11. Методи за обучение. Обучение с учител - параметрични и непараметрични подходи. Линеен дискриминант на Фишер. Обучение без учител – minimax, kMeans, Izodata [5, 12].

(Виж материалите страница ПР_Част1 44 – 52, , http://www2.it.lut.fi/kurssit/04-05/010586001/lectures.html - лекции 3,4 и 5)

Методи за обучение

Теорията на Бейс за статистическа класификация дава прави добра класификация само ако вероятностната плътност на класовете е известна.
Обикновено тя не е известна и тогава се налага да се направят оценки на базата на някакви данни и някакво обучение.
Методи за обучение на класификатора се делят на две групи – с учител и без учител.

- С учител
  Ако имаме класове C1,C2,…Cn и сега от тях изберем чрез експерименти Ti – множество от елементи от Ci – тия клас. Като:
  $$T = \bigcup_{i=1}^{n} T_i$$
  T се нарича тестово множество. Освен това както вече споменахме:
  $$T_i \subset C_i \subset W$$
  За всеки пример x от тестовото множество можем да дефинираме 2-ка (x, i), която да указва, че примера x ϵ Ti.
- Самообучение x ϵ T с W и знаем само това на входа.

Интуитивни методи

1. k-те центъра – имаме k на брой еталонни обекта, някаква метрика например Евклидова или Статистическа(Мехаланобис) и представянето на x:
   $$x_{rep} = \frac{1}{|T_i|} \sum_{x_i \epsilon T_i} x_i$$
   Изолиниите на гаусиана означават събития с еднаква вероятност. За разстоянието по Мехаланобис може да е изпълнено $D_M(x,\mu) = D_M(y,\mu)$, докато за евклидовото $D_E(x,\mu) \leq D_E(y,\mu)$. В зависимост от статистиката разстоянието по Мехаланобис разтегля или смачква евклидовото разстояние. Основната идея все пак си остава:
   $$y = Argmin\ D(x, x_{rep})$$
2. NN
   За всяко x и xi, където xi ϵ T намираме $y = Argmin\ D(x, x_i)$.
   Проблем би било, ако имаме такова разположение на класовете, че x да е по-близо до центъра на един клас, но да се намира в друг.

3. kNN (k-Nearest Neighbour)

Отново имаме представители от всички класове, но този път не само по един.

a. Пресмятат се всички налични разстояния
b. Сортират се
c. Избират се k – на брой на близки разстояния

$$y = Argmax\ k_i, \text{където } i = 1,2 \dots, n$$

ki е броят на елемнтите от класа Ti в избраните k най-близки. (Може да се зададе и по-сложна функция от чистия сбор)

Необходими данни:

- n-те точки за центровете(или други)
- обучаващите примери - |T| на брой.

Методи за обучение с учител

При неясна вероятностна плътност, но знание за разпределението. Остава чрез настройване на параметрите и тяхна оценка да достигне до правдобна класификация. При настройване на параметри казваме, че имаме така наречения параметричен подход, защото се опитваме да обучим системата да намери правилните параметри. Нека да имаме да дефинирае модела на разпределение като M(θ_1, θ_2, …θ_n) (за пример може да бъде даден гаусиан $N_n(\mu, \sum, x)$ – където μ и ∑ могат да бъдат настроени според някакви тестови данни).

Един параметричен подход е подхода за максимално правдоподобие.
Какво имаме до момента – p(x|Ci) има нормално разпределение, всеки един клас е описан от параметричен вектор от модела – на Ci – θi и тестово множество T.
Като тестовите примери са независими.
Въвежда ме функция на правдоподобието:

$$p(x|\theta) = p(x_1, x_2, \dots, x_n|\theta) = \prod_{k=1}^{N} p(x_k|\theta)$$

Прилагам стандартната стъпка за подхода на максималното правдоподобие логаритмуване на гаусиана и получаваме:

$$\frac{\sum_{x \epsilon T_i}(x_1, x_2, \dots x_n)}{|x_i|} = (\mu_1, \mu_2, \dots \mu_n)$$

За матрицата ∑ имаме:

$$\Sigma = \left|\left|\sigma_{ij}\right|\right| = \sum_{x \epsilon T_i} (\mu_i - x_i)(\mu_j - x_j)$$

След това използваме клаоическия подход на Бейс, при който приемаме, че p(θ) има фиксирано разпределение.

$\sigma_{ij} = \frac{\Sigma}{|T_i|-1}$ – също състоятелна, но отместена оценка.

Непараметрични
Когато не знаем дори функцията на разпределението и нямаме как да приложим параметричен подход. Тогава прилагаме така наричения непараметричен подход. При него единственото, което знаем са тествоите примери.
Можем да приложим Хистограмен подход – в двумерното пространство разделяме абцисата на равни интервали с дължина h.
Вероятността за всяко една парче от хистограмата е:

$$P \approx \frac{k}{N}$$

Където к е броя на примерите вътре в парчето от хистограмата а N броя на всички примери.
Необходимо допускане е, че функцията на плътността е еднаква в рамките на всеки правоъгълник от хистограмата. Така получаваме:

$$\hat{p}(x) \equiv \hat{p}(\hat{x}) \approx \frac{k}{hN}, |x - \hat{x}| \leq \frac{h}{2}$$

Оценката изглежда добра при малко h и голям брой примери N.

Parzen Windows
Същия подход само, че за многомерно пространство. Тук вместо правоъгълници със страна h, ще използваме хиперкубове със страна h и обем V_l – където l измерението на пространството.
При избор на x построяваме хиперкуба с център x и изчисляваме елементите в него:

n = |Ti| , ki ϵ Ti,

$$p(x) = \frac{k_i}{nV_l}$$

n – броя на всички експерименти
ki – броя на тези, които са вътре в куба
Vn – обема на куба

Ако имаме празни клетки то имаме нужда от още експерименти.

Нека дефинираме функцията $\varphi(x) = \{^{1,\ |ti|\text{е вътре в куба}}_{0,\ \text{иначе}}$, където ti е елемент на Ti.

- ⇨ $ki = \sum_{i=1}^{n} \varphi(\frac{(xi-x)}{h})$, тъй като φ е пръксната това води до прекъснатост на p(x), затова можем да дефинираме φ по-добре:
- ⇨ $\varphi(x) \geq 0$ и $\int \varphi(x)dx = 1$, това е легитимна оценка на функция на вероятностната плътност. Такива функции са известни като kernel функции или Parzen Windows. Типичен пример е Гаусиан – N(0,I).

Избора на h е доста сериозен проблем при ограничен брой тестови данни.
За това може да се използва подход като при k-nearest neighbor:
Вместо да се броят примерите в хиперкубове се избират k на брой най-близки примери и се изчислява обема на получената фигура.
Така в по-гъстите участъци обема ще е по-малък, а в по-редките по-голям.

$$p(x) = \frac{k}{nV(x)}$$

Линеен дискриминант на Фишер

За да се определи един класификатор, първо трябва да изберем подходяща проекционна права, която да разделя класовете по възможно най-добрия начин.
Линейната дискриминатна функция, която отговаря за проектирането може да бъде представена във вида:

$$y = w^T x$$

Където дължината на вектора w е ограниче до ||w||=1. За да измери разграничеността на два класа Фишер въвежда:

$$J(w) = \frac{s_B^2}{s_W^2}$$

Където SB е между класова дисперсия, а SW е вътре класова дисперсия.
Намирайки проекционната права L ние можем да построим построим посредство дискриминантната равнина между класовете.

Извираме за модели гаусиани, които проектираме върху L. Така задачата от n-мерно простратрво се свежда до 1-мерно. Метода е средно между линеен дискриминантен анализ и вероятностен подход.

$$S_B^2 = w_1(x-\mu_1)^2 + w_2(x-\mu_2)^2$$
$$S_w^2 = \sigma_1^2 + \sigma_2^2$$

Критерия за оптималност на Фишер е $J(w)$ да е максимално.
При еднаква априорна вероятност за класовете => w1 = w2.

⇨ $J(w) = \frac{(\mu_1-\mu_2)^2}{\sigma_1^2+\sigma_2^2}$

Като разпишем за n-мерното пространво получаваме:

$$J(w) = \frac{w^T S_B w}{w^T S_W w}$$

Което е максимално при $W = (\mu_1 - \mu_2)\, S_w^{-1}$, а ако са известни ковариационние матрици проблема се свежда до проблем за собствените стойности.

Непараметрични методи за определяне броя на класовете в T

Minimax
Дадено ни е само множеството T и някаква метрика (D(N0,N1) – евклидово разстояние).

1) Избираме произволно ядро – N0 е T (потенциално ядро на клас C0)
2) N1 е T, такова че D(N1, N0) = max (потенциално 2-ро ядро)
3) Разделяме T на два класа {N0} и {N1} по min D.
4) N2 е T, такова че като изберем по един кандидат от C0 и C1, то
   N20 е C0 – кандидат 1 е D(N20,N0) е max.
   N21 е C1 – кандидат 2 е D(N21,N1) е max.
   N2 = min {N21, N20}
5) Разпределяме T в k класа по min D(Ni, x) за i=1,2,…k.
6) За Nk+1 имаме $N_{k+1} = Argmin_{\{N_i\}}(Argmax\ D(N_i, N_{k+1\,i}))$
7) Остава само да кажем кога спираме – когата средното разстояние м/у класовете е под някакъв праг:

$$G_k = \frac{1}{\frac{k(k-1)}{2}} \sum_{j<i<k}^{k} D(N_i^k, N_j^k)$$

   Спираме при Gk+1 < ½ Gk.

Метод на k-те среди
Нека имаме вече k на брой ядра.

1) Правим разпределение в T на k класа Ci={Ni} за i=1,2,...,k по минимално разстояние.
2) За всяко Ci изчисляваме новите центрове $\widetilde{N_i} = \frac{1}{|N_i|}\sum_{x_i \epsilon C_i} x_i$
3) Или спираме или се връщаме на 1 с новите центрове.

IZODATA
Може не само да увеличаваме, но и да намаляваме ядрата.
Имаме k ядра и 2 вида стъпки – за нарастване на ядрата и за съкращаването им.

1) Преразпределяне на T на k класа Ci={Ni} за i=1,2,...,k по минимално разстояние.
2) Стъпка за намаляване на k – обединение на 2 класа.

a. Намираме двата най-близки класа min D(Ni, Nj)
b. Ако D(Ni,Nj) < e – обединяваме иначе не, като e да кажем, че някаква допустима дисперсия

3) Разбиване на класа Ci на два класа – Ci0 и Ci1
Имаме статистиката на класа и ковариантната матрица. Можем да я приведем в диагонална форма и да подредим сигмите по големина:

$$\sigma_{i1}$$

$$\sigma_{i2}$$

$$\sigma_{in}$$

Ако се окаже, че в тази редица има голям скок в разликите м/у сигмита ще е основателно да направим разделението иначе няма смисъл.