2. KONTROLNA ZADAĆA

1. *20%*

a) Hipoteza *h(x)="Životinja x je iz razreda sisavaca"*, dobivena učenjem ID3 algoritmom na temelju vizualnih značajki, provjerava se na skupu primjera za provjeru. Skup sadrži 45 primjera. Ustanovljeno je da je hipoteza sedam sisavaca klasificirala pogrešno, te da ujedno kao sisavca klasificirala žabu gatalinku običnu (*hyla arborea*) i pjegavog daždevnjaka (*salamandra salamandra*), koji, međutim, oboje pripadaju razredu vodozemaca. Nađi 95%-tni interval pouzdanosti za pravu vrijednost proporcije pogreške $error_D(h)$ hipoteze *h(x)*.

b) Koliko minimalno primjera moramo dodati u skup primjera za učenje ako želimo da interval iz *a)* dijela zadatka, uz istu pouzdanost i istu procjenu pogreške, ne bude širi od 0.2?

Za rješavanje ovog i idućeg zadatka koristi se slijedećim isječkom iz statističke tablice za vjerojatnosti $P(Z \leq z) = \Phi(z)$ u standardnoj normalnoj razdiobi $N(0, 1)$. Vrijednost argumenta z mijenja se u redcima tablice na mjestu desetice, a u stupcima na mjestu stotice. Npr. $P(Z \leq 1{,}82)=0{,}9656=96{,}56\%$. Pri tome je dovoljno odabrati vrijednost najbližu traženoj.

z	0,00	0,01	0,02	0,03	0,04	0,05	0,06	0,07	0,08	0,09
1,5	0,9332	0,9345	0,9357	0,9370	0,9382	0,9394	0,9406	0,9418	0,9429	0,9441
1,6	0,9452	0,9463	0,9474	0,9484	0,9495	0,9505	0,9515	0,9525	0,9535	0,9545
1,7	0,9554	0,9564	0,9573	0,9582	0,9591	0,9599	0,9608	0,9616	0,9625	0,9633
1,8	0,9641	0,9649	0,9656	0,9664	0,9671	0,9678	0,9686	0,9693	0,9699	0,9706
1,9	0,9713	0,9719	0,9726	0,9732	0,9738	0,9744	0,9750	0,9756	0,9761	0,9767

Primijeti da je u tablici riječ o jednostranom intervalu. Pri pretvorbi jednostranog intervala u obostrani, i obrnuto, posluži se skicom funkcije gustoće vjerojatnosti.

---

2. *20%*

Čitač optičkog bar-koda, prema navodima proizvođača, s namirnica očitava pogrešan kod u samo 2% slučajeva. Stoji li tvrdnja proizvođača u slučaju kada se kupcu dogodi da od 32 namirnica koje kani kupiti, kodove za dvije namirnice blagajnica mora unijeti ručno, jer čitač griješi. (Napravi dvostrani test hipoteze uz odabrani nivo značajnosti $\alpha = 5\%$.)

---

3. *20%*

a) Navedi na koja dva osnovna načina možemo riješiti problem nedostajuće vrijednosti atributa kod računanja informacijske dobiti za ID3 algoritam.

b) Objasni što je to prekomjerno naučena hipoteza. Zašto takve hipoteze nisu poželjne? Kako izbjegavamo prekomjernu naučenost kod hipoteza u obliku stabla odluke?

c) Temeljni ID3 algoritam radi s diskretnim vrijednostima atributa, no može se modificirati za rad s atributima koji imaju kontinuirane vrijednosti. Ukratko navedite korake u postupku rada ID3 algoritma s kontinuiranim vrijednostima.

d) U skladu s *c)* dijelom zadatka, na navedenom primjeru nađite onu vrijednost kontinuiranog atributa *Vlaga* koja je najpogodnija za dinamičko kreiranje booleovog atributa pri izgradnji ID3 stabla.

Vlaga (%)	70	75	80	85	95
Ciljni atribut	NE	NE	DA	DA	NE

---

4. *20%*

a) Pretpostavimo da smo za algoritam vektora potpore odabrali linearnu funkciju, tj. $k(x_i , x_j)= (x_i , x_j)$. Usporedi algoritam vektora potpore s algoritmom perceptrona u smislu konačnog rješenja (hiper-ravnine) uz pretpostavku da imamo linearno odjeljive uzorke.

b) Skiciraj u ravnini dvije grupe linearno odjeljivih uzoraka te skiciraj rješenje koje nastaje algoritmom *SVM*. Kakvo svojstvo ima to rješenje?

c) Pretpostavimo da smo za *SVM* sada odabrali polinomijalnu ili Gaussovu jezgrenu funkciju te da uzorci nisu linearno odjeljivi. U čemu je sada glavna razlika između algoritma vektora potpore i algoritmom perceptrona?

d) Što predstavljaju jezgrene funkcije $K(x_i , x_j)$ u algoritmu vektora potpore?

---

5. *20%*

a) Navedi metode učenja na temelju primjera i objasni po čemu se one razlikuju od ostalih metoda (npr. ID3 algoritma ili algoritma eliminacije kandidata).

b) Koja je induktivna pristranost algoritma *k* najbližih susjeda?

c) Ukratko objasni metodu aproksimacije funkcije pomoću radijalnih baznih funkcija (*RBF*). Koji je uobičajeni izbor za jezgrenu funkciju? Grafički prikaži interpretaciju *RBF* hipoteze kao neuronsku mrežu. Kako se trenira takva mreža?

d) Navedi i objasni nedostatke metoda učenja na temelju primjera u odnosu na ostale metode.

A. *20%*

Poznate su apriorne vjerojatnosti dviju hipoteza, a to su H_1 da osoba ima rak, odnosno H_2 da osoba nema rak, tj. $P(rak) = 0.008$ i $P(\neg rak) = 0.992$. Test na rak klasificira točno pozitivne slučajeve u 98% slučajeva, negativne u 97 %, tj. vrijedi:

P(test+ | rak) = 0,98	P(test- | rak) = 0.02
P(test- | ¬ rak) = 0,97	P(test+ | ¬ rak) = 0.03

Pretpostavimo da je nekoj osobi test na rak dao pozitivan rezultat. Nađi maksimalnu aposteriornu hipotezu (h_{MAP}).

■