UNIZG FER, **Strojno učenje**, ak. god. 2011/12.

Domaća zadaća 3

Zadano:	**13.12.2011.**
Rok predaje:	**23.12.2011.** do **17.00 sati**

Napomena: Rješenju treba priložiti izvorne kodove programa.

1. Na predavanjima smo objasnili kako se linearna regresija može upotrijebiti za klasifikaciju. Osnovna ideja jest naučiti model $h_j(\mathbf{x})$ koji (u idealnom slučaju) daje $h(\mathbf{x}) = \tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} = 1$ za primjere koji pripadaju klasi $\mathcal{C}_j$, a $h(\mathbf{x}) = \tilde{\mathbf{w}}^{\mathrm{T}}\tilde{\mathbf{x}} = 0$ za primjere koji ne pripadaju toj klasi. Rješenje $\mathbf{w}_j$ koje minimizira kvadratnu pogrešku klasifikacije za klasu $\mathcal{C}_j$ dobiva se pseudoinverzom matrice primjera:

$$\tilde{\mathbf{w}}_j = (\tilde{\mathbf{X}}^{\mathrm{T}}\tilde{\mathbf{X}})^{-1}\tilde{\mathbf{X}}^{\mathrm{T}}\mathbf{y} = \tilde{\mathbf{X}}^{+}\mathbf{y}\,.$$

Želimo naučiti klasifikator za tri klase u prostoru $\mathcal{X} = \mathbb{R}^2$. Koristit ćemo tri klasifikatora i metodu jedan-naspram-ostali. Skup primjera za učenje je

$$\mathcal{D} = \left\{(\mathbf{x}^{(i)}, y^{(i)})\right\}_{i=1}^{5} = \left\{(5,3),1),((5,-1),1),((-3,0),2),((-6,-4),2),((-4,6),3)\right\}.$$

(a) Izračunajte težine $\tilde{\mathbf{w}}_j$ za sve tri klase i skicirajte u prostoru $\mathbb{R}^2$ primjere $\mathcal{D}$ i pravce $h_1(\mathbf{x}) = 0$, $h_2(\mathbf{x}) = 0$ i $h_3(\mathbf{x}) = 0$.

(b) Na drugoj skici prikažite granice između klasa koje odgovaraju hipotezi

$$h(\mathbf{x}) = \operatorname*{argmax}_{\mathcal{C}_j} h_j(\mathbf{x})\,.$$

(c) U koju će klasu biti klasificiran primjer $\mathbf{x} = (1,3)^{\mathrm{T}}$? Koja je vjerojatnost da primjer pripada toj klasi?

(d) Kakav će utjecaj na položaj granica očekujete da će imati uvođenje primjera $(\mathbf{x}, y) = ((10,10),1)$? (Ne morate iznova računati; odgovorite opisno.) Je li taj efekt poželjan ili ne, i zašto?

2. Logistička regresija je diskriminativan linearan klasifikacijski model koji ima probabilističku interpretaciju.

(a) Izvedite izraz za pogrešku unakrsne entropije krenuvši od log-izglednosti na skupu za učenje.

(b) Po čemu možete zaključiti da je ovo diskriminativan model?

(c) Kako konvergencija algoritma gradijentnog spusta ovisi o parametru η? Kakav utjecaj ima broj primjera N na parametar η? Kako bismo mogli definirati funkciju pogreške, a da poništimo takav utjecaj?

(d) Neka je skup primjera za učenje

$$\mathcal{D} = \left\{((1,1),1),((-1,-1),0)\right\}.$$

Koliko iznosi pogreška unakrsne entropije, a koliko vektor gradijenta $\nabla E(\tilde{\mathbf{w}})$ za $\tilde{\mathbf{w}} = (0,0,0)^{\mathrm{T}}$? Hoće li algoritam gradijentnog spusta konvergirati? Obrazložite odgovor.

(e) Hoće li algoritam gradijentnog spusta u situaciji iz zadatka 2d konvergirati ako koristimo regulariziranu funkciju pogreške? Zašto?

(f) U općenitom slučaju, ako povećamo vrijednost parametra regularizacije λ, hoće li se vrijednost empirijske pogreške povećati ili smanjiti? Obrazložite odgovor.

(g) Neka je skup primjera za učenje

$$\mathcal{D} = \left\{((2,5),0),((2,4),0),((2,-5),0),((-2,-5),1),((-2,-4),1),((-2,5),1))\right\}.$$

Skicirajte ovaj skup i granicu koju dobivamo modelom logističke regresije (granicu možete lako nacrtati ako se prisjetite koja je veza između logističke regresije i generativnog modela). Komentirajte: kako je moguće da logistička regresija ne uspijeva ispravno klasificirati linearno odvojive primjere?

3. U dnevniku "Vjesnik" novinski tekstovi svrstavaju se u jednu od nekoliko kategorija (npr. *Svijet*, *Sport*, *Gospodarstvo*). Klasifikacija tekstova po temama jedan je od zadataka koji se može riješiti korištenjem više binarnih klasifikatora i to tako da svaki klasifikator razlikuje tekstove koji pripadaju određenoj kategoriji od svih drugih tekstova.

Vaš zadatak je izrada binarnog klasifikatora za kategoriju *Svijet*, tj. klasifikatora koji će sve tekstove koji pripadaju kategoriji *Svijet* klasificirati kao pozitivne ($y = 1$), a sve tekstove koji ne pripadaju toj kategoriji kao negativne ($y = 0$). Ovaj zadatak možete rješavati u programskom jeziku po izboru. Obavezno priložite izvorni kôd programskog rješenja.

Skup primjera sadrži 100 novinskih članaka iz kategorije *Svijet* (pozitivni primjeri) i 100 iz svih ostalih kategorija (negativni primjeri). Podijeljen je na skupove za učenje, provjeru i ispitivanje u omjeru 80:60:60. Novinski članci pripremljeni su za klasifikaciju na način opisan u nastavku (v. sliku 1).

Priprema tekstova

U prvom koraku iz teksta svakog članka izbačene su interpunkcije i sve su riječi napisane malim slovom. Tako obrađeni tekstovi mogu se vidjeti u datoteci `korak1.txt` (u svakom se retku nalazi po jedan članak).

Pretvorba u vektorski zapis

Pokazalo se da za potrebe klasifikacije tekstova po temama nije potrebno paziti na redoslijed riječi. Zapis u kojemu se za svaki dokument pamti samo broj pojavljivanja pojedinih riječi zove se *vreća riječi* (engl. *bag of words*). U datoteci `korak2.txt` nalaze se članci u takvom zapisu (zbog jednostavnosti obrade riječi koje se ne nalaze u primjerima za učenje izbačene su iz skupa za provjeru i ispitnog skupa).

Sve različite riječi koje se javljaju u primjerima za učenje poredane su abecedno i zapisane u datoteku `rjecnik.txt`. U idućem koraku svaka riječ zamjenjuje se rednim brojem te iste riječi u rječniku. Dimenzija prostora je jednaka broju riječi u rječniku, tj. 9982. Datoteka `korak3.txt` sadrži rijetki zapis vektora.

Obrada značajki

Riječi koje se često javljaju u različitim dokumentima manje su korisne za klasifikaciju teksta. Za potrebe klasifikacije teksta često se svakoj riječi dodjeljuje težina idf (engl. *inverse document frequency*):

1. Početni tekst:

```
Pedesetšestogodišnji šumski radnik smrtno je stradao u ponedjeljak oko
podneva u mjestu Ruševica, općina Cetingrad, kada se na njega srušilo stablo.
Njegov četrdesetjednogodišnji kolega stradao je prilikom sječe stabala u šumi
pokraj mjesta Furjan.  Zadobio je teške tjelesne ozljede.
```

2. Zadržane samo riječi, velika slova pretvorena u mala:

```
pedesetšestogodišnji šumski radnik smrtno je stradao u ponedjeljak oko
podneva u mjestu ruševica općina cetingrad kada se na njega srušilo stablo
njegov četrdesetjednogodišnji kolega stradao je prilikom sječe stabala u šumi
pokraj mjesta furjan zadobio je teške tjelesne ozljede
```

3. Vreća riječi:

```
cetingrad:1 furjan:1 je:3 kada:1 kolega:1 mjesta:1 mjestu:1 na:1 njega:1
njegov:1 oko:1 općina:1 ozljede:1 pedesetšestogodišnji:1 podneva:1 pokraj:1
ponedjeljak:1 prilikom:1 radnik:1 ruševica:1 se:1 sječe:1 smrtno:1
srušilo:1 stabala:1 stablo:1 stradao:2 teške:1 tjelesne:1 u:3 zadobio:1
četrdesetjednogodišnji:1 šumi:1 šumski:1
```

4. Zapis u obliku rijetkog vektora:

```
623:1 1455:1 2397:3 2520:1 2709:1 3504:1 3508:1 3663:1 4172:1 4173:1 4614:1
4748:1 4977:1 5110:1 5267:1 5385:1 5490:1 6187:1 6621:1 7030:1 7164:1 7296:1
7478:1 7659:1 7664:1 7667:1 7798:2 8226:1 8272:1 8494:3 9319:1 9744:1 9913:1
9914:1
```

5. Vektor nakon skaliranja značajki:

```
623:4.38 1455:4.38 2397:0.00 2520:1.49 2709:3.69 3504:1.90 3508:2.77
3663:0.04 4172:1.90 4173:2.08 4614:1.29 4748:4.38 4977:3.69 5110:4.38
5267:3.69 5385:3.69 5490:1.82 6187:2.77 6621:4.38 7030:4.38 7164:0.13
7296:4.38 7478:3.69 7659:4.38 7664:3.69 7667:4.38 7798:8.76 8226:3.69
8272:4.38 8494:0.00 9319:4.38 9744:4.38 9913:3.69 9914:4.38
```

Slika 1: Pretvaranje teksta u vektorski zapis.

$$\text{idf}(w) = \log \frac{\text{broj primjera za učenje}}{\text{broj primjera u kojima se pojavljuje riječ } w}\,.$$

Na primjer riječ "a" ima vrlo malu težinu, a riječ "je" težinu 0 jer se pojavljuje u svakom primjeru. Težine riječi nalaze se u datoteci `idf.txt` (težina u retku i odgovara dimenziji i i riječi u i-tom retku datoteke `rjecnik.txt`). Konačni vektori (kojima je i-ta komponenta pomnožena s $\text{idf}(w_i)$) nalaze se u rijetkom zapisu u datoteci `korak4.txt` te u gustom zapisu u datoteci `korak5.txt`.

Zadatak

Učitajte primjere iz datoteka `X_train.txt` (9982 stupaca i 80 redaka) i `y_train.txt` (80 redaka). U i-tom retku datoteke `X_train.txt` nalazi se i-ti primjer u vektorskom zapisu, a u i-tom retku datoteke `y_train.txt` nalazi se klasifikacija tog primjera (1 = kategorija *Svijet*, 0 = ostali tekstovi).

Ako radite u sustavima Matlab/Octave možete za učitavanje matrica `X_train` i `y_train` koristiti ovaj isječak kôda:

```
load X_train.txt
load y_train.txt
```

(a) Napišite (ne u kôdu nego u izvještaju) funkciju pogreške koju optimirate u ovisnosti o primjerima za učenje i parametru λ.

(b) Implementirajte metodu gradijentnog spusta kojom ćete pronaći mimumum te funkcije. Uz hiperparametar $\lambda = 0$ pronađite parametre logističke regresije koji minimiziraju pogrešku na skupu za učenje (koristeći matrice `X_train` i `y_train`). *Uputa:* pogrešku možete smanjiti ako povećate broj iteracija i smanjite parametar η.

Napišite u izvještaju vrijednost funkcije pogreške na skupu za učenje (uz dobiveni vektor težina). U datoteku `w_a.txt` zapišite vektor težina tako da u retku i piše vrijednost komponente s indeksom $i - 1$ (tj. u prvom retku piše vrijednost težine w_0, u drugom težine w_1, itd.). Uz ovako naučen klasifikator broj pogrešno klasificiranih primjera na skupu za učenje trebao bi biti 0.

(c) U ovom podzadatku potrebno je pronaći najbolju vrijednost hiperparametra λ korištenjem skupa za učenje i skupa za provjeru.

Odaberite 20-ak vrijednosti parametra λ (obavezno uključite $\lambda = 0$). Raspon parametra λ je od 0 do ∞, pa se preporučuje da odabrani parametri pokrivaju nekoliko redova veličina (kako vrijednosti manje od 1, tako i vrijednosti znatno veće od 1). Za svaki od odabranih vrijednosti parametra λ pronađite najbolji vektor težina korištenjem skupa za učenje. Pomoću svakog od naučenih vektora težina klasificirajte primjere iz skupa za provjeru (datoteke `X_validate.txt` i `y_validate.txt`) i zapišite postotak pogrešno klasificiranih primjera. Broj pogrešno klasificiranih primjera trebao bi biti manji od 20%.

U izvještaju napravite tablicu koja će sadržavati sve vrijednosti λ, postotak pogrešno klasificiranih primjera na skupu za provjeru za svaku vrijednost od λ te vrijednosti koje su korištene u metodi gradijentnog spusta (η i broj iteracija).

(d) Koristeći parametar λ iz prošlog podzadatka, kojim se minimizira broj pogrešno klasificiranih primjera na skupu za provjeru, potrebno je naučiti klasifikator korištenjem spojenih skupova za učenje i provjeru. Težine tako naučenog klasifikatora zapišite u datoteku `w_d.txt`. Klasificirajte primjere iz skupa za ispitivanje (datoteke `X_test.txt` i `y_test.txt`) i napišite postotak pogrešno klasificiranih primjera. Postotak pogrešno klasificiranih primjera ovako naučenog klasifikatora trebao bi biti manji od klasifikatora naučenog uz $\lambda = 0$.

Za provjeru se mogu pogledati riječi koje odgovaraju najpozitivnijim komponentama vektora težina naučenog klasifikatora. To su riječi koje su najvažnije za dotičnu kategoriju. Za kategoriju *Svijet* za 50 najvažnijih riječi trebali biste dobiti sljedeće:

```
američki, američkom, austrijski, bih, crkve, dnevnik,
glasnogovornik,italiji, izbora, izjava, izjavio, između, kancelar,
kancelara, katoličke,koalicija, koalicije, krize, ljudima, mjeseci,
napad, nedjelju, njemačke,objavila, odnose, osuđen, pakistana,
petak, potres, pravni, predsjednika,predvodio, pregovora,
priopćenju, protiv, rekao, republike, srj, stanje,state, stranke,
supruga, sveta, svjedočio, tajnik, utorak, vlade, vlasti, vojska,
vođa
```