A

Zadanie	1	2	3	4	5	6	7	8	9	10	$\sum$
Punkty (maks)	(2)	(2)	(2)	(2)	(4)	(6)	(8)	(8)	(12)	(12)	(40)

Uwaga: ta część zostanie wypełniona przez egzaminatora!

Nazwisko i Imię:

Nr. indeksu:

Systemy uczące się – SUS'2012

— Czas pisania pracy: 75 minut —

1. [4 punkty] W tabeli poniżej przedstawiony jest zbiór treningowy, w którym każdy przykład ma cztery atrybuty (a_1, a_2, a_3, a_4) i decyzję (c).

   a. Jaka jest entropia tego zbioru danych?

   b. Jaki jest oczekiwany przyrost informacji (Information Gain) po obserwacji atrybutu a_1?

a_1	a_2	a_3	a_4	c
0	0	1	0	+
0	1	1	1	+
0	0	0	0	−
1	1	0	0	−

2. [4 punkty] Co to jest boosting?

   ☐ a) Proces, w którym przestrzeń hipotez jest explicite kolejno uzupełniana nowymi elementami.

   ☐ b) Niekorzystny efekt przeuczenia, gdy w modelu jest zbyt wiele stopni swobody.

   ☐ c) Technika budowy klasyfikatorów poprzez łączenie wielu słabych uczniów.

   ☐ d) Korzystny efekt wpływający na tempo uczenia.

3. [4 punkty] Co się stanie, gdy zwiększymy liczbę ukrytych węzłów w dwuwarstwowej sieci neuronowej?

   ☐ a) Będzie w stanie reprezentować bardziej skomplikowane wzorce.

   ☐ a) Będzie mniej skłonna do przeuczenia.

   ☐ a) Będzie lepiej generalizowała/uogólniała.

   ☐ a) Sieć będzie szybciej zbiegała.

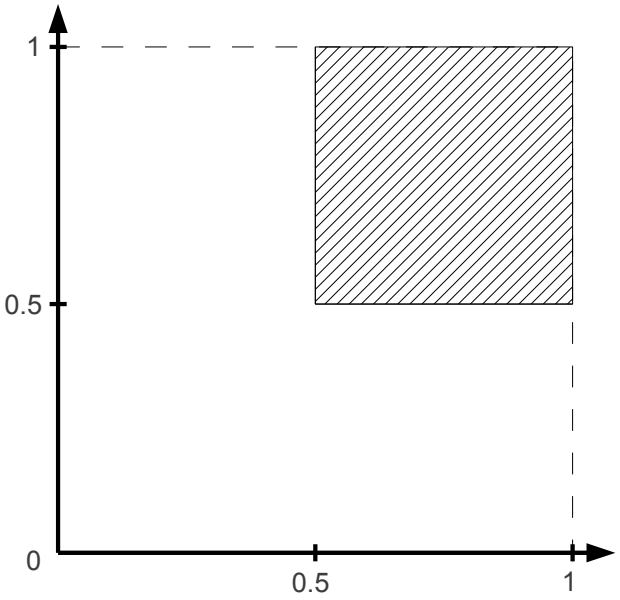
4. [8 punkty] Dany jest następujący zbiór punktów w $\mathbb{R}^2$: $\{A = (5,3), B = (4,0), C = (6,0), D = (1,0), E = (3,0), F = (1,1), G = (6,2), H = (2,3), I = (1,2), J = (4,4), K = (3,1), L = (4,1), M = (6,1), N = (3,4)\}$. Wypisz wektor grupowania na dwa podzbiory (przypisz poszczególnym punktom numer grupy), który jest wynikiem zastosowania hierarchicznego algorytmu aglomeracyjnego z *minimum* jako funkcją łączącą (*single linkage*) (w metryce Euklidesowej).

5. [8 punkty] Dane są obiekty $\{(0,0,+),(1,0,+),(0,3,-),(1,3,-),(0,6,+),(1,6,+)\}$. Oszacujemy skuteczność algorytmu k-NN przy $k=3$ (z metryką Euklidesową) metodą 3-CV, czyli walidacją krzyżową z trzema równolicznymi podzbiorami.

   a. Jaka jest najmniejsza możliwa skuteczność algorytmu? Odpowiedź uzasadnij.

   b. Jaka jest największa możliwa skuteczność algorytmu? Odpowiedź uzasadnij.

6. [8 punkty] Skonstruuj sieć neuronową która dla danych wejściowych $(x,y) \in [0,1] \times [0,1]$ rozpoznaje obszar zaznaczony na rysunku.



Zakładamy, że funkcją aktywacji jest:

$$f(w) = \begin{cases} 1 & \text{dla } w \geqslant 0 \\ -1 & \text{dla } w < 0 \end{cases}$$

Podaj najprostszą możliwą strukturę sieci oraz wagi poszczególnych neuronów.

7. [12 punkty] Zadana jest tablica danych:

has shell	aquatic	predator	venomous	type
F	F	T	F	mollusc
T	T	T	F	mollusc
T	F	F	F	insect
T	F	F	T	insect
T	T	F	F	insect
F	F	T	F	insect
F	F	T	T	mollusc
T	F	T	T	mollusc
F	T	F	F	mollusc

a. Wypisz najlepszą regułę długości 1.

a. Wypisz najlepszą regułę długości 2.
Wskazówka: Regułę taką można wygenerować algorytmem CN2 dla $K \geqslant 3$.

Do obliczania jakości reguł wykorzystaj estymatę Laplace'a zdefiniowaną jako:

$$L(\tau) = \frac{S_d(\tau) + 1}{S(\tau) + N_d},$$

gdzie $S(\tau)$ odpowiada liczbie obiektów wspierających regułę τ, $S_d(\tau)$ to liczba obiektów poprawnie klasyfikowanych przez τ, a N_d to liczba klas decyzyjnych.

8. [8 punkty] Na podstawie tablicy decyzyjnej z Zadania 7 sklasyfikuj obiekt $x = (T, T, F, T, ?)$ przy pomocy algorytmu NaiveBayes.

9. [12 punkty] Rozpatrzmy przestrzeń $\mathbb{R}^n_+$, czyli przestrzeń n-wymiarowych wektorów o współczynnikach dodatnich. Niech K_θ będzie funkcją definiowaną jak następująco:

$$K_\theta : \mathbb{R}^n_+ \times \mathbb{R}^n_+ \longrightarrow \mathbb{R}$$
$$(\mathbf{a}, \mathbf{b}) \longmapsto K_\theta(\mathbf{a}, \mathbf{b}) = \sum_{i=1}^{n} \delta_\theta(a_i, b_i)$$

gdzie $\mathbf{a} = \langle a_1, \ldots a_n \rangle$, $\mathbf{b}\langle b_1, \ldots\ b_n \rangle$ są wektorami z $\mathbb{R}_+^n$, a

$$\delta_\theta(a_i, b_i) = \begin{cases} 1 & \text{jeśli } a_i > \theta \text{ oraz } b_i > \theta \\ 0 & wpp \end{cases}$$

Innymi słowami, $K_\theta(\mathbf{a}, \mathbf{b})$ jest liczbą pozycji, w których oba wektory $\mathbf{a}$ i $\mathbf{b}$ mają większą wartość niż próg θ.

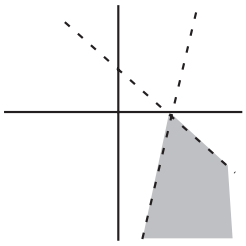
a. Pokazać, że $K_\theta(\mathbf{a}, \mathbf{b})$ jest funkcją jądrową (ang. kernel function) dla pewnego zanurzenia przestrzeni $\mathbb{R}_+^n$ w inną przestrzeń.

b. Naszym celem jest szukanie dobrej wartości θ dla odpowiedniego klasyfikatora SVM. Jak się zachowuje funkcja $K_\theta(\mathbf{a}, \mathbf{b})$?
(*wsk.* podaj właściwości liczbowe tej miary kiedy zmieniamy wartość θ od 0 do ∞.

c. Zaproponuj (heurystyczny) algorytm ustalenia optymalnej wartości dla θ

10. [12 punkty] Niech $\mathcal{X} = \mathbb{R}^2$, czyli $\mathcal{X}$ jest przestrzenią wszystkich punktów na płaszczyźnie.

Rozpatrujemy klasę $\mathcal{K}$ zawierającą hipotezy definiowane jako części wspólne dwóch perceptronów (półpłaszczyzn). Każda hipoteza może być traktowana jako wypukły kąt na płaszczyźnie (por. rysunek obok).



a. Pokazać, że trzy punkty mogą być rozbite (ang. shattered) przez hipotezy z $\mathcal{K}$;

b. Podaj przykład zbioru punktów na płaszczyźnie, który nie jest rozbity przez klasę $\mathcal{K}$;

c. Oblicz wymiar Vapnika-Chervonenkisa $VCdim$ dla klasy hipotez $\mathcal{K}$. Odpowiedź uzasadnij.