**WD:**

**LAB. 2**  **Wymiarowość vs. ilość danych vs. złożoność klasyfikatora**

Wykonaj zadanie w notatniku używając repozytorium *scikIt learn* lub użyj cokolwiek (np. Orange)

1. Wybierz zbiór danych: ***Bone marrow mononuclear cells with AML*** (M=1000, N= 1004) z UCI dataset, (cechy to wartości liczbowe oraz etykiety kategoryczne 1,2 jako numery klas), lub inny podobny, który **zawiera dużą liczbę przykładów M>=1000 oraz cech N>=1000 tak, że M~N (M równe - z małym błędem - N)**. W przypadku innego zbioru z wieloma klasami połącz je tak by otrzymać mniej więcej zbalansowane dwie klasy.
2. Zaznajom się ze zbiorem, co reprezentuje, co chcemy uzyskać i na jakie pytania możemy próbować odpowiedzieć.
3. Dlaczego według teorii taki zbiór nie “kwalifikuje” się do użycia modeli ML (w bezpośredni sposób bez żadnej manipulacji cechami!)? Narysuj histogram odległości w analizowanym zbiorze i mapę odległości. Podaj argumenty „dlaczego”.
4. Zwizualizuj zbiór (t-SNE). Czy widzisz bardzo pomieszane klasy? Czy te klasy są dobrze rozseparowane?
   * Jeżeli pierwsze to wybrałeś nieprawidłowy zbiór.
   * Jeżeli to drugie to: Podaj najlepszy wynik klasyfikacji przy pomocy wybranego modelu (tu MLP, patrz niżej).
   * Zoptymalizuj złożoność modelu?
   * Podaj jakość klasyfikacji jako funkcję złożoności modelu dla zmniejszającej się ilości cech K. Wyrzucaj cechy tylko te poniżej pewnego (małego progu). Czyli cechy rzeczywiście bardzo mało istotne. Tabela max 4x4.
   * Odpowiedz teraz na pytanie: dlaczego pomimo 3) otrzymujemy „uporządkowaną” strukturę zbioru, a nie zmieszane klasy (czego moglibyśmy się spodziewać)? Zacytuj te wyniki które uzasadniają twoją odpowiedź.

* *Do wyboru cech użyj rankingu cech uzyskanego przy pomocy miary Information Gain* (IG).
* *Jako modelu użyj wielowarstwowej sieci neuronowej MLP. Na podstawie małego zbioru walidacyjnego* ***USTAL*** *meta-parametry modelu. Złożoność modelu będzie monotoniczną funkcja ilości wag, ale także ilości warstw. Wybierz 3-4 modele od najprostszego np. (1), do najbardziej złożonego np. (100,50,20). W nawiasach ilość neuronów w warstwach ukrytych.*

1. Jak się zmienia jakość klasyfikacji (twojego najlepszego klasyfikatora) w funkcji zmniejszającej się ilości przykładów i wzrostu ilości nieistotnych cech? W tym drugim przypadku startujesz z 5 najlepszych cech i dodajesz te najmniej istotne do pewnego progu IG.
2. Zaprezentuj wszystkie swoje wnioski.

**WD\_Lab2 🡪BJ\_Lab-3**

**BJ: Lab-3 Tasks**

**BJ do Lab-3**: ***Komórki jednojądrzaste szpiku kostnego (BM) z AML*** (M=1000, N=1004)…

1. Poznaj zestaw danych Bone Marrow… (BM) (zobacz także mój plik…)

Które klasy Zadań można zrealizować za pomocą BM d-set (?)

2. Przeczytaj pliki BJ na BM, t-SNE, NN i innych…

3. Zaimplementuj zadania klasyfikacyjne za pomocą BM d-set:

3.1 Użyj kNN dla zbioru d BM.

3.2 Użyj t-SNE dla BM d-zbioru.

3.4 Używaj sieci NN z różnymi solwerami i parametrami, warstwami ukrytymi…. Porównaj wyniki z wynikami widżetu SGD.

3.5 Spróbuj użyć widżetów Rank i/lub PCA wraz z sieciami NN (uważaj na przepływ informacji)

4. Analizuj różne Metryki wraz z Matrycą Confusion (dla różnych scenariuszy wdrażania sieci NN).

5. Spróbuj znaleźć odpowiednią („najlepszą”/dobrą) klasyfikację w ramach BM d-set.

6. Przygotuj raport (z wnioskami)

**BJ to Lab-3: *Bone Marrow(BM) mononuclear cells with AML*** (M=1000, N= 1004) **…**

1. Learn the ***Bone Marrow… (BM)*** dataset (see also my file…)

Which classes of Tasks can be implemented using ***BM*** d-set (?)

2. Read files from BJ on ***BM***, t-SNE, NNs and others…

3. Implement classification tasks with ***BM*** d-set:

3.1 Use kNN for ***BM*** d-set.

3.2 Use t-SNE for ***BM*** d-set.

3.4 Use NNs with different solvers and parameters, hidden layers,…. Compare the outputs with ones of SGD widget.

3.5 Try to use Rank widgets or/and PCA along with NNs (be careful concerning information flows)

4. Analyze different Metrics along with Confusion Matrix (for different scenarios of NNs implementation).

5. Try to find an appropriate (the “best”/good) classification within ***BM*** d-set.

6. Prepare a Report (with the Conclusions)