A

Zadanie	1	2	3	4	5	6	7	8	9	10	∑
Punkty											
(maks)	(2)	(2)	(2)	(2)	(2)	(6)	(6)	(6)	(6)	(6)	(40)

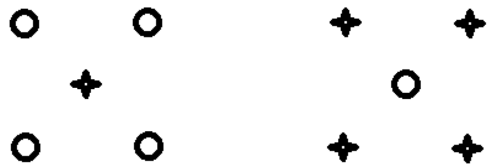
Uwaga: TA CZĘŚĆ ZOSTANIE WYPEŁNIONA PRZEZ EGZAMINATORA!

NAZWISKO I IMIĘ: NR. INDEKSU:

Systemy Uczące się – SUS'2011

— Czas pisania pracy: 75 minut —

1. Chcemy oszacować błąd klasyfikatorów typu kNN za pomocą algorytmu walidacji Leave-One-Out (1-CV). Używamy do tego celu następującego zbioru danych w $\mathbb{R}^2$ (z metryką euklidesową).



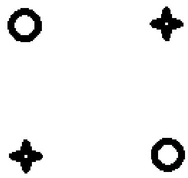
Wybierz T (tak) lub N (nie) przy następujących stwierdzeniach:

a) 1NN jest lepszym klasykatorem niż 3NN (jak zwykle);

b) 1NN jest gorszy niż 9NN;

c) Na tych danych 1NN nie może być lepszy niż jakikolwiek inny klasyfikator;

2. Dla następującego zbioru danych na płaszczyźnie, który z wymienionych klasyfikatorów może mieć zerowy błąd treningowy?



a) 3NN;

b) SVM (z jądrem kwadratowym);

c) Dwupoziomowe drzewo decyzyjne (z użyciem cięć na pojedyńczych atrybutach)

3. Który z poniższych algorytmów można zaimplementować dla uczenia "online", t.j. mając na wejściu nieskończony strumień przykładów, ale mając do dyspozycji skończoną pamięć?

a) kNN (z metryką Euklidesową i dowolnym parametrem k).

b) SVM, jeśli tylko punkty należące do różnych klas są separowalne.

c) AdaBoost.

4. Wyniki testów dwóch filtrów antyspamowych F_1 i F_2 na tej samej próbce testowej są przedstawione w poniższych tabelach:

	Prawdziwa decyzja	
	spam	*not spam*
$F_1 = spam$	15	15
$F_1 = not\ spam$	5	65

	Prawdziwa decyzja	
	spam	*not spam*
$F_2 = spam$	10	0
$F_2 = not\ spam$	10	80

☐ a) $FP(F_1) > FP(F_2)$;

☐ b) $TN(F_1) > TN(F_2)$;

☐ c) Który z filtrów jest lepszy? Odpowiedź uzasadnij

5. Które z poniższych algorytmów klastrowania nie wymagają znajomości wartości atrybutów poszczególnych obserwacji, a jedynie macierzy odległości między obserwacjami?

☐ a) k-Means (algorytm k-centroidów lub k-średnich)

☐ b) Grupowanie hierarchiczne z odległością między klastrami *complete linkage* (najbardziej odległe obserwacje).

☐ c) SVM.

6. Dany jest zbiór S zawierający 6 punktów na płaszczyźnie (z odległością euklidesową): $a = (0,0), b = (8,0), c = (16,0), d = (0,6), e = (8,6), f = (16,6)$. Chcemy znaleźć 3 klastry za pomocą algorytmu k-means (k-centroidów). Podział zbioru S na trzy klastry nazywamy *początkowym* jeśli jest wyznaczony przez jakiś początkowy układ środków, zaś *stabilnym* jeśli algorytm k-means już nie zmieni tego podziału. Wykonanie algorytmu k-means sprowadza podziały początkowe do jednego z podziałów stabilnych.

- Wyznacz wszystkie podziały stabilne dla S;

- Dla każdego podziału stabilnego dla S, wyznacz liczbę podziałów początkowych, z których możemy otrzymać ten podział stabilny;

- Jaka jest maksymalna liczba iteracji potrzebna do osiągnięcia podziału stabilnego z podziału początkowego?

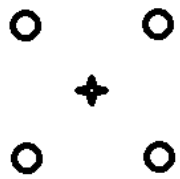
7. Rozpatrzmy tablicę decyzyjną dla problemu rozpoznawania grzybów trujących. Obiekty od 1 do 8 są treningowymi, lecz obiekty od 9 do 11 – testowymi.

LP.	ciężki?	śmierdzący?	kropeczki?	gładki?	trujący?
1	0	0	0	0	0
2	0	0	1	0	0
3	1	1	0	1	0
4	1	0	0	1	1
5	0	1	1	0	1
6	1	1	0	0	1
7	0	0	1	1	1
8	0	0	0	1	1
9	1	1	1	1	?
10	0	1	0	1	?
11	1	1	0	0	?

(a) Chcemy skonstruować drzewo decyzyjne dla tej tablicy. Który test, według Ciebie, powinien być używany w korzeniu drzewa? Odpowiedź uzasadnij

(b) Policz wszystkie potrzebne rozkłady prawdopodobieństw warunkowych i klasyfikuj obiekty 9, 10 i 11 metodą Naive Bayes.

8. Rozpatrzmy algorytm AdaBoost z użyciem prostych cięć na atrybutach jako słabe klasyfikatory na następującym zbiorze danych:



(a) Który obiekt będzie miał zwiększoną wagę po pierwszej iteracji?

(b) Ile iteracji trzeba wykonać, aby osiągnąć zerowy błąd treningowy?

9. (a) Dla funkcji jądrowej $K(\mathbf{x}, \mathbf{y}) = \langle \mathbf{x}, \mathbf{y} \rangle + 4(\langle \mathbf{x}, \mathbf{y} \rangle)^2$, gdzie $\mathbf{x}, \mathbf{y}$ są wektorami w przestrzeni dwuwymiarowej, znaleźć wartość k oraz odpowiednie zanurzenie $\Phi : \mathbb{R}^2 \to \mathbb{R}^k$ tak, aby

$$\langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle = K(\mathbf{x}, \mathbf{y})$$

(b) Niech $K_1, K_2 : \mathbb{R}^n \times \mathbb{R}^n \longrightarrow \mathbb{R}$ będą funkcjami jądrowymi. Udowodnij, że

$$K(\mathbf{x}, \mathbf{y}) = a_1 K_1(\mathbf{x}, \mathbf{y}) + a_2 K_2(\mathbf{x}, \mathbf{y})$$

też jest funkcją jądrową dla dowolnych dodatnich liczb rzeczywistych a_1, a_2.

10. Niech $\mathcal{H}_1$ będzie przestrzenią hipotez o wymiarze Vapnika-Chervonenkisa $VCdim(\mathcal{H}_1) = 3$ i niech $\mathcal{H}_2 = \{f_1, ..., f_k\}$ będzie dowolnym zbiorem klasyfikatorów. Niech $\mathcal{H}_3 = \mathcal{H}_1 \cup \mathcal{H}_2$, oraz niech $m = VCdim(\mathcal{H}_3)$.

(a) Uzasadnij, że $m < \infty$.

(b) Udowodnij (np. za pomocą lematu Sauera), że $2^m \leqslant m^3 + k$.

(c) Za pomocą oszacowania z (b) znajdź górne ograniczenie dla $VCdim(\mathcal{H}_3)$, gdy $k = 20$.