
  - zobrazením histogramu pre vybraný stĺpec pred a po normalizácii. **(1b)**
- Zhodnoťte náročnosť problému:
  - Vyhodnoťte úspešnosť pri náhodnom klasifikátore. **(1b)**
  - Natrénujte a vyhodnoťte klasifikáciu pomocou logistickej regresie. **(1b)**
- Natrénujte a vyhodnoťte neurónovú sieť:
  - Zvoľte si architektúru a parametre trénovania. **(1b)**
  - Používajte validačnú množinu a early stopping. **(1b)**
  - Sledujte vývoj chyby pre trénovaciu a validačnú množinu na grafe (aj do dokumentácie). **(1b)**
  - Vyhodnoťte váš klasifikátor pre obe množiny pomocou celkovej úspešnosti a konfúznej matice. **(2b)**
- Nájdite dobré parametre pre vašu sieť.
  - Meňte architektúru (počet alebo počet neurónov, typ kriteriálnej funkcie). **(1b)**
  - Meňte nastavenia trénovania (solver, rýchlosť učenia). **(1b)**
  - Výsledky z vašich experimentov prehľadne zobrazte do tabuľky. Pre najlepší výsledok zobrazte priebeh, úspešnosť a konfúznu maticu. **(2b)**

– Záznamy z jednotlivých trénovaní si uchovávajte a nájdite v nich príklady pretrénovania/podtrénovania, prirýchleho/pripomalého učenia alebo iné neštandardné priebehy. Zobrazte ich priebeh, úspešnosť a konfúznu maticu a analyzujte svoje pozorovania. **(2b)**

Nepovinné úlohy

- Pri analýze dát získajte a vizualizujte aj iné štatistické ukazovatele ako medián, spodný a horný kvartil, ... **(1b)**
- Analyzujte biele a červené víno samostatne. **(1b)**
- Predpovedajte, či je hodnota pH v odporúčanom rozmedzí podľa WHO (6.5 až 8.5).**(1b)**
- Sledujte dôležitosť vstupných parametrov a ukážte ich vplyv na priebeh trénovania. **(2b)**
- Granularizujte váš grid search aspoň v jednej iterácii. **(1b)**

Popis stĺpcov - voda

1. *pH value*: PH is an important parameter in evaluating the acid–base balance of water. It is also the indicator of acidic or alkaline condition of water status. WHO has recommended maximum permissible limit of pH from 6.5 to 8.5. The current investigation ranges were 6.52–6.83 which are in the range of WHO standards.
2. *Hardness*: Hardness is mainly caused by calcium and magnesium salts. These salts are dissolved from geologic deposits through which water travels. The length of time water is in contact with hardness producing material helps determine how much hardness there is in raw water. Hardness was originally defined as the capacity of water to precipitate soap caused by Calcium and Magnesium.
3. *Solids (Total dissolved solids - TDS):* Water has the ability to dissolve a wide range of inorganic and some organic minerals or salts such as potassium, calcium, sodium, bicarbonates, chlorides, magnesium, sulfates etc. These minerals produced un-wanted taste and diluted color in appearance of water. This is the important parameter for the use of water. The water with high TDS value indicates that water is highly mineralized. Desirable limit for TDS is 500 mg/l and maximum limit is 1000 mg/l which prescribed for drinking purpose.

4. *Chloramines*: Chlorine and chloramine are the major disinfectants used in public water systems. Chloramines are most commonly formed when ammonia is added to chlorine to treat drinking water. Chlorine levels up to 4 milligrams per liter (mg/L or 4 parts per million (ppm)) are considered safe in drinking water.

5. *Sulfate*: Sulfates are naturally occurring substances that are found in minerals, soil, and rocks. They are present in ambient air, groundwater, plants, and food. The principal commercial use of sulfate is in the chemical industry. Sulfate concentration in seawater is about 2,700 milligrams per liter (mg/L). It ranges from 3 to 30 mg/L in most freshwater supplies, although much higher concentrations (1000 mg/L) are found in some geographic locations.

6. *Conductivity*: Pure water is not a good conductor of electric current rather's a good insulator. Increase in ions concentration enhances the electrical conductivity of water. Generally, the amount of dissolved solids in water determines the electrical conductivity. Electrical conductivity (EC) actually measures the ionic process of a solution that enables it to transmit current. According to WHO standards, EC value should not exceeded 400 µS/cm.

7. *Organic carbon*: Total Organic Carbon (TOC) in source waters comes from decaying natural organic matter (NOM) as well as synthetic sources. TOC is a measure of the total amount of carbon in organic compounds in pure water. According to US EPA ¡ 2 mg/L as TOC in treated / drinking water, and ¡ 4 mg/Lit in source water which is use for treatment.

8. *Trihalomethanes*: THMs are chemicals which may be found in water treated with chlorine. The concentration of THMs in drinking water varies according to the level of organic material in the water, the amount of chlorine required to treat the water, and the temperature of the water that is being treated. THM levels up to 80 ppm is considered safe in drinking water.

9. *Turbidity*: The turbidity of water depends on the quantity of solid matter present in the suspended state. It is a measure of light emitting properties of water and the test is used to indicate the quality of waste discharge with respect to colloidal matter. The mean turbidity value obtained for Wondo Genet Campus (0.98 NTU) is lower than the WHO recommended value of 5.00 NTU.

10. ***Potability***: Indicates if water is safe for human consumption where 1 means Potable and 0 means Not potable.

Popis stĺpcov - víno

1. *fixed acidity*: most acids involved with wine or fixed or nonvolatile (do not evaporate readily) [tartaric acid - g / dm^3]

2. *volatile acidity*: the amount of acetic acid in wine, which at too high of levels can lead to an unpleasant, vinegar taste [acetic acid - g / dm^3]

3. *citric acid*: found in small quantities, citric acid can add 'freshness' and flavor to wines [g / dm^3]

4. *residual sugar*: the amount of sugar remaining after fermentation stops, it's rare to find wines with less than 1 gram/liter and wines with greater than 45 grams/liter are considered sweet [g / dm^3]

5. *chlorides*: the amount of salt in the wine [sodium chloride - g / dm^3]

6. *free sulfur dioxide*: the free form of SO2 exists in equilibrium between molecular SO2 (as a dissolved gas) and bisulfite ion; it prevents microbial growth and the oxidation of wine [mg / dm^3]

7. *total sulfur dioxide*: amount of free and bound forms of S02; in low concentrations, SO2 is mostly undetectable in wine, but at free SO2 concentrations over 50 ppm, SO2 becomes evident in the nose and taste of wine [mg / dm^3]

8. *density*: the density of water is close to that of water depending on the percent alcohol and sugar content [g / cm^3]

9. *pH*: describes how acidic or basic a wine is on a scale from 0 (very acidic) to 14 (very basic); most wines are between 3-4 on the pH scale

10. *sulphates*: a wine additive which can contribute to sulfur dioxide gas (S02) levels, which acts as an antimicrobial and antioxidant [potassium sulphate - g / dm^3]

11. *alcohol*: the percent alcohol content of the wine [% by volume]

12. *type*: type of wine (0 - white, 1 - red)

    Output variable (based on sensory data):

13. ***quality*** (0 - low quality, 1 - high quality)