# Otkrivanje biomarkera i klasifikacija

DATA MINING FOR GENOMICS AND PROTEOMICS

DARIUS M. DZIUDA

Chapter 3

Istraživanje podataka u bioinformatici, 2021/2022.

G. Pavlović-Lažetić

# Otkrivanje biomarkera i klasifikacija

- Otkrivanje biomarkera i klasifikacija – pregled
- Otkrivanje biomarkera
- Validacija i izveštavanje
- Izbor svojstava (Feature Selection)
- Diskriminantna analiza i izbor svojstava
- SVM i izbor svojstava
- Nasumične šume i izbor svojstava

# Otkrivanje biomarkera i klasifikacija - pregled

- Implementacija i korišćenje metoda nenadgledanog učenja u istraživanju podataka, u opštem slučaju, postigli su zrelost u oblasti biomedicinskih istraživanja
- Drugačije je sa nadgledanim učenjem, posebno u oblasti otkrivanja biomarkera
- Mada postoje dobro ustanovljene metode – SVM, stabla odlučivanja, ...,
- Nedovoljno istražen korak odabira  svojstava
- Otkrivanje biomarkera uključuje:
  - Odabir svojstava
  - Izgradnju klasifikacionog modela
  - Validaciju modela (poželjno na nezavisnom skupu podataka za testiranje)
  - Implementaciju klasifikacionong modela (poželjno sa vizuelizacijom diskriminativnog prostora)
  - Objašnjenje bioloških procesa u osnovi diferencijacije klasa

# Pregled

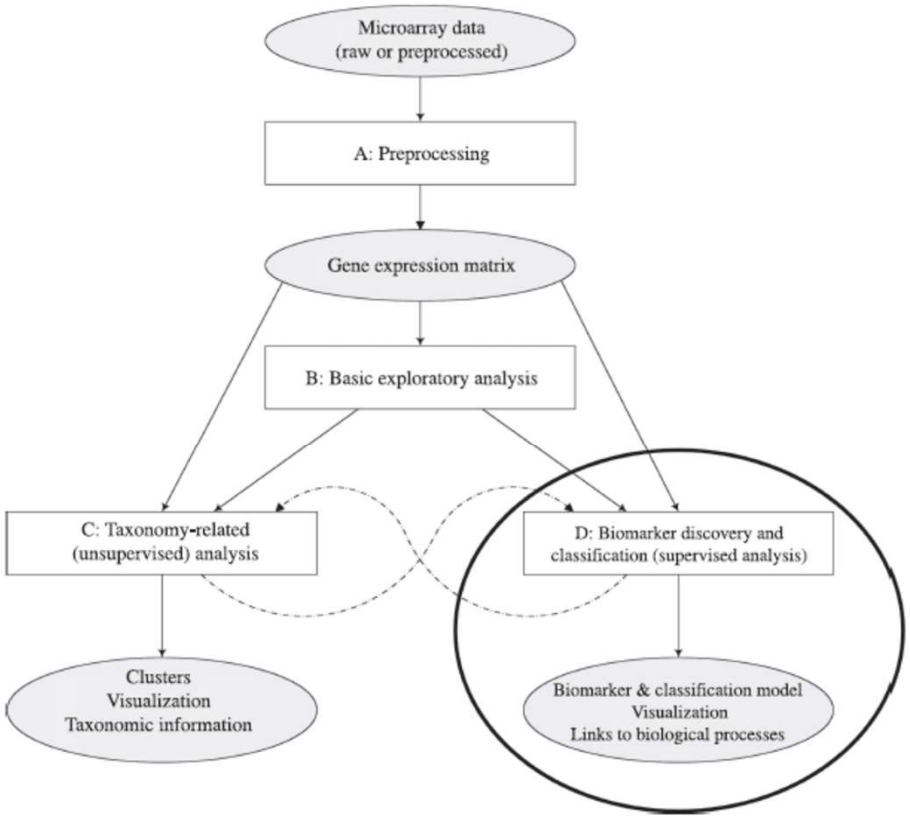



**Figure 3.1:** Elements of microarray gene expression data analysis—an example. The focus of this chapter is on the element D: Biomarker discovery and classification.

# Šta su biomarkeri

- (jedna od definicija) Biomarkeri su merljive promene parametara bioloških sistema (izgrađenih od organskih i neorganskih jedinjenja ) koje se prate pomoću biomonitoringa
- Biomarkeri se definišu kao indikatori razlika u biohemijskim ili ćelijskim elementima ili procesima, u strukturi ili funkcijama bioloških sistema ili uzoraka.
- Vrste:
  - Molekularni biomarkeri (npr. promene na DNK/RNK)
  - Organizmički biomarkeri (npr. anatomske/ histološke/ citološke promene/status)
  - Populacijski biomarkeri (npr. kvalitativni sastav biocenoza)
- U osnovnim i kiliničkim istraživanjima
  - Široka potkategorija medicinskih znakova – objektivnih indikatora medicinskog stanja koji se opažaju spolja i mogu se tačno i ponovljivo meriti (različiti od medicinskih simptomima koje opažaju sami pacijenti)
  - National Institutes of Health Biomarkers Definitions Working Group: biomarker je karakteristika koja se objektivno meri i evaluira kao indikator normalnog biološkog procesa, patogenog procesa ili farmakološkog odgovora na terapijsku intervenciju
  - World Health Organization (WHO), UN: biomarker je bilo koja supstanca , struktura ili proces koji se može meriti u organizmu ili njegovim proizvodima i utiče na ili predviđa učestolst  ishoda ili bolesti

# Koraci otkrivanja biomarkera

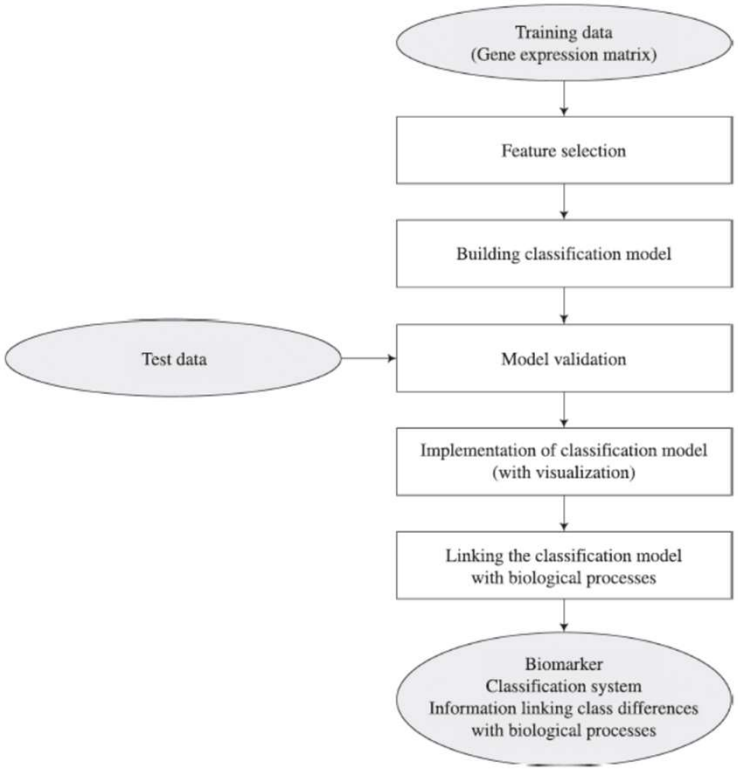



**Figure 3.2:** Elements of biomarker discovery. Please note that this is a simplification of the biomarker discovery process. Often, biomarker discovery includes various iterations and combinations of these elements. In Chapter 4, we will describe how to combine these elements into a process allowing the identification of robust and interpretable multivariate biomarkers.

- Slika ilustruje primer i pojednostavljenje procesa otkrivanja biomarkera
- Često uključuje i iteracije nekih koraka ili njihovih kombinacija

# Ciljevi otkrivanja biomarkera

- Glavni ciljevi otkrivanja biomarkera:
  - Identifikovanje malih podskupova promenljivih (svojstava) koje se mogu koristiti za efikasnu klasifikaciju novih uzoraka
  - Obezbeđenje cost-effective klasifikatora koji se lako mogu ugraditi u kliničku praksu
  - Povezivanje identifikovanih biomarkera i razlika u klasama sa biološkim procesima u osnovi
  - Vizualizacija diskriminativnog prostora i rezultata klasifikacije

# Matrica ekspresije gena... opet

- *N* kolona predstavlja biološke uzorke, *p* vrsta predstavlja gene (promenljive, skupove proba)
  - (proba: supstanca, npr. DNK ili njen deo, koji je obeležen, npr. radioaktivno ili na drugi način, i koji se koristi za otkrivanje druge supstance u uzorku)
- Podaci su prethodno obrađeni, proverenog kvaliteta i sa eliminisanim šumom
- Minimum metapodataka je informacija o klasama dodeljenim uzorcima

**TABLE 3.1: Gene Expression Matrix, *J* Classes, *N* Samples, *p* Variables**

| | Class 1 | | | | | | Class 2 | | | | Class *J* | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | Sample 1 | Sample 2 | Sample 3 | Sample 4 | Sample 5 | ... | ... | ... | ... | ... | Sample $N-1$ | Sample $N$ |
| 1053_at | 6.5248 | 5.7101 | 6.2165 | 5.9984 | 6.2434 | | | | | | 6.8442 | 6.9399 |
| 1316_at | 5.8610 | 5.4029 | 6.5561 | 6.0639 | 5.5503 | | | | | | 5.4944 | 5.1969 |
| 1494_f_at | 5.5814 | 6.0217 | 6.1789 | 6.6498 | 5.9189 | | | | | | 5.9639 | 6.4087 |
| 1729_at | 4.2750 | 4.5192 | 5.0687 | 4.1301 | 4.5680 | | | | | | 6.0076 | 6.4880 |
| 200595_s_at | 9.5081 | 8.6934 | 7.9957 | 8.6349 | 9.0861 | | | | | | 9.1352 | 8.8701 |
| 200600_at | 9.7710 | 8.9780 | 9.1160 | 10.3930 | 9.9285 | | | | | | 10.0923 | 10.5569 |
| ... | | | | | | | | | | | | |
| ... | | | | | | | | | | | | |
| ... | | | | | | | | | | | | |
| 81811_at | 6.4634 | 6.9962 | 6.9654 | 6.8983 | 6.4436 | | | | | | 6.3183 | 4.6599 |
| 89948_at | 8.2129 | 8.1692 | 8.0640 | 7.8943 | 6.5841 | | | | | | 6.1843 | 5.7522 |
| 91617_at | 6.0796 | 5.6061 | 2.8094 | 4.6165 | 5.1416 | | | | | | 6.5173 | 6.0670 |
| 91682_at | 7.6769 | 6.2937 | 7.0522 | 7.1586 | 6.5065 | | | | | | 5.4893 | 6.6147 |
| 91684_g_at | 6.1999 | 6.3268 | 6.5926 | 5.3147 | 5.9832 | | | | | | 6.0896 | 4.9364 |
| 91952_at | 7.2515 | 8.6567 | 7.1778 | 7.2326 | 6.2506 | | | | | | 6.5232 | 6.6493 |

nazad

# Reevaluacija kvaliteta podataka

- Mada su podaci proverenog kvaliteta, dobro je ponovo ispitati sledeće aspekte pre korišćenja kao trening skupa
  - Kvalitet pojedinačnih merenja
  - Kvalitet dodele uzoraka klasama
  - Broj uzoraka u svakoj klasi  (da li je dovoljan za statistička zaključivanja?)
  - Homogenost (ili razuman stepen heterogenosti) klasa
  - Nasumičnost i nezavisnost uzetih uzoraka
  - Koliko dobro uzorci reprezentuju populaciju koju želimo da diferenciramo? Vrlo značajno za generalizaciju rezultata – primenu na istraživanu populaciju a ne samo na trening skup

# Reevaluacija kvaliteta podataka

- Ako je broj klasa nepoznat, nije dobro koristiti prethodno klasterovanje kao osnov za određivanje broja klasa
  - Na primer, ako se u okviru projekta istraživanja podataka u otkrivanju biomarkera postavi sledeći problem:
  - Postoje zanimljivi podaci i namera da se pronađu biomarkeri koji razlikuju dve (ili više) bolesti (ili fenotipa)
  - Ne znaju se sve dijagnoze (ili skup dijagnoza nije siguran)
  - Da li je moguće prvo klasterovati uzorke a zatim izgraditi klasifikatore za identifikovane klase (grupe)?
  - Klasterovanje može da bude vođeno promenljivim koje nisu povezane sa klasama koje želimo da razlikujemo
  - Za mnogodimenzione podatke, interpretacija rezultata klasterovanja je teška i daleko od nivoa kvaliteta koji se zahteva za dobre trening podatke

# Otkrivanje biomarkera i klasifikacija

- Otkrivanje biomarkera i klasifikacija – pregled
- Otkrivanje biomarkera
- Validacija i izveštavanje
- Izbor svojstava (Feature Selection)
- Diskriminantna analiza i izbor svojstava
- SVM i izbor svojstava
- Nasumične šume i izbor svojstava

# Otkrivanje biomarkera

- U kontekstu analize genske ekspresije, otkrivanje biomarkera znači
  - Identifikovanje optimalnog podskupa promenljivih (gena) koje značajno diferenciraju klase i mogu da se upotrebe za tačnu predikciju pripadnosti klasi
  - Najčešće nalaženje dobrog biomarkera znači potragu za skupom promenljivih koje zajedno, kao skup, mogu da razdvajaju klase
  - Heuristički proces ili algoritam koji vodi takvom optimalnom podskupu naziva se ***odabirom (selekcijom) svojstava***

# Nadgledana analiza genske ekspresije - pojmovi

- *Svojstvo = promenljiva*
  - Termin *svojstvo* se koristi kao sinonim za originalnu ulaznu *promenljivu* (na primer, skup proba koji predstavlja gen i pridružen je vrsti u matrici genske ekspresije)
  - U opštem slučaju, svojstvo može da se odnosi na originalnu promenljivu, njihovu kombinaciju ili na promenljivu konstruisanu od originalnih promenljivh
  - U sklopu biomedicinskih aplikacija pogodno je graditi klasifikatore koji direktno koriste neke od originalnih promenljivih
    - Kada originalne promenljive definišu dimenzije diskriminativnog prostora, moguća je neposrednija biološka interpretacija rezultata klasifikacije
    - Stoga će se pod odabirom svojstava podrazumevati odabir optimalnog podskupa originalnih promenljivih

# Nadgledana analiza genske ekspresije - pojmovi

- *Uzorak = biološki uzorak*
  - Bioinformatika je interdisciplinarna oblast
  - Neophodno je uskladiti terminologiju
  - Uzorak će se koristiti u smislu biološkog a ne statističkog uzorka
  - Statistički uzorak će odgovarati grupi bioloških uzoraka koji su odabrani iz – i predstavljaju – u trening skupu - jedne (jednu) od populacija koje treba istraživati

# Nadgledana analiza genske ekspresije - pojmovi

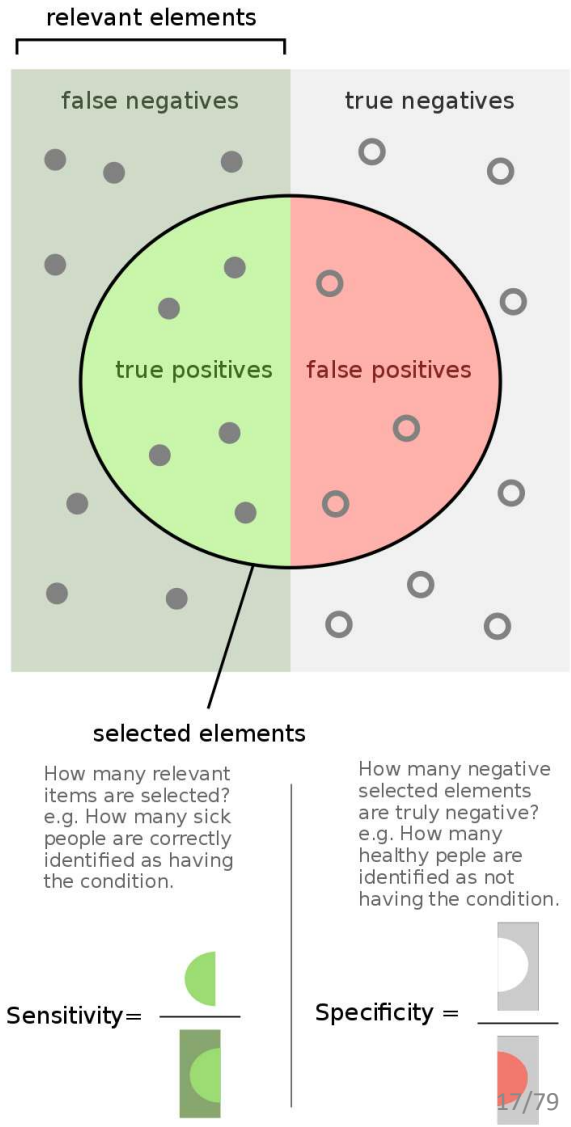
- *Trening podaci*—podaci koji predstavljaju biološke uzorke (npr. pacijenti, tkiva, objekti, opservacije)
  - Karakterišu se izmerenim nivoom nekog broja promenljivih (skupovi proba, geni, egzoni)
  - Poznata je, potvrđena i visoke tačnosti dodela fenotipskoj klasi (bolesti, stanja bolesti, prognoze, odgovori na lečenje)
  - Posle obrade podataka o genskoj ekspresiji, trening skup podataka predstavljen je matricom genske ekspresije
- *Test podaci*—podaci nezavisni od trening podataka (ili bar koji nisu korišćeni za treniranje sistema)
  - Koriste se za ocenu efikasnosti klasifikacionog sistema, posebno stepen kompromisa između preprilagođavanja i generalizacije
- *Preprilagođavanje*—sposobnost klasifikacionog sistema da savršeno ili skoro savršeno klasifikuje trening uzorke
  - ali daje slabe rezultate pri klasifikovanju novih uzoraka

# Nadgledana analiza genske ekspresije - pojmovi

- *Generalizacija*—sposobnost klasifikacionog sistema da korektno klasifikuje uzorke koji nisu uključeni u trening skup
  - najbolje može da se oceni evaluacijom performansi klasifikacionong sistema na nezavisnim test podacima
- *Biomarker* (*opštija definicija*)—biohemijska karakteristika  koja može da se upotrebi za
  - dijagnostikovanje bolesti,
  - predviđanje ishoda,
  - odabir terapije,
  - procenu efiksnosti ili toksičnosti kandidata za lek ili – uopšte
  - predviđanje članstva klase
- *Multivarijantni biomarker (kontekst-specifičnija definicija*)—skup gena sa zadovoljavajućom diskriminativnom moći koji mogu da se upotrebe za stvaranje klasifikacionong sistema visoke osetljivosti i visoke specifičnosti
  - skup gena čiji zajednički obrazac ekspresije je prediktivan za pripadnost klasi

# Osetljivost – specifičnost (sensitivity – specificity)

**Sensitivity** (True Positive rate) meri odnos pozitivnih koji su ispravno identifikovani (tj. odnos onih koji su ispravno identifikovani da ispunjavaju uslov od svih koji zaista ispunjavaju neki uslov)
**Specificity** (True Negative rate) meri odnos negativnih koji su ispravno identifikovani (tj. odnos onih koji su ispravno identifikovani da ne ispunjavaju uslov od svih koji zaista ne ispunjavaju neki uslov)



# Nadgledana analiza genske ekspresije - pojmovi

- *Optimalni multivarijantni biomarker*—ekonomičan multivarijantni biomarker koji obezbeđuje najbolji kompromis između preprilagođavanja i generalizacije
  - Analiza genske ekspresije bavi se hiljadama promenljivih
  - Puna pretraga najboljeg podskupa gena je neizvodljiva
  - Heurističke metode moraju da se koriste za idetifikovanje optimalnih biomarkera
  - Mali skup gena sa zadovoljavajućom i po mogućstvu velikom diskriminativnom moći
- *Klasifikacija vs. predikacija*
  - Kao i termin *uzorak*, termini *predikacija* i *klasifikacija* često se u biomedicinskim istraživanjima koriste drugačije nego u statistici
  - U statistici pridruženi su različitim tipovima promenljivih odgovora
  - Kontinualnim u predikciji, kategoričkim u klasifikaciji
  - U biomedicinskim istraživanjima često se koriste sinonimno

# Primeri tipova biomarkera

- Dijagnostički biomarkeri—indikatori prisustva bolesti ili prisustva specifičnog stanja ili poditpa bolesti
- Prognostički biomarkeri—indikatori verovatnoće specifičnih ishoda tretmana - terapije
- Biomarkeri za personalizovanu medicinu
  - Npr. biomarkeri za odabir terapije ili za minimizovanje rizika od štetnih reakcija na lek
  - Indikatori verovatnoće specifičnog ishoda za razmatranu terapijsku opciju ili lekove
- Biomarkeri toksičnosti—indikatori nivoa toksičnosti leka
  - Često se koriste u toku procesa otkrivanja leka i u toku kliničkih ispitivanja
- Biomarkeri efikasnosti—koriste se u otkrivanju lekova za odabir sastojaka koji najviše obećavaju
- Farmakodinamički i farmakogenomski biomarkeri—indikatori odnosa između odgovora na lek i doze
  - Koriste se za određivanje doze koja odgovara optimalnom odgovoru

# Biomarkeri za personalizovanu medicinu

- Genomski, proteomski i metabolomski biomarkeri imaju veliki potencijal za podršku personalizovanoj medicini
- Ciljevi otkrivanja biomarkera za personalizovanu medicinu
  - Izbor terapije
  - Minimizovanje rizika od štetnih reakcija na lekove
- Prilagođavanje terapije stanju pacijenta posebno značajno u tretmanu karcinoma
- Preterana terapija pacijenata koji imaju mali rizik povratka bolesti može da izazove stanja indukovana terapijom kao što je leukemija
- Izbor najadekvatnijeg protokola lečenja za datog pacijenta može se poboljšati identifikovanjem karakteristika obrazaca ekspresije pridruženih različitim odgovorima na veći broj tretmana
- Za identfikovanje takvih obrazaca potreban je veliki repozitorijum podataka o ekspresiji za pacijente sa poznatom dijagnozom, terapijom i ishodom
- Algoritmi za izbor multivarijantnih svojstava mogu se koristiti za identifikovanje genomskih ili proteomskih biomarkera sa visokom klasifikacionom efikasnošću

# Preprilagođavanje i generalizacija

- Tipični skupovi podataka genske ekspresije sadrže 5000–20,000 promenljivih
- Puna pretraga koja bi garantovala pronalaženje najboljeg podskupa promenljivih ne može se implementirati (složenost $O(2^p)$.
- Mnogi problemi vezani za odabir svojstava su NP-teški
- Neophodan istinski multivarijantni pristup
- Po završetku odabira svojstava imamo jedan ili više potencijalnih biomarkera
- Veliki biomarkeri imaju tendenciju prepilagođavanja trening podacima
- Mali biomarkeri često nemaju dovoljno veliku diskriminativnu moć
- Biomarker i izbor modela – kompromis između preprilagođavanja modela i generalizacije
- Model: “not so simple that it cannot explain the differences between the categories, yet not so complex as to give poor classification of novel patterns”

# Poželjna veličina biomarkera

- Istinski multivarijantni biomarker: ne više od 10 promenljivih
- Ako se zna da su klase koje treba razlikovati – heterogene, može da treba veći broj promenljivih za konstrukciju efikasnog biomarkera (ako je uopšte moguć)
- Ako rezultujući biomarker ima više od 20 promenljivih, treba preispitati
  - Kvalitet trening skupa podataka
  - Homogenost klasa (ili razumnu heterogenost)
  - Dizajn eksperimenta
  - Pretpostavke studije
- Ako studija izveštava o biomarkeru sa više od 30 promenljivih
  - Problemi sa primenjenim pristupima

# Otkrivanje biomarkera i klasifikacija

- Otkrivanje biomarkera i klasifikacija – pregled
- Otkrivanje biomarkera
- Validacija i izveštavanje
- Izbor svojstava (Feature Selection)
- Diskriminantna analiza i izbor svojstava
- SVM i izbor svojstava
- Nasumične šume i izbor svojstava

# Izveštavanje o rezultatima validacije

- Ocenjena efikasnost klasifikacionong sistema može da se prikaže kao izračunata *tačnost* (*accuracy*): procenat korektno klasifikovanih uzoraka
- Stopa pogrešne klasifikacije (*misclassification rate*): *1-tačnost*
- Mere su dovoljne samo kada su klase približnih veličina (slične a*priorne* verovatnoće*)*
- Na primer, razmatraju se dva podtipa jedne bolesti $\Delta$, $\delta_1$ i $\delta_2$
- Ako samo 1% obolelih od $\Delta$ pate od $\delta_2$, naivni klasifikator, koji svaki uzorak svrstava u klasu $\delta_1$, imaće tačnost od 99%
- Ipak, pogrešno će klasifikovati 100% pacijenata koji pate od $\delta_2$
- Detaljnije mere kvaliteta klasifikacije su potrebne

# Tipovi klasifikacije i matrice konfuzije

- Binarna klasifikacija – dve klase, npr. bolest / ne-bolest tj. bolest-pozitivan / bolest-negativan

**TABLE 3.2: Confusion Matrix for a Binary Classifier**

| | | Predicted Class: Disease (positive) | Predicted Class: No Disease (negative) |
|---|---|---|---|
| True Class | Disease (positive) | *True Positive (TP)* | *False Negative (FN)* |
| | No Disease (negative) | *False Positive (FP)* | *True Negative (TN)* |

The rows represent true classes of the test cases. The columns represent classes predicted for them by the classifier.

# Tipovi klasifikacije i matrice konfuzije

- *True Positive* (*TP*)—broj pozitivnih test slučajeva koji su korektno klasifikovani („pogoci“)
- *True Negative* (*TN*)—broj negativnih test slučajeva koji su korektno klasifikovani („koraktna odbacivanja“)
- *False Positive* (*FP*)—broj negativnih test slučajeva nekorektno klasifikovanih („false alarms“)
- *False Negative* (*FN*)—broj pozitivnih test slučajeva koji su nekorektno klasifikovani („promašaji“)
- Broj pozitivnih $P = TP + FN$,
- Broj negativnih $N = TN + FP$

# Parcijalne mere uspešnosti binarne klasifikacije

- *Osetljivost (sensitivity)*: verovatnoća da će klasifikator predvideti „bolest“ kada je tačna klasa – „bolest“
  - Zove se i *stopa stvarno pozitivnih, TPR* ili stopa pogodaka

$$Sensitivity = TPR = \frac{TP}{TP + FN} = \frac{TP}{P}$$

- *Specifičnost (specificity)*: verovatnoća da će klasifikator predvideti „ne-bolest“ kada je tačna klasa „ne-bolest“
  - Zove se i *stopa stvarno negativnih, TNR*

$$Specificity = TNR = \frac{TN}{TN + FP} = \frac{TN}{N}$$

- *Stopa lažno pozitivnih*
  - *False Positive Rate, FPR*

$$FPR = \frac{FP}{FP + TN} = \frac{FP}{N}$$
$$= 1 - Specificity$$

- *Stopa lažno negativnih*
  - *False Negative Rate, FNR*

$$FNR = \frac{FN}{FN + TP} = \frac{FN}{P}$$
$$= 1 - Sensitivity$$

# Parcijalne mere uspešnosti binarne klasifikacije

- *Stopa lažnog otkrivanja* – procenat test slučajeva klasifikovanih kao pozitivni koji su lažno pozitivni
  - *False Discovery Rate*, FDR)

$$FDR = \frac{FP}{FP + TP}$$

- *Tačnost* (*Accuracy*)
  - procenat test slučajeva koji su tačno klasifikovani

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = \frac{TP + TN}{P + N}$$

- *Stopa pogrešne klasifikacije* (*Misclassification Rate*)
  - Procenat svih pogrešno klasifikovanih test slučajeva

$$Misclassification\ Rate = \frac{FP + FN}{TP + TN + FP + FN} = \frac{FP + FN}{P + N}$$
$$= 1 - Accuracy.$$

- *Pozitvna prediktivna vrednost*
  - *Positive Predictive Value*, PPV
  - Procenat slučajeva koji su klasifikovani kao pozitivni a koji su stvarno pozitivni

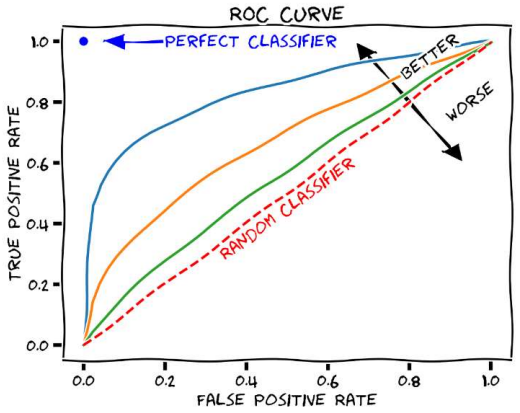
$$PPV = Precision = \frac{TP}{TP + FP}$$
$$= 1 - FDR$$

- *Negativna prediktivna vrednost*
  - *Negative Predictive Value*, NPV
  - Procenat slučajeva koji su klasifikovani kao negativni a koji su stvarno negativni

$$NPV = \frac{TN}{TN + FN}.$$

# Naivni binarni klasifiaktor - revizija

- Obeležimo podtip $\delta_1$ bolesti $\Delta$ kao negativni ishod a $\delta_2$ kao pozitivni
- Specificity=100%
- Sensitivity=0%
- Neprihvatljivo, posebno kada je bolest - $\delta_2$ - mnogo opasnija klasa od $\delta_1$
- Treba uzeti u obzir moguće različite cene različitih stopa pogrešne klasifikacije
- Može da se uključi u proces učenja, kroz težine trening slučajeva ili usklađivanje veličina klasa
- Traži se najprihvatljivija nagodba između osetljivosti i specifičnosti
- Vizualizacija ove nagodbe za različite vrednosti parametara binarne klasifikacije:
  - Receiver Operator Characteristic (ROC) kriva
- Obično: osetljivost naspram 1-specifičnost
- Površina ispod ROC krive (**Area Under The Curve**, AUC) služi za poređenje modela



# Višeklasni klasifikatori

$N_i$ – broj test slučajeva čija je tačna klasa *i*
(zbir vrste *i*)

$$N_i = \sum_{j=1}^{J} C_{ij}$$

$P_j$ – broj test slučajeva klasifikovanih u klasu *j*
(zbir kolone *j*)

$$P_j = \sum_{i=1}^{J} C_{ij}$$

*N* – ukupan broj slučajeva u test skupu

$$N = \sum_{i=1}^{J} N_i = \sum_{j=1}^{J} P_j = \sum_{i=1}^{J} \sum_{j=1}^{J} C_{ij},$$

**TABLE 3.3: Confusion Matrix for a Multiclass Classifier**

| | | Predicted Class | | | |
|---|---|---|---|---|---|
| | | Class 1 | Class 2 | ... | Class *J* |
| True Class | Class 1 | $C_{11}$ | $C_{12}$ | ... | $C_{1J}$ |
| | Class 2 | $C_{21}$ | $C_{22}$ | ... | $C_{2J}$ |
| | ... | ... | ... | ... | ... |
| | Class *J* | $C_{J1}$ | $C_{J2}$ | ... | $C_{JJ}$ |

The rows of the confusion matrix represent true classes. The columns represent the predicted classes.

# Višeklasni klasifikatori

- Osetljivost (Sensitivity) klase *k* – procenat test slučajeva klase *k* koji su korektno klasifikovani u klasu *k*
  - *k=1,2,...,J*

$$Sensitivity(k) = \frac{C_{kk}}{N_k}$$

- Specifičnost (Specificity) klase *k* – procenat test slučajeva koji nisu u klasi *k* koji su klasifikovani u ne-klasu *k*
  - *k=1,2,...,J*

$$Specificity(k) = \frac{N - P_k}{N - N_k}.$$

- Sveukupna tačnost i stopa pogrešne klasifikacije višeklasnog klasifikatora

$$Accuracy = \frac{\sum_{k=1}^{J} C_{kk}}{N},$$

$$Misclassification\ Rate = 1 - Accuracy.$$

# Otkrivanje biomarkera i klasifikacija

- Otkrivanje biomarkera i klasifikacija – pregled
- Otkrivanje biomarkera
- Validacija i izveštavanje
- Izbor svojstava (Feature Selection)
- Diskriminantna analiza i izbor svojstava
- SVM i izbor svojstava
- Nasumične šume i izbor svojstava

# Izbor svojstava

- „Manje je više“ – suština izbora svojstava
- Za tipični skup podataka mikronizova genske ekspresije broj promenljivih je u hiljadama
- Iscrpno pretraživanje multivaijantnog markera koji ima do, na primer, 10 gena je izuzetno skupo
- Cilj izbora svojstava – naći mali podskup promenljivih koji značajno razdvaja razlikovane populacje (predstavljene klasama uzoraka u trening skupu)
- Potrebne efikasne heurističke metode za
  - (i) uklanjanje irelevantnih i redundantnih promenljivih,
  - (ii) pronalaženje optimalnog skupa promenljivih – minimizovanjem veličine skupa i maksimizovanjem njegove diskriminativne moći

# Izbor svojstava

- Stabilnija rešenja se obično postižu ansambl klasifikatorima
  - Za klasifikaciju novog uzorka koriste se svi ili podskup identifikovanih biomarkera
  - Rezultati klasifikacije baziraju se na shemi glasanja
  - Bolji pristup korišćenju identifikovanih potencijalnih biomarkera može da bude korišćenje ansambl klasifikatora za glasanje o samim svojstvima (umesto o klasama)
- Univarijantni vs. multivarijantni pristupi
  - Npr, uzimati jedan po jedan gen i određivati kako genska ekspresija diskriminiše klase
    - samo ako su promenljive nekorelisane ili
    - Iz nekog razloga nas interesuje istraživanje izolovanog gena
    - Ne važi u slučaju otkrivanja biomarkera iz podataka genske ili proteinske ekspresije

# Univarijantni vs. multivarijantni pristupi

- Koregulacije i interakcije između gena (ili proteina) su značajne
- Geni koji su daleko od vrha univarijantno uređene liste, čija ih *p*-vrednost (ili druga mera statističke značajnosti) čini univarijantno beznačajnim mogu da nose značajnu diskriminativnu informaciju kada se njihov obrazac ekspresije kombinuje sa obrascem ekspresije drugih gena

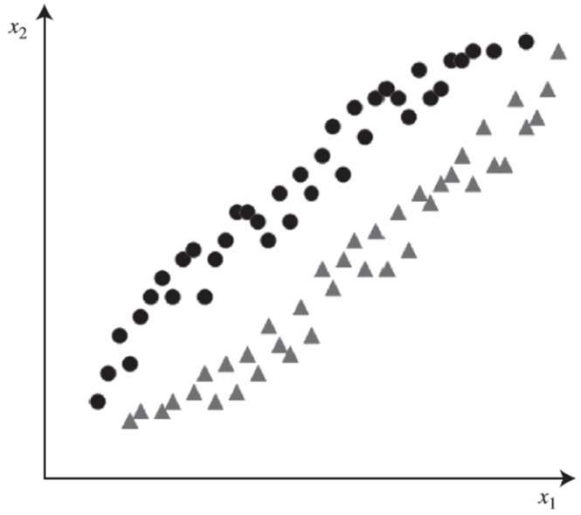



**Figure 3.4:** A data set with two classes of samples. There are only two variables. Neither of them is univariately significant for the class discrimination. However, as a set of two variables they can perfectly separate the classes. (See color insert.)

# Univarijantni vs. multivarijantni pristupi

- Multivarijantni pristupi evaluiraju skupove promenljivih
- Razmatraju interakcije među promenljivim
- Sposobni su da pronađu male skupove promenljivih sa visokom diskriminativnom moći  (ako takav skup postoji)
- Obično, neke od promenljivih u skupu su univarijanto značajne dok druge nisu

# Taksonomija metoda za izbor svojstava (Feature Selection)

- Kriterijumi
  - Modeli pretrage definisani odnosima (ili odsustvom odnosa) između izbora svojstava i algoritma klasifikacije (filter, omotač, hibridni, ugnježđeni modeli)
  - Pristup učenju (nadgledane i nenadgledane metode)
  - Da li se uzimaju u obzir interakcije između promenljivih (univarijantne i multivarijantne metode)
  - Strategije pretrage (iscrpna, sekvencijalna, nasumična, hibridna)
- Neke kategorije—univarijantne i nenadgledane metode—neprikladne su za otkrivanje biomarkera na bazi podataka genske ili proteinske ekspresije

# Stabilnost biomarkera

- Kada se izbor svojstava vrši na raznim verzijama trening skupa korišćenjem različitih algoritama izbora svojstava, rezultujući multivarijantni biomarkeri su najverovatnije različiti
- Ne znači da je rešenje nestabilno
- Stabilnost biomarkera nije isto što i jednakost skupova
- Za stabilnost rešenja bitno je da se biomarkeri sastoje od gena koji ukazuju na iste skupove bioloških procesa koji razlikuju klase
- Stabilnost biomarkera može da se definiše u terminima ekvivalencije bioloških procesa predstavljenih skupovima gena koji su izabrani u biomarkere
- Identifikacija bioloških procesa – netrivijalan zadatak, sledi za otkrivanjem biomarkera
- Zbog toga, u smislu stabilnosti, razmatra se da li se biomarkeri sastoje od gena koji predstavljaju iste primarne obrasce ekspresije pridružene razlikama klasa

# Otkrivanje biomarkera i klasifikacija

- Otkrivanje biomarkera i klasifikacija – pregled
- Otkrivanje biomarkera
- Validacija i izveštavanje
- Izbor svojstava (Feature Selection)
- Diskriminantna analiza i izbor svojstava
- SVM i izbor svojstava
- Nasumične šume i izbor svojstava

# Diskriminantna analiza (DA)

- Stara metoda (1930-te dvoklasna, 1940-te višeklasna)
- Prvo u biologiji i medicini
- Dva tipa:
  - prediktivna DA – fokusira se na predviđanje pripadnosti klasi
  - deskriptivna DA – fokus na interpretaciji klasnih razlika, tj. linearne diskriminantne funkcije pridružene razlikama klasa

# Osnovne pretpostavke

- Nezavisnost bioloških uzoraka
- Multivarijantna normalnost – normalna raspodela svake promenljive i njihove linearne kombinacije
- U tipičnim podacima genske ekspresije, hiljade promenljivih – nepraktično testirati normalnost
- Obično se proverava samo univarijantna normalnost ili
- Podaci se preprocesiraju tako da se povećaju izgledi za multivarijantnom normalnom raspodelom
- Homogenost kovarijansi klasa
  - Rezultat: linearna diskriminantna analiza (LDA)
  - Ako matrice kovarijansi klasa nisu jednake, kvadratna diskriminantna analiza (QDA)

# Osnovne pretpostavke

- Nema singulariteta ili multikolinearnosti
  - Singularitet se odnosi na poptpuno redundantnu promenljivu
  - Multikolinearnost se odnosi na visoko korelisane promenljive
  - Uz singularitet, matrice kovarijanse nemaju inverz
  - Uz multikolinearnost, rezultat inevrtovanja matrice je nestabilan
- Nema ekstremnih elemenata van granica koji bi imali veliki uticaj na srednju vrednost i povećali varijabilnost

# Algoritam učenja

- Trening skup: $N$ tačaka podataka (bioloških uzoraka) i $p$ promenljivih $x_1, \ldots, x_p$ – npr. nivoi ekspresije $p$ gena
- Svakoj tački podataka dodeljena je jedna od $J$ klasa (populacija)
- Svaka klasa $j$ uključuje $n_j$ tačaka tako da je $N = \sum_{j=1}^{J} n_j$.
- Trening tačka koja pripada klasi $j$: $p$-dimenzioni vektor $x_{ji}$, $\mathbf{x}_{ji} = \begin{bmatrix} x_{1ji} \\ x_{2ji} \\ \vdots \\ x_{pji} \end{bmatrix}$
- Matrica ekspresije gena X može se predstaviti u obliku

$$\mathbf{X} = \left(\begin{array}{ccc|ccc|c|ccc} x_{111} & \cdots & x_{11n_1} & x_{121} & \cdots & x_{12n_2} & \cdots & x_{1J1} & \cdots & x_{1Jn_J} \\ x_{211} & \cdots & x_{21n_1} & x_{221} & \cdots & x_{22n_2} & \cdots & x_{2J1} & \cdots & x_{2Jn_J} \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{p11} & \cdots & x_{p1n_1} & x_{p21} & \cdots & x_{p2n_2} & \cdots & x_{pJ1} & \cdots & x_{pJn_J} \end{array}\right)$$

- (Matrica ekspresije gena... Opet)

# Algoritam učenja

- Vektor srednje vrednosti trening tačaka klase *j*: $\overline{\mathbf{x}}_j = \frac{1}{n_j}\sum_{i=1}^{n_j}\mathbf{x}_{ji}$
  - Nepristrasna ocena vektora srednje vrednosti populacije $\boldsymbol{\mu}_j$ za klasu *j*
- Vektor srednje vrednosti svih *N* trening tačaka:
  - Ocena sveukupnog vektora srednje vrednsti $\boldsymbol{\mu}$ svih klasa zajedno $\overline{\overline{\mathbf{x}}} = \frac{1}{N}\sum_{j=1}^{J}\sum_{i=1}^{n_j}\mathbf{x}_{ji} = \frac{1}{N}\sum_{j=1}^{J} n_j\overline{\mathbf{x}}_j$
- Matrica ocenjenih kovarijansi za klasu *j* $\mathbf{S}_j = \frac{1}{n_j - 1}\sum_{i=1}^{n_j}(\mathbf{x}_{ji} - \overline{\mathbf{x}}_j)(\mathbf{x}_{ji} - \overline{\mathbf{x}}_j)^T$

# Algoritam učenja

- Pod pretpostavkom da su biološki uzorci nezavisni i normalno raspoređeni unutar klasa sa zajedničkom matricom kovarijansi $\Sigma$, zajednička ocena za $\Sigma$ je

$$\mathbf{S} = \frac{1}{N-J}\sum_{j=1}^{J}\sum_{i=1}^{n_j}(\mathbf{x}_{ji} - \overline{\mathbf{x}}_j)(\mathbf{x}_{ji} - \overline{\mathbf{x}}_j)^T$$

- ili

$$\mathbf{S} = \frac{1}{N-J}\sum_{j=1}^{J}(n_j - 1)\mathbf{S}_j$$

# Diskriminativna moć

- Potrebno je definisati *diskriminativna moć* skupa *p* promenljivih kao meru separacije klasa u *p*-dimenzionom prostoru
- Interpretacija: vrednost test statistike koja meri odstupanje od nulte hipoteze $H_0$ jednakosti *J* centroida klasa $\boldsymbol{\mu}_j, j = 1, \ldots, J$ (*p*-dimenzioni vektori koji predstavljaju populacijske srednje vrednosti)

$$H_0: \quad \boldsymbol{\mu}_1 = \boldsymbol{\mu}_2 = \cdots = \boldsymbol{\mu}_J$$
$$H_a: \quad \boldsymbol{\mu}_j \neq \boldsymbol{\mu}_k \text{ for some } j \neq k$$

- Kako odabrati metriku (meru) diskriminativne moći?

# Diskriminativna moć

- Jedna od mera multivarijantne diskriminativne moći skupa od *p* promenljivih je *Lawley-Hotelling trag statistika* $T^2$:

$$T^2 = T^2(x_1, \ldots, x_p) = tr(\mathbf{HE}^{-1}).$$

- **H** je *p*x*p* matrica „hipoteza“ koja opisuje varijabilnost između klasa $\mathbf{H} = \sum_{j=1}^{J} n_j(\bar{\mathbf{x}}_j - \bar{\bar{\mathbf{x}}})(\bar{\mathbf{x}}_j - \bar{\bar{\mathbf{x}}})^T$

- **E** je *p*x*p* matrica „grešaka“ koja opisuje varijabilnost unutar klasa $\mathbf{E} = \sum_{j=1}^{J} \sum_{i=1}^{n_j} (\mathbf{x}_{ji} - \bar{\mathbf{x}}_j)(\mathbf{x}_{ji} - \bar{\mathbf{x}}_j)^T$

(E=(*N*-*J*)S, S - ocena matrice kovarijansi $\Sigma$)

- *tr*(**A**) je trag kvadratne matrice **A**; za matricu *k*x*k*, $tr(\mathbf{A}) = \sum_{i=1}^{k} a_{ii}$
- $\mathbf{E}^{-1}$ je inverz matrice **E**

# Statistika $T^2$

- Maksimizovanje mere diskriminativne moći $T^2$ znači maksimizovanje varijacije između klasa u odnosu na varijacije unutar klasa
- Tačna raspodela $T^2$ je kompleksna
- Aproksimacije korišćenjem $\chi^2$ ili *F* raspodele, na primer $$F = \frac{t(N - J - p - 1) + 2}{t^2 b} tr(\mathbf{H}\mathbf{E}^{-1}).$$

- U slučaju da je nulta hipoteza tačna, prethodna aproksimacija $T^2$ ima *F* raspodelu

# Statistika $T^2$

- Uvek važi $T^2 \geq 0$;
- $T^2=0$ znači da su klase nerazdvojive pomoću $p$ promenljivih
- Povećana vrednost $T^2$ odgovara povećanoj separabilnosti klasa
- $T^2$ vrednosti mogu da se koriste za direktno poređenje diskriminativne moći skupova podataka sa različitim brojem promenljivih i različitim brojem bioloških uzoraka
- $T^2$ metrika je monotona, tj. $$T^2(x_1) \leq T^2(x_1, x_2) \leq \cdots \leq T^2(x_1, \ldots, x_p)$$

- $T^2$ može da se računa samo kada je $p< N-J-1$
- U slučaju podataka genske ekspresije, za npr. $p$=5000, $N=10$ i $J=3$, metrika može da se koristi ali ne za svih $p$ promenljivih istovremeno
- Cilj: otkrivanje biomarkera, tj. malog skupa promenljivih koji
  - dovoljno razdvaja klase
  - može efikasno da se koristi u klasifikovanju novih slučajeva

# Statistika $T^2$

- Pretpostavka: imamo optimalni multivarijantni biomarker od *p* promenljivih i *p* manje od *N-J-1* (npr. *p* = 10 promenljivih)
- Kako da izgradimo klasifikacioni sistem baziran na multivarijantnom biomarkeru
- Klasifikator može da se upotrebi i za validaciju biomarkera a zatim i za klasifikaciju novih uzoraka
- Klasifikacija novog uzorka: klasa najbližeg centroida (vektora srednje vrednosti klase)
- Na primer, Mahalanobis rastojanje $D_j^*$ između tačke $\mathbf{x}=[x_1,\ldots,x_p]^T$ koja se klasifikuje i centroida klase *j*, $\bar{\mathbf{x}}_j$,

$$D_j^* = \sqrt{(\mathbf{x}-\bar{\mathbf{x}}_j)^T\mathbf{S}^{-1}(\mathbf{x}-\bar{\mathbf{x}}_j)}.$$

- Uzorak se klasifikuje u klasu koja odgovara najmanjem $D_j^*$

# Statistika $T^2$ – smanjenje dimenzionalnosti

- Za biomarkere sa brojem promenljivih $p>3$, nije moguća grafička reprezentacija celog $p$-dimenzionog diskriminativnog prostora
- Dimenzionalnost prostora može da se smanji rešavanjem generalizovanog problema sopstvenih vrednosti $\mathbf{Hv} = \lambda \mathbf{Ev}$
- Problem ima $t=\min(p, J-1)$ nenultih sopstvenih vrednosti $\lambda_1, \lambda_2, \ldots, \lambda_t$

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_t$$

- Za $J \leq p$, rešenje problema omogućuje transformaciju $p$-dimenzionog prostora originalnog biomarkera u $t$-dimenzioni prostor koji ima $t$ linearnih diskriminativnih funkcija
- Rezultujući $t$-dimenzioni prostor predstavlja istu diskriminativnu informaciju kao i originalni $p$-dimenzioni prostor
- t diskriminativnih funkcija su linearne kombinacije $p$ promenljivh tog biomarkera

# Statistika $T^2$ – smanjenje dimenzionalnosti

- Na primer, prva diskriminativna funkcija $f_1$ pridružena najvećoj sopstvenoj vrednosti $\lambda_1$ i njenom *p*-dimenzionom sopstvenom vektoru $\mathbf{v_1}$

$$\mathbf{v}_1 = \begin{bmatrix} v_{11} \\ v_{21} \\ \vdots \\ v_{p1} \end{bmatrix}$$

- definisana je sa

$$\begin{aligned} f_1 &= v_{11}x_1 + v_{21}x_2 + \cdots + v_{p1}x_p \\ &= \sum_{l=1}^{p} v_{l1}x_l \\ &= \mathbf{v}_1^T\mathbf{x}. \end{aligned}$$

- Ova funkcija transformiše proizvoljnu *p*-dimenzionu tačku **x** u jednu dimenziju definisanu sa $f_1$ – *svojstvo (feature)*, nova promenljiva
- Elementi vektora $\mathbf{v_1}$ su težine ove linearne transformacije pridružene originalnim promenljivim $x_1,\ldots, x_p$

# Statistika $T^2$ – smanjenje dimenzionalnosti

- Matrica $\mathbf{V}_{pxt}$ čije su kolone – *t* sopstvenih vektora $\mathbf{v}_1, \mathbf{v}_2,\ldots,\mathbf{v}_t$ pridruženih sopstvenim vrednostima

$$\mathbf{V} = \begin{bmatrix} v_{11} & v_{12} & \cdots & v_{1t} \\ v_{21} & v_{22} & \cdots & v_{2t} \\ \vdots & \vdots & \ddots & \vdots \\ v_{p1} & v_{p2} & \cdots & v_{pt} \end{bmatrix}.$$

- Matrica *V* uključuje težine za svih *t* linearnih diskriminativnih funkcija $f_1, \ldots f_t$
- Uzorak koji se klasifikuje (npr. pacijent koji se dijagnostikuje), predstavljen vektorom **x** u *p*-dimenzionom prostoru *p* promenljivih biomarkera,

$$\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_p \end{bmatrix}$$

- može se predstaviti vektorom **w** u prostoru *t* svojstava definisanih pomoću *t* diskriminativnih funkcija, gde $\mathbf{w} = \mathbf{V}^T\mathbf{x}$

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_t \end{bmatrix}$$

# Statistika $T^2$ – smanjenje dimenzionalnosti

- diskriminativna moć t svojstava rezultujućeg prostora ista je kao i diskriminativna moć p originalnih promenljivih biomarkera

$$T^2(w_1, \ldots, w_t) = T^2(x_1, \ldots, x_p).$$

- Klasifikacioni model se gradi u ovom t-dimenzionom prostoru
- Klase – *t*-dimenzione hipersfere
- Uzorci za klasifikaciju – *t*-dimenzioni vektori
- Obično je $J \leq 4$
- Moguće je svu diskriminativnu informaciju predstaviti u 3 dimenzije ili manje
- Za binarnu klasifikaciju ($J=2$), diskriminativni prostor je samo jedna osa

# Statistika $T^2$ – klasifikacija

- Da bi se novi uzorak klasifikovao, ocenjuje se verovatnoća njegove pripadnosti svakoj od *J* klasa
- Pod pretpostavkom da je vektor **w**=[$w_1$,...,$w_t$]$^T$ koji predstavlja nepoznati uzorak – centar $\mu_0$ hipotetičke klase *j*=0, izvodimo *J* testova značajnosti za *j*=1,...,*J*

$$H_0: \quad \boldsymbol{\mu}_0 = \boldsymbol{\mu}_j$$

$$H_a: \quad \boldsymbol{\mu}_0 \neq \boldsymbol{\mu}_j$$

- Može da se koristi test statistika

$$F_j = \frac{n_j}{n_j + 1} \cdot \frac{N - J - t + 1}{t(N - J)} (\mathbf{w} - \overline{\mathbf{w}}_j)^T (\mathbf{w} - \overline{\mathbf{w}}_j).$$

- gde je $\overline{\mathbf{w}}_j$ centroid klase *j* u *t*-dimenzionom diskriminativnom prostoru

$$\overline{\mathbf{w}}_j = \mathbf{V}^T \overline{\mathbf{x}}_j.$$

# Statistika $T^2$ – klasifikacija

- $F_j$ statistika ima $F$ raspodelu sa $t$ i $N$-$J$-$t$+1 stepeni slobode
- Za nivo značajnosti $\alpha$, uzorak pripada klasi $j$ ako je $F_j \leq F_\alpha$
- Dakle, uzorak može da pripada jednoj, ni jednoj, ili većem broju klasa
- Alternativa: uzorak pripada klasi sa najmanjom vrednošću $F_j$

- Hipersfera sa poluprečnikom $R_j$ sadrži (1-$\alpha$)·100% uzoraka koji pripadaju klasi $j$

$$R_j = \sqrt{F_\alpha \cdot \frac{n_j + 1}{n_j} \cdot \frac{t(N - J)}{N - J - t + 1}}$$

# Statistika $T^2$ – klasifikacija

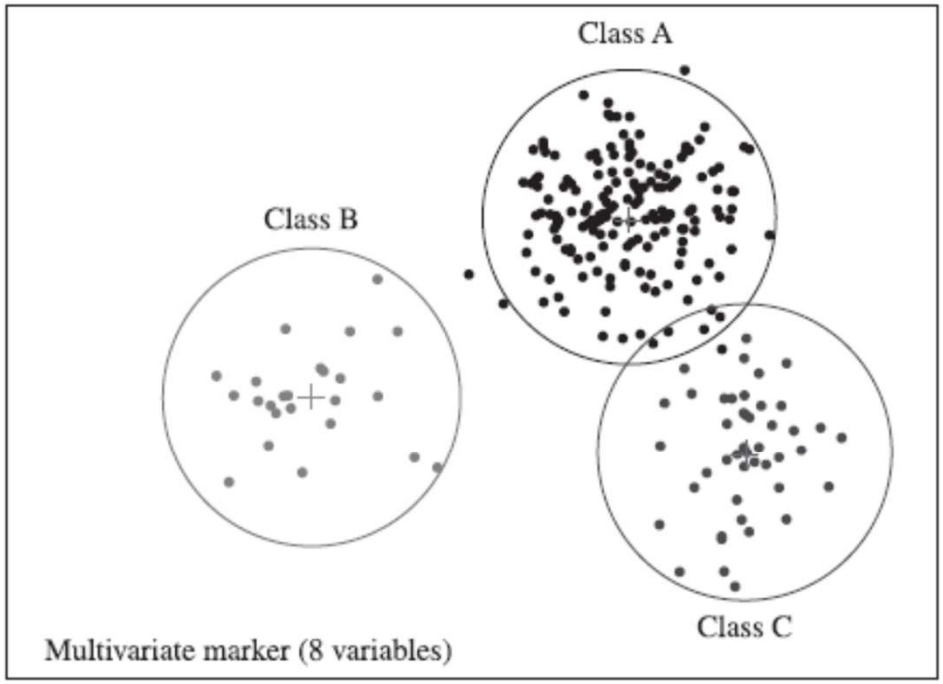



**Figure 3.9:** Discriminatory space of a classification model built on an eight-gene multivariate biomarker ($p = 8$). For this three-class model ($J = 3$), the discriminatory space is two-dimensional, $t = \min(p, J - 1) = 2$. The circles represent constant density boundaries enclosing 95 percent of the probability for each class. Points represent samples from the training data set. (Graphics from the *MbMD* data mining software.)[20] (See color insert.)

# Postepeni hibridni izbor svojstava sa $T^2$

- U polaznom skupu je *p'* promenljivih
- Početi sa jednom promenljivom, diskriminativno najmoćnijom ili nasumičnom
- Na sledećim koracima po jedna promenljiva se dodaje ili izbacuje iz biomarker skupa – u cilju maksimizovanja diskriminativne moći skupa *p* promenljivih (*p*=1,2,3,...)
- Diskriminativna moć svakog skupa računa se $T^2$ metrikom razdvajanja klasa
- Na svakom koraku, prvo se primenjuje dodavanje – jedna od *p'* – *p* +1 promenljivih, koja maksimizuje $T^2$ diskriminativnu moć skupa *p* promenljivih, dodaje se skupu prethodno odabranih *p*-1 promenljivih
- Zatim (za *p*>2) se traži podskup od *p*-1 promenljivih koji je optimalniji od prethodnog *p*-1 skupa
- Kriterijumi kraja $T^2 \geq stop\_\ T^2$ ili $p=stop\_p$ (za parametre $stop\_\ T^2$, $stop\_p$ )

# Postepeni hibridni izbor svojstava - pseudokod

**TABLE 3.4: C-Style Pseudocode of the Stepwise Hybrid Feature Selection Algorithm**

```
1  // Stepwise hybrid feature selection
2  // input:    trainingData(N, p, J)  training data set: p variables, N samples, J classes
3  //           stop_p, stop_T2        stopping criteria
4  //           randomFlag             starting point flag
5  // output:   currentSet             an optimal subset
6  pool ← trainingData(N, p, J);
7  currentSet ← NULL;
8  poolSize = p; markerSize = 0; currentT2 = 0.0;
9  for (step = 1;     markerSize < stop_p && currentT2 < stop_T2   &&
10                    (markerSize < (N-J-2) || (N-J-2) == 0) && markerSize < p;
11                    step++ ) {
12   maxGain = 0.0;
13   markerSize++;
14   if (markerSize == 1 && randomFlag) {
15      selectedVar = selectRandomVar(pool); // random first variable
16      maxGain = calculateT2(selectedVar);
17   } else {
18         // Forward selection: add the variable that maximizes T^2 of this step.
19         for (i=1; i<=poolSize; i++) {
20            deltaT2 = calculateT2( currentSet + poolVar (i)) - currentT2;
21            if (deltaT2 >= maxGain) {
22               maxGain = deltaT2;
23               selectedVar = poolVar(i);
24            }
25         }
26   }
27   currentT2 += maxGain;
28   poolToMarker(selectedVar);      // move selected variable from pool to currentSet
29   poolSize--;
30   pEmpirical = bootstrap(currentT2); // Bootstrap estimate of the T^2 p-value
31   pF = pValueF(currentT2);           // p-value from F distribution (w/Bonferroni)
32   // Calculate multivariate significance of the added variable
33   member_F = memberSignificance(currentSet, selectedVar);
34   // Backward optimization: if elimination of any variable results in T^2 greater
35   // than that of the previous step, eliminate one that minimally decreases T^2.
36   minLoss = currentT2;
37   if (markerSize > 2 ) {
38      for (i=1; i<markerSize; i++) {
39         deltaT2 = currentT2 – calculateT2( currentSet – setVar(i));
40         if (deltaT2 <= minLoss) {
41            minLoss = deltaT2;
42            removedVar = setVar(i);
43         }
44      }
```

*(Continued)*

# Postepeni hibridni izbor svojstava - pseudokod

**TABLE 3.4:** *Continued*

```
45          if (minLoss < maxGain) {
46              markerToPool(removedVar); // move variable from currentSet to pool
47              poolSize++; markerSize−−;
48              currentT2 −= minLoss;
49          }
50          saveSet(currentSet);      // save set optimal for the current cardinality
51      }
52  }
53  return currentSet;
```

After the algorithm finishes, we are presented with the marker of size *stop_p* or with a marker with fewer variables but satisfactory discriminatory power of $T^2 \geq stop\_T^2$. The optimal sets for all considered cardinalities are also available for eventual selection of one of them as our optimal marker.

# Otkrivanje biomarkera i klasifikacija

- Otkrivanje biomarkera i klasifikacija – pregled
- Otkrivanje biomarkera
- Validacija i izveštavanje
- Izbor svojstava (Feature Selection)
- Diskriminantna analiza i izbor svojstava
- SVM i izbor svojstava
- Nasumične šume i izbor svojstava

# SVM

- Linearno separabilne klase

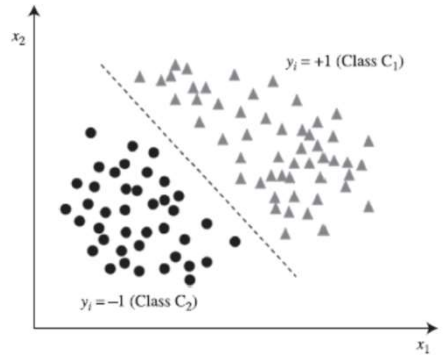



**Figure 3.10:** A training data set with two input variables $(x_1, x_2)$ and two classes; $p = 2, J = 2$. The classes are linearly separable.

- SVM sa tvrdom marginom

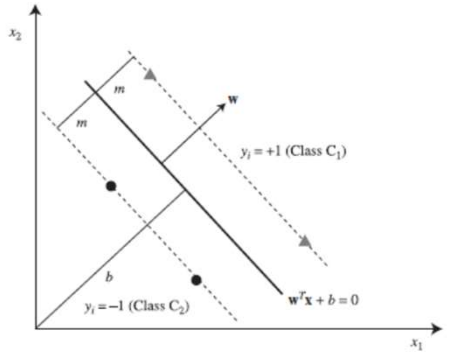



**Figure 3.11:** Separating hyperplane $\mathbf{w}^T\mathbf{x} + b = 0$. The margin $m$ is the functional distance between the separating hyperplane and the training data point(s) nearest to the hyperplane. Elements of the vector $\mathbf{w}$ are weights associated with the $p$ variables. The scalar $b$ defines the offset of the hyperplane from the origin. Euclidean distances corresponding to $m$ and $b$ are $m/\|\mathbf{w}\|$ and $|b|/\|\mathbf{w}\|$.

- Optimizacioni problem

$$\underset{\mathbf{w},b}{\text{minimize}} \; \|\mathbf{w}\|^2$$

$$\text{subject to} \;\; y_i(\mathbf{w}^T\mathbf{x}_i + b) \geq 1 \quad \text{for all training points } \mathbf{x}_i, i = 1, \ldots, N.$$

# SVM – dualna reprezentacija optimizacionog problema

$$\text{maximize} \quad W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2} \sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to} \quad \sum_{i=1}^{N} y_i \alpha_i = 0,$$

$$\alpha_i \geq 0, \quad \text{for } i = 1, \ldots, N.$$

Lagranžijan, tačka prevoja

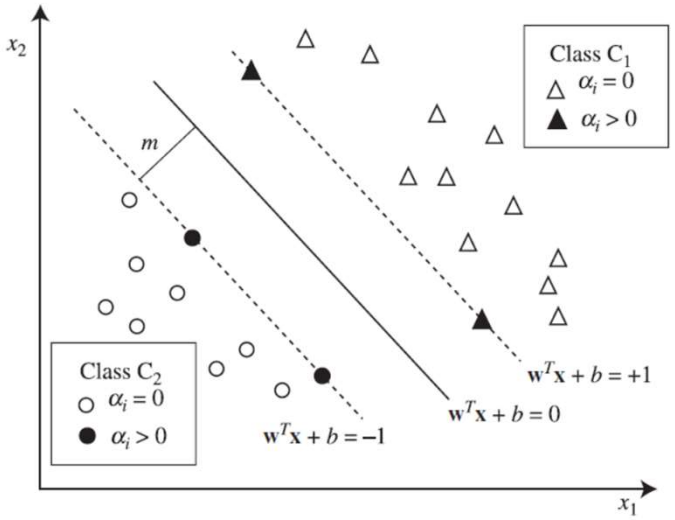



**Figure 3.13:** Only the training data points that lie on one of the support hyperplanes $\mathbf{w}^T\mathbf{x} + b = \pm 1$ may have nonzero values of the Lagrange multipliers $\alpha_i$, $\alpha_i > 0$. The training data points with nonzero $\alpha_i$ are called *support vectors*.

$$f(\mathbf{x}) = sign(\mathbf{w}^{*T}\mathbf{x} + b^*)$$

$$= sign\left(\sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i^T \mathbf{x} + b^*\right).$$

# SVM – meka margina

SVM - meka margina

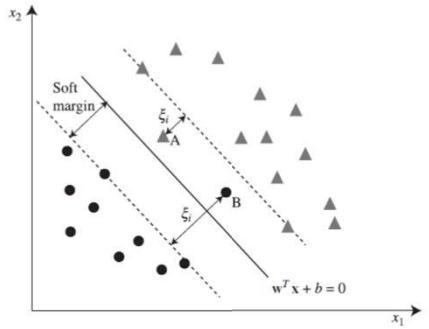



**Figure 3.14:** The slack variables $\xi_i$ are measures of the margin violations for the training data points. Training points violating the margin may be correctly classified (like the point A). If they are, however, on the wrong side of the separating hyperplane (like the point B), they are misclassified and have $\xi_i > 1$. The Euclidean distance corresponding to a $\xi_i$ equals $\xi_i/\|\mathbf{w}\|$.

Dualna reprezentacija

$$\text{maximize} \quad W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

$$\text{subject to} \quad \sum_{i=1}^{N} y_i \alpha_i = 0,$$

$$0 \le \alpha_i \le C, \quad \text{for } i = 1, \ldots, N.$$

Karush-Kuhn-Tucker uslovi:

$$\alpha_i[\, y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1 + \xi_i] = 0 \text{ for } i = 1, \ldots, N$$
$$\xi_i(\alpha_i - C) = 0 \text{ for } i = 1, \ldots, N.$$

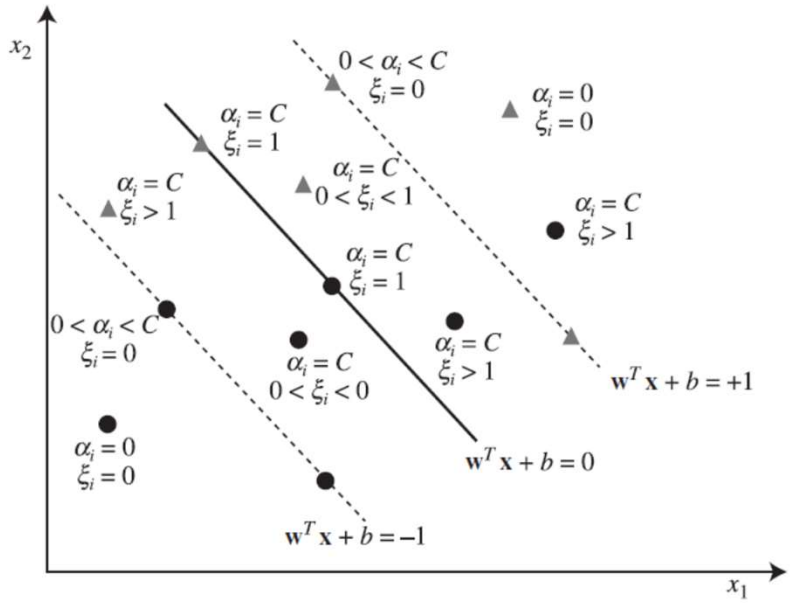



**Figure 3.15:** This figure represents a soft-margin SVM. The training data points with $\xi_i > 0$ violate their margin. It they have $\xi_i > 1$, they are on the wrong side of the separating hyperplane and are misclassified. They all have $\alpha_i = C$ and are called bounded support vectors (since the $\alpha_i$ multipliers associated with them reach the upper bound value defined by $C$). The support vectors that are on the support hyperplanes are called unbounded support vectors; they have $0 < \alpha_i < C$ and $\xi_i = 0$.

# SVM - kerneli

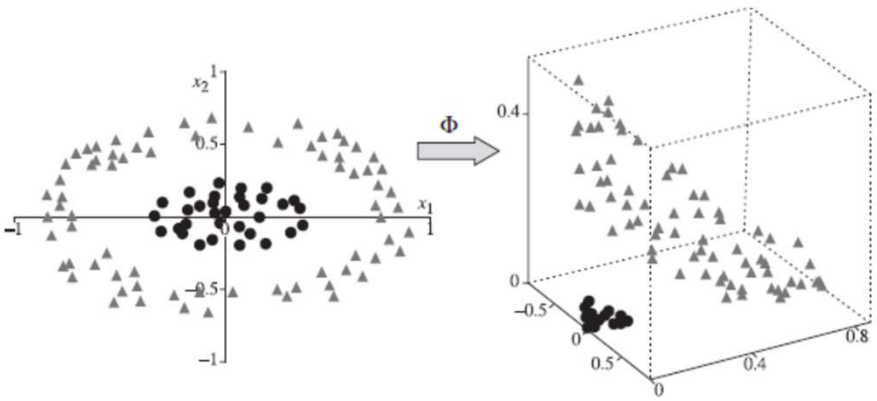



**Figure 3.18:** Mapping the data into a higher-dimensional space can make the classes linearly separable.

$$\Phi\colon (x_1, \ldots, x_p) \longrightarrow (z_1, \ldots, z_s). \qquad z_i = \Phi(x_i), \quad z = \Phi(x),$$

$$\Phi\colon (x_1, x_2) \longrightarrow (z_1, z_2, z_3),$$

$$z_1 = x_1^2; \quad z_2 = \sqrt{2}x_1x_2; \quad z_3 = x_2^2$$

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{N} \alpha_i y_i \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}) + b\right)$$

Kernel: $K(\mathbf{x}_i, \mathbf{x}) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}),$

$$f(\mathbf{x}) = sign\left(\sum_{i=1}^{N} \alpha_i y_i K(\mathbf{x}_i, \mathbf{x}) + b\right)$$

$$\begin{aligned}\mathbf{z}_i^T \mathbf{z}_j &= z_{1i}z_{1j} + z_{2i}z_{2j} + z_{3i}z_{3j} \\ &= x_{1i}^2 x_{1j}^2 + 2x_{1i}x_{2i}x_{1j}x_{2j} + x_{2i}^2 x_{2j}^2 \\ &= (\mathbf{x}_i^T \mathbf{x}_j)^2,\end{aligned}$$

$$K(\mathbf{x}_i, \mathbf{x}_j) = \Phi(\mathbf{x}_i)^T \Phi(\mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j)^2,$$

Polinomijalni kernel:

$$K(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \mathbf{x}_j + 1)^d, \quad d > 0,$$

# SVM i izbor svojstava: rekurzivna eliminacija svojstava

- Početi sa trening skupom koji uključuje sve promenljive
- Ponavljati dok tekući podskup promenljivih ne postane prazan (ili kriterijum zaustavljanja ispunjen)
  - Izgraditi optimalni klasifikator za tekući podskup promenljivih
  - Za svaku promenljivu u tekućem podskupu oceniti njen multivarijantni značaj
  - Izbaciti svojstvo sa najmanjom vrednošću multivarijantnog značaja

# SVM i izbor svojstava: rekurzivna eliminacija svojstava

- Proces eliminacije se završava ili kada se eliminišu sva svojstva ili kada se dobije predefinisani rezultat
- U oba slučaja imamo sekvencu ugnježdenih podskupova svojstava
- Rezultati imaju oblik rangirane liste gena (svojstava) sa poslednjim eliminisanim genom na vrhu, a prvim eliminisanim – na dnu liste
- Rekurzivna eliminacija svojstava rangira podskupove gena (ne pojedinačne gene)
- Ugnježdeni podskup veličine m (m od p do 1) može se pronaći iz liste uzimajući vršnih m gena (ako je proces okončan eliminacijom svih gena)
- SVM rekurzivna eliminacija svojstava (SVM-RFE) koristi činjenicu da je optimalna razdvajajuća hiperravan pridružena vektoru **w** težina svojstava, $w_k$,

$$\mathbf{w} = \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_p \end{bmatrix},$$

- $k$=1, . . . , $p$ i $p$ je broj svojstava (promenljivih, dimenzija, gena, . . .)

# SVM i izbor svojstava: rekurzivna eliminacija svojstava

- Svojstvo ortogonalno vektoru **w** (sa $w_k$ jednako 0) ne doprinosi klasifikaciji
- Elementi $w_k$ vektora **w** mogu da budu pozitivni ili negativni
- Multivarijantni značaj svojstva *k* može da se definiše kao apsolutna vrednost težine $|w_k|$ (ili, na primer, $w^2_k$)
- Na svakom koraku rekurzivne elminacije svojstava, svojstvo sa najmanjim multivarijantnim značajem $|w_k|$ može da se eliminiše
- Sa više hiljada promenljivih u tipičnim podacima genske ekspresije, treniranje SVM *p* puta je skupo
- Moguće generalizacije – eliminacija više od jednog svojstva po koraku
  - Eliminacjom svih svojstava sa $|w_k|$ ispod nivoa praga
  - Uklanjanjem predefinisanog procenta tekućeg broja svojstava
- Istovremeno uklanjanje većeg broja svojstava zanemaruje neke interakcije između svojstava i može da pogorša rezultate
- Razumni kompromis između tačnosti i cene računanja: uklanjanje grupe svojstava (i do 50%) u prvih nekoliko iteracija a zatim preći na eliminaciju po jedne promenljive

# SVM – glavne karakteristike u otkrivanju biomarkera

- Maksimizovanje margine
  - Optimalna hiperravan optimizuje marginu između klasa i u slučaju tvrde i u slučaju meke margine
- Dualnost
  - Dualna forma optimizacionong problema omogućuje da treniranje (i zatim klasifikacija) zavise samo od skalarnog proizvoda vektora ulaznog prostora
  - Dualni problem se optimizuje u N-dimenzionom prostoru Lagranžovih multiplikatora (N – broj trening tačaka)
  - Za tipične podatke mikronizova genske ekspresije, gde je $N<<p$, značajno smanjenje broja promenljivih u poređenju sa primalnim problemom
- Kerneli
  - Kernel funkcije omogućuju proširenje linearnih SVM na nelinearne slučajeve
  - Koriste originalne vektore ulaznog prostora i vraćaju skalarni proizvod njihovih slika u prostoru svojstava
  - Omogućuju virtuelnu optimizaciju u visoko-dimenzionom (čak i beskonačnom) prostoru svojstava bez eksplicitnog preslikavanja trening podataka u prostor svojstava
  - ***Izbor odgovarajućeg kernela može da bude vremenski zahtevan proces***
- Konveksnost
  - Razdvajajuća hiperravan se pronalazi optimizovanjem konveksne funkcije, nema lokalnih optimuma i rešenje je jednistveno (za dati podskup gena)

# Otkrivanje biomarkera i klasifikacija

- Otkrivanje biomarkera i klasifikacija – pregled
- Otkrivanje biomarkera
- Validacija i izveštavanje
- Izbor svojstava (Feature Selection)
- Diskriminantna analiza i izbor svojstava
- SVM i izbor svojstava
- Nasumične šume i izbor svojstava

# Nasumične šume

- Bootstrap (samopokretanje, inicijacija)
  - Ove metode pripadaju tehnikama ponovnog uzorkovanja – generisanje veštačkih skupova podataka radi boljeg ocenjivanja statističkih svojstava prediktivnog sistema ili populacije
  - Najčešće se podrazumeva Efronov neparametarski bootstrap – uzorci se nasumično biraju iz uzorka sa vraćanjem i iste su veličine kao i originalni uzorak
  - Nepoznati parametri populacije mogu da se ocene uprosečavanjem ocena iz svih bootstrap uzoraka
  - Pristup može da se koristi za ocenu greške pogrešne klasifikacije klasifikacionog sistema
  - Može da se koristi za izgradnju većeg broja klasifikatora koji se zatim koriste za klasifikaciju uzoraka iz originalnog trening skupa
  - Ocena je realnija ako se svaki bootstrap klasifikator koristi za klasifikovanje samo onih uzoraka originalnog skupa koji nisu korišćeni baš u njegovom treniranju

# Nasumične šume

- Pakovanje (bagging)
- Pojačavanje (boosting)
  - Ideja: kombinovanje slabih klasifikatora (tek nešto bolje od nasumičnog pogađanja) u izgradnji moćnog ansambl klasifikatora
  - Polazi se od slabog (baznog) klasifikatora;  pojačavanje gradi niz klasifikatora koji se treniraju na modifikovanom trening skupu
  - Modifikacije mogu da uključe dodelu različitih težina trening uzorcima
  - U svakoj iteraciji naglasak se daje pogrešno klasifikovanim uzorcima u prethodnom koraku
  - Ansambl klasifikatora je sastavljen od svih sekvencijalno izgrađenih klasifikatora
  - Klasifikacija se bazira na težinskom glasanju
  - Najpopularniji boosting algoritam, AdaBoost, 1997 Freund i Shapire

# Gini indeks nečistoće

- U stablima odlučivanja
  - Pored entropije čvora,
  - još jedan kriterijum odluke po kojoj promenljivoj da se cepa unutrašnji čvor,
  - treba definisati meru nečistoće čvorova – dece
- Pretpostavka: *J* klasa, čvor $\alpha$ sa $N_\alpha$ uzoraka
- $\hat{p}_j(\alpha)$ udeo $N_\alpha$ u svakoj od klasa *j*, j=1,2,...,*J*
- Mera nečistoće čvora $\alpha$, $i(\alpha)$, biće 0 ako svih $N_\alpha$ uzoraka pripadaju jednoj istoj klasi
- Mere nečistoće:
  - Entropijska nečistoća
  - Entropija varijanse i njena generalizacija – Gini indeks
  - Nečistoća stope pogrešne klasifilacije

# Mere nečistoće čvora

- Entropijska nečistoća

$$i_{entropy}(\alpha) = -\sum_{j=1}^{J} \hat{p}_j(\alpha) \log_2 \hat{p}_j(\alpha).$$

- Nečistoća stope pogrešne klasifilacije

$$i_{misclassification}(\alpha) = 1 - \max_j \hat{p}_j(\alpha).$$

- Gini indeks nečistoće

$$i_{Gini}(\alpha) = \sum_{j=1}^{J} \sum_{j' \neq j}^{J} \hat{p}_j(\alpha) \hat{p}_{j'}(\alpha).$$

- Entropija varijanse – specijalni slučaj Gini indeksa za $J$=2

$$i_{Gini}(\alpha) = \hat{p}_1(\alpha) \hat{p}_2(\alpha).$$

# Mere nečistoće čvora

- Za cepanje čvora $\tau$ na dva deteta-čvora $\alpha$, $\beta$, bira se promenljiva koja nudi maksimalno smanjenje mere nečistoće
- Za binarno stablo i *J* klasa, smanjenje $\Delta i_{Gini}(\tau)$ za Gini meru nečistoće $i_{Gini}(\tau)$ u čvoru $\tau$, računa se kao

$$\Delta i_{Gini}(\tau) = i_{Gini}(\tau) - i_{Gini}(\alpha)\hat{p}_\alpha - i_{Gini}(\beta)\hat{p}_\beta,$$

- $i_{Gini}(\alpha)$ $i_{Gini}(\beta)$ su Gini indeksi nečistoće čvorova – dece, a

$\hat{p}_\alpha$ $\hat{p}_\beta$ udeli trening uzoraka u čvoru $\tau$ dodeljenih čvorovima $\alpha$, $\beta$

# Nasumične šume

- Predavanje Klasifikacija u bioinformaici (slajdovi 89-92)
  - Karakteristike
  - Algoritam učenja
- Koriste mnogo stabala i mnogo promenjivih za klasifikaciju novog uzorka
  - Nepodesno za biomarkere
  - Neophodan izbor svojstava (Feature Selection)

# Nasumične šume i izbor svojstava

- Koraci u izračunavanju značaja $I_k$ promenljive $k$ u svakom stablu $t$ u šumi ($t$=1,2,..., $n_{tree}$)
  - Klasifikovati OOB uzorke za stablo $t$ i prebrojati glasove za korektnu klasu (broj korektno klasifikovanih OOB uzoraka)
  - Nasumično permutovati vrednosti promenljive $k$ u OOB uzorcima, provući ih kroz stablo i prebrojati glasove za korektnu klasu
  - Oduzeti broj glasova za korektnu klasu za OOB uzorke sa permutovanom promenljivom $k$ od broja glasova za korektnu klasu u originalnim OOB podacima
  - Rezultujuća razlika je značaj $I_k(t)$ promenljive $k$ za stablo $t$
  - Srednja vrednost ove razlike po svim stablima – skor značaja $I_k$ promenljive $k$, $$I_k = \frac{1}{n_{tree}} \sum_{t=1}^{n_{tree}} I_k(t).$$
  - Deljenjem standardnom greškom $\sigma/\sqrt{n_{tree}}$ dobije se standardizovani index
  - Statistička značajnost skora značaja $I_k$ $$z_k = \frac{I_k}{\frac{\sigma}{\sqrt{n_{tree}}}}$$

# Nasumične šume i izbor svojstava

- Značaj promenljive može da se računa i preko smanjenja Gini indeksa nečistoće $\Delta i_{Gini}(\tau, k)$ kada se čvor $\tau$ cepa na bazi promenljive *k*
- Značaj promenljive *k* u stablu *t*, $I_k(t)$, $t$=1,2,..., $n_{tree}$, može da se računa kao suma smanjenja svih indeksa nečistoće u stablu *t* za promenljivu *k*

$$I_k(t) = \sum_{\tau \in t} \Delta i_{Gini}(\tau, k).$$

- Uprosečavanjem značaja promenljive $I_k(t)$ po svim stablima u šumi dobija se značaj promenljive $I_k$

- Često konzistentan sa merama na bazi permutacija

# Nasumične šume, izbor svojstava i biomarkeri

- Koristi se neka od mera značaja promenljive
- Izbor svojstava – iterativna procedura slična rekurzivnoj eliminaciji  svojstava
- U svakoj iteraciji isključuju se najmanje značajne promenljive i gradi se nova šuma na bazi preostalih promenljivih
- Kriterijum eliminacije promenljivih: neki prag značajnosti ili procenat promenljivih
- Od niza šuma, bira se jedna i njen skup promenljivih, na bazi
  - Broja promenljivih
  - Stope greške OOB šume
  - Kombinacije prethodnog
  - Na primer, izbor skupa promenljivih pridruženih stablu koje ima najmanji broj gena među svim stablima sa stopom greške OOB u okviru predefinisaog broja standardnih devijacija od minimalne stope greške svih šuma

# Učenje linearne SVM

- Lagranžova formulacija problema (Lagranžijan) objedinjuje funkciju koja se minimizuje i ograničenja

$$L(\mathbf{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\mathbf{w}^T\mathbf{w} - \sum_{i=1}^{N} \alpha_i [\, y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1]$$

  - Prednost: trening podaci se pojavljuju samo u obliku skalarnog proizvoda, što će omogućiti uopštenje na nelinearno razdvojive podatke
  - Vrednost $\alpha$ - Lagranžov množilac - (tj. vektor ($\alpha_i$)) određuje se iz uslova $\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial \mathbf{w}} = 0$ $\frac{\partial L(\mathbf{w}, b, \boldsymbol{\alpha})}{\partial b} = 0$
  - Zamenom $\alpha$ i $b$ dobije se minimum ciljne funkcije
  - Svi $\alpha_i$ su pozitivni jer su sva ograničenja ≥ 0
- Rešavanjem prethodnih jednačina dobije se $w = \sum_{i=1}^{N} \alpha_i y_i x_i$ i $\sum_{i=1}^{N} \alpha_i y_i = 0$
- Zamenom u funkciju $L$, dobija se dualna reprezentacija Lagranžijana $W(\boldsymbol{\alpha}) = \sum_{i=1}^{N} \alpha_i - \frac{1}{2}\sum_{i,j=1}^{N} \alpha_i \alpha_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$

  pri uslovu $\sum_{i=1}^{N} y_i \alpha_i = 0$ i $\alpha_i \geq 0.$

nazad

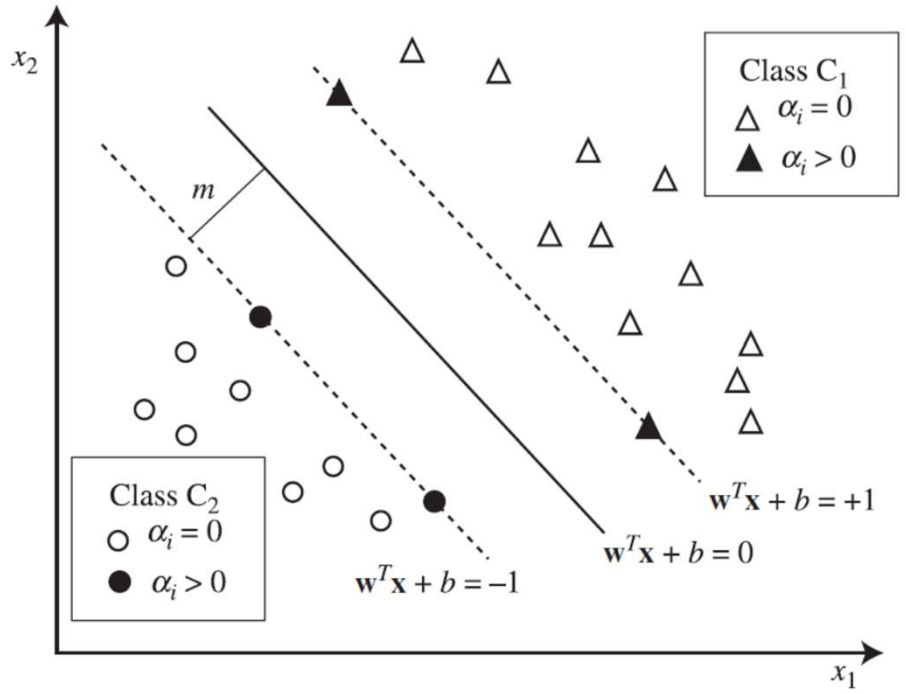
# Dualna reprezentacija

- Rešenja treba da zadovolje tzv. Karush-Kuhn-Tucker (KTT) uslove komplementarnosti:

$$\alpha_i[\, y_i(\mathbf{w}^T\mathbf{x}_i + b) - 1] = 0 \quad \text{for } i = 1, \ldots, N$$

- Lako se vidi da
  - $\alpha_i$ može da bude pozitivno samo kada je $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) = 1$
  - Sve trening tačke za koje je $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) > 1$ moraju da imaju $\alpha_i = 0$
- Trening tačke na hiperravnima $\boldsymbol{w}^T\boldsymbol{x}_i + b = +/-1$ za koje je $\alpha_i > 0$ zovu se ***podržavajući vektori***
- Vektor rešenja w* može da se sračuna kao

$$\mathbf{w}^* = \sum_{i=1}^{N} \alpha_i y_i \mathbf{x}_i$$



# Meka margina razdvajanja

- Slek promenljive (labave promenljive, promenljive „olabavljenja“)
- Formule za tvrdu marginu zamenjuju se formulama sa labavim promenljivim:

$$\min_{w,b} \frac{1}{2}\|w\|^2 + C\sum_i \xi_i$$

$$\textit{subject to: } y_i(\langle w, x_i\rangle + b) \geq 1 - \xi_i, \quad \xi_i \geq 0$$

- Konstanta $C > 0$ zadaje relativni značaj maksimizovanja margine i minimizovanja labavih veličina (greške klasifikacije)
- Ova formulacija se naziva SVM sa mekom marginom

nazad