**Zimski ispitni rok 2019/2020 – Strojno učenje**

**1. Zadatak – Osnovni koncepti**

a) Ukratko objasnite 3 osnovne komponente svakog algoritma strojnog učenja te ih na primjeru modela linearne regresije povežite s konceptima pristranosti jezikom i pristranosti preferencijom.

b) Razmatramo model *Ɦ* u *X* = *{1,2,…,8}* koji sadrže hipoteze oblika *h(x|w) = 1{x ≥ w}*, pri čemu w є ℝ. Raspolažemo skupom primjera *D = {(x(i), y(i))} = {(1,0), (2,0), (4,0), (6,1), (7,1),(8,1)}.* Formalno definirajte prostor inačica te odredite |*VSꞪ,D*|.

**2. Zadatak – Linearna regresija**

a) Treniramo L2-regularizirani model univarijantne linearne regresije, *h(x) = w0 + w1x + w2x2* , nad podatcima koji su u stvarnosti generirani polinomom 1.stupnja. U ravnini w1-w2 skicirajte (1) izokonture neregularizirane funkcije pogreške, (2) izokonture L2-regularizacijskog izraza te (3) izokonture L2-regularizirane pogreške.

b) Ukratko opišite kako se algoritam linearne regresije može upotrijebiti za binarnu klasifikaciju. Objasnite koji je glavni nedostatak takvog postupka kroz primjer i kroz skicu funkcije gubitka tog algoritma (graf *L* u ovisnosti o *yh(x)*).

**3. Zadatak – Logistička regresija**

a) Napišite pseudokod algoritma stohastičkog gradijenta spusta L2-regularizirane logističke regresije s linijskim pretraživanjem. Konvergira li uvijek ovaj postupak pri ovakvoj izvedbi? Objasnite.

b) Izvedite pogrešku *poopćene* unakrsne entropije *E(W|D)* kao negativnu log-izglednost na skupu za učenje. Napišite sve pretpostavke na kojima se ovaj izvod temelji.

c) Definirajte model multinomijalne logističke regresije. Ako koristimo naučeni model multinomijalne logističke regresije (bez preslikavanja) s težinama *w1 = (1,2,2,3), w2 = (4,4,1,0)* i *w3 = (-2,-3,4,5)*, izračunajte predikciju modela za *x = (1,1,1).*

**4. Zadatak – SVM, jezgrene i neparametarske metode**

a) Izvedite, korak po korak, primarni problem (tvrde) maksimalne margine, a zatim dođite do primarnog problema meke maksimalne margine. Koja je motivacija iza uvođenja meke margine?

b) Definirajte radijalnu bazu (RBF). Je li ova familija jezgrenih funkcija osjetljiva na razlike u skalama značajki? Zašto?

c) Primjenjujemo model k-NN na skup podataka *D* koji se sastoji od 3 klase, pri čemu je 40 primjera jedne klase, 20 druge i 70 treće. Kolika će biti pogreška učenja ovog modela na skupu *D* ako koristimo k = 1 i ako koristimo k = 130? Objasnite.

**5. Zadatak – Procjenitelji, Bayesov klasifikator i probabilistički grafički modeli (6 bodova)**

a) Definirajte Pearsonov koeficijent korelacije. Koliko bi otprilike iznosio Pearsonov koeficijent između varijabli X i Y, ako X predstavlja troškove grijanja, a Y temperaturu zraka (prema vremenskoj prognozi)? Zašto?

b) Definirajte kriterij uzajamne informacije *I(X, Y)*. Razmatramo familiju modela polunaivnog Bayesovog klasifikatora *Ɦα* kod kojeg se združivanje varijabli provodi za sve parove varijabli *(xi, yi)* za koje *I(X, Y) ≥ α.* Skicirajte pogreške učenja i ispitivanja modela *Ɦα* kao funkcije praga *α* (dvije krivulje na istoj slici).

c) Bayesovom mrežom modeliramo vjerojatnost oboljenja od kardiovaskularnih bolesti. Mreža sadrži 4 varijable: spol osobe (S), koliko često osoba odlazi u teretanu (T), je li osoba pušač (P)te varijablu koja govori o kakvom se riziku radi (R). Pritom vrijedi S є {muški, ženski}, P є {False, True}, T є {1, 3, 5} i R є {nizak, umjeren, visok}. Zajednička razdioba faktorizirana je kao *P(S, T, P, R)* = *P(S)P(P)P(T|S,R)P(R|T).* Primjenom Laplaceovog MAP-procjenitelja procijenite *P(T|S,P)* na danom skupu podataka *D*:

|  |  |  |  |
| --- | --- | --- | --- |
| **S** | **P** | **T** | **R** |
| Ženski | True | 1 | Visok |
| Ženski | True | 5 | Umjeren |
| Muški | False | 3 | Nizak |
| Ženski | False | 1 | Umjeren |
| Muški | True | 5 | Nizak |
| Ženski | False | 1 | Nizak |

**6. Zadatak – Vrednovanje klasifikatora i odabir značajki (4 boda)**

a) Od N = 1000 primjera, klasifikator je za prvu, drugu i treću klasu ispravno pozitivno klasificirao njih 590, 146, odnosno 134. Od preostalih 130 neispravno je klasificiranih primjera, 30 ih je klasificirao u drugu umjesto u prvu, 60 u drugu umjesto u treću, a 40 u treću umjesto u prvu klasu. Izračunajte makro-*F1.*

b) Definirajte unakrsnu provjeru i objasnite zašto ju koristimo. Zatim napišite pseudokod ugniježđene višestruke unakrsne provjere(eng. nested k-fold cross validation) i detaljno obrazložite njene prednosti nad „običnom“ unakrsnom provjerom.

**7. Zadatak - Grupiranje (4 boda)**

a) Napišite izraz za model miješane gustoće i model Gaussove mješavine. Kada model Gaussove mješavine degradira na algoritam k-sredina?

b) Raspolažemo skupom neoznačenih primjera: *D = {a=(0, 3), b=(4, 1), c=(3, 5), d=(1, 5)}.* Primijenite hijerarhijsko aglomerativno grupiranje (HAC) s euklidskom udaljenošću i prosječnim povezivanjem. Ispišite međukorake algoritma (matrice udaljenosti), skicirajte pripadni dendogram te na njemu navedite udaljenosti na kojima se odvija stapanje.