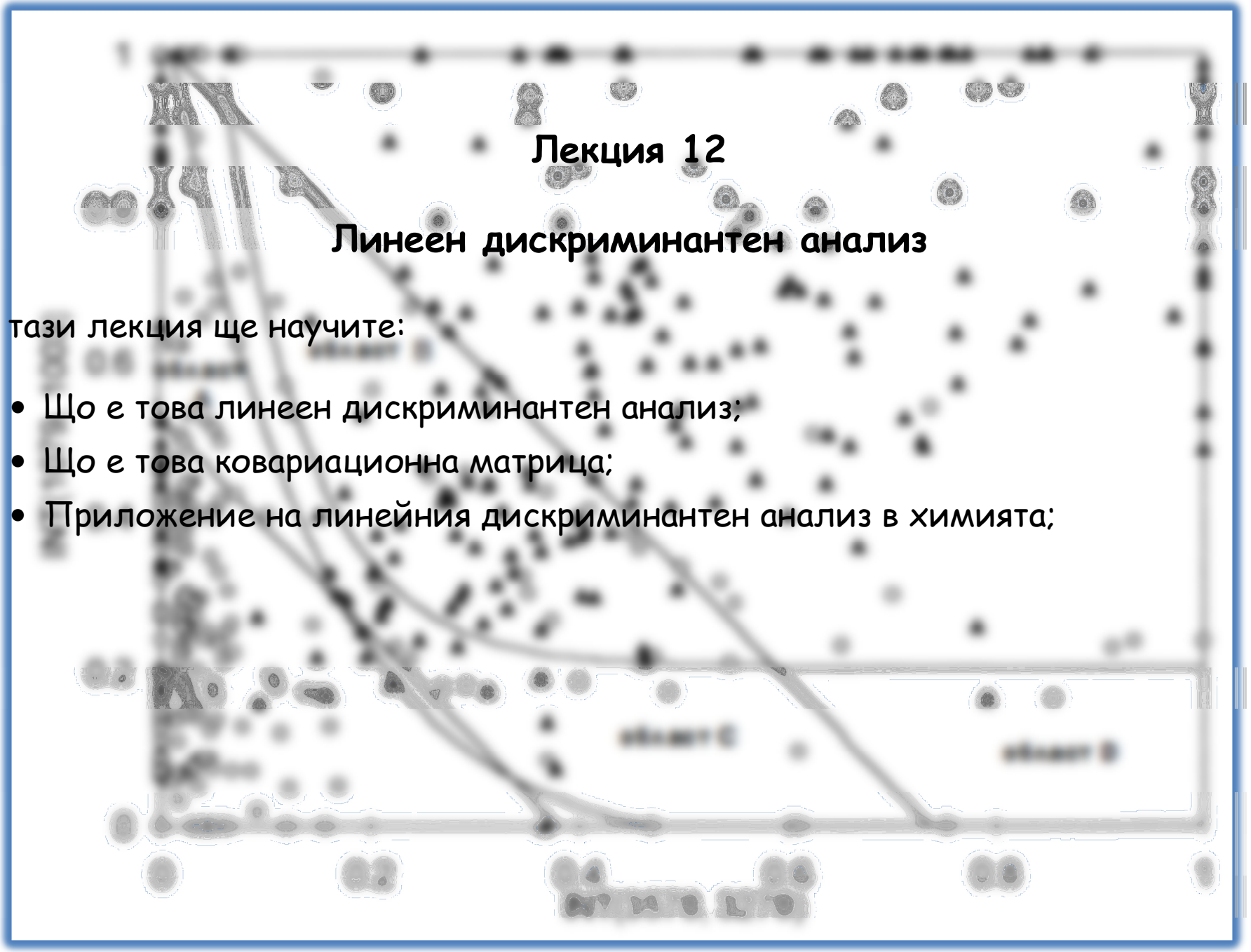


Лекция 12

Линеен дискриминантен анализ

В тази лекция ще научите:

- Що е това линеен дискриминантен анализ;
- Що е това ковариационна матрица;
- Приложение на линейния дискриминантен анализ в химията;

Линейният дискриминантен анализ (ЛДА)

- се използва в статистиката, разпознаването на образи и обучението на машини,

- за да се намери линейна комбинация от признаци;

- получената линейна комбинация е нов признак и той характеризира много по-добре разделянето на два или повече класа.

В тази лекция ще покажем приложението на метода за два класа.

Неговото приложение за повече класове е подобно и ако усвоите този материал, то лесно се разбира.

Алгоритъмът на ЛДА е

- бърз;

- еднозначен;

- ако класовете са линейно разделими, той напълно разделя двата класа;

- ако класовете са линейно неразделими в пространството на признаците, методът осигурява нова променлива, която най-добре разделя класовете, въпреки тяхното припокриване.

ЛДА е подобен на ANOVA (analysis of variance) и регресионния анализ, които се опитват да обяснят една зависима променлива от признаците (променливите), които характеризират извадката от данни.

ЛДА е подобен на анализа на главните компоненти (principal component analysis, PCA), защото подобно на него ЛДА намира линейна комбинация от изходните променливи.

Да припомним някои понятия

- един образ (обект) се характеризира с няколко променливи (признаци);

- образът е многомерен вектор: признаците са координатите на вектора;

- **скаларно произведение** на два вектора,

$$\mathbf{x} = (x_1, x_2, \ldots x_N) \text{ и } \mathbf{w} = (w_1, w_2, \ldots w_N)$$

се получава като се умножат съответните координати на векторите и се съберат

$$\mathbf{x}\ \mathbf{w} = x_1 w_1 + x_2 w_2 + \ldots + x_N w_N = \Sigma\, x_n w_n = \Sigma\, w_n x_n$$

- ако координатите на втория вектор са постоянни коефициенти, то горната сума е **линейна комбинация** от координатите на първия вектор.

Да припомним някои понятия

- **Ковариация** на две случайни величини X и Y се дава с израза

$$\mathbf{cov(X,Y) = M\{[X - M(X)][Y - M(Y)]\}}$$

където с **M()** е означено съответното математическо очакване.

Ако имаме M измервания на двете величини

#	1	2	3	… m …	M
X	x_1	x_2	x_3	x_m	x_M
Y	y_1	y_2	y_3	y_m	y_M

Оценка за ковариацията е изразът

$$\sigma_{x,y} = [\Sigma (x_n - \bar{x})(y_n - \bar{y})] / (M - 1),$$

където се сумира по броя измервания (от 1 до M), а $\bar{x}$ и $\bar{y}$ са съответните средни стойности.

Нека имаме обучаваща извадка на M образа с N признака.

Нека първите M_1 обекти са от първи клас, а вторите M_2 – от втори.

$$M_1 + M_2 = M$$

Това на практика представлява една матрица, **X**, с M реда и с N колони.

Произволен елемент $x_{m,n}$ на тази матрица преставлява стойността на $n^{ти}$ признак на $m^{ти}$ образ.

За обектите от първи клас и за тези можем да изчислим съответните средни образи $\bar{\mathbf{x}}_1$ и $\bar{\mathbf{x}}_2$, както и тяхната разлика $\bar{\Delta} = \bar{\mathbf{x}}_1 - \bar{\mathbf{x}}_2$.

$$\bar{\mathbf{x}}_{1\,p} = (\Sigma\, x_{m1,p}) / M_1 ; \;\; p = 1, 2, \ldots N$$

$$\bar{\mathbf{x}}_{2\,p} = (\Sigma\, x_{m2,p}) / M_2 ; \;\; p = 1, 2, \ldots N$$

$$\bar{\Delta}_p = \bar{\mathbf{x}}_{2\,p} - \bar{\mathbf{x}}_{1\,p} ; \;\; p = 1, 2, \ldots N$$

За образите на двата класа можем да съставим т.н. ковариационни матрици Σ_1 и Σ_2, чийто елементи са равни на

$$\sigma 1_{p,q} = [\Sigma (x_{m1,p} - \bar{x}_p) (x_{m1,q} - \bar{x}_q)] / (M_1 - 1);\ p,q = 1, 2, \ldots N$$

и

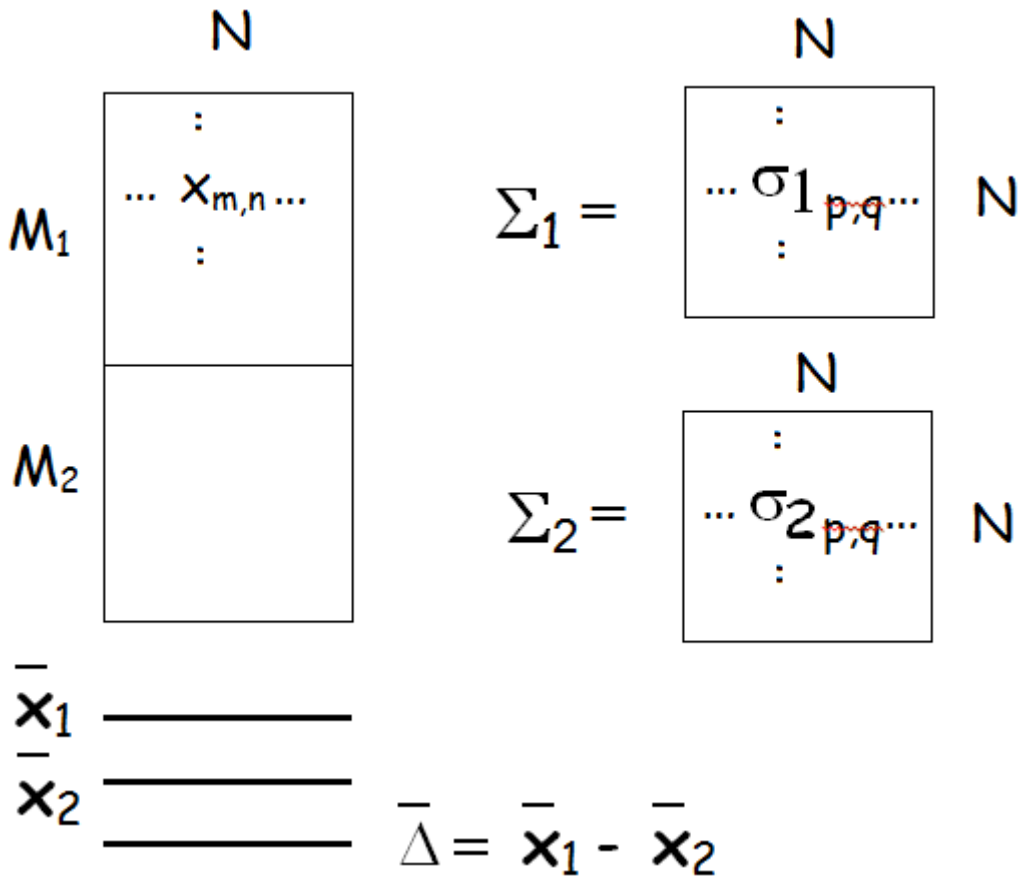
$$\sigma 2_{p,q} = [\Sigma (x_{m2,p} - \bar{x}_p) (x_{m2,q} - \bar{x}_q)] / (M_2 - 1);\ p,q = 1, 2, \ldots N$$

В първата сума се сумира по m_1 от 1 до M_1, а във втората по m_2 от 1 до M_2.

Обърнете внимание, че ковариацията се изчислява със съответните колони!

И двете ковариационни матрици са симетрични квадратни матрици с размерност, равна на броя признаци (променливи) N.

Тези изчисления графично могат да се онгаледят по следния начин



Трите вектора, два за средните образи и един за тяхната разлика, представляват матрици с размерност 1 на N=

Новата координата се дава като линейна комбинация от координатите на образите в обучаващата извадка:

$$d = \mathbf{x}\,\mathbf{w} = x_1 w_1 + x_2 w_2 + \ldots + x_N w_N = \Sigma\, x_n w_n = \Sigma\, w_n x_n$$

Обърнете внимание, че d е число, а не вектор!

Как се изчислява векторът **w**?

Той има такива координати w_n, че следният израз е максимален за обучаващата извадка

$$F = (\,\overline{d}_2 - \overline{d}_1\,)^2/(s_1^2 - s_2^2),$$

където $\overline{d}_2$ и $\overline{d}_1$ са средните стойности на променливата d за обектите от клас 2 и клас 1 в обучаващата извадка, а s_1^2 и s_2^2 – съответните стандартни отклонения на променливата d за тези класове.

Математиците са доказали, че векторът **w** се получава по следната формула

$$\mathbf{w} = \overline{\Delta} \, (\Sigma_1^{-1} \text{ и } \Sigma_2^{-1})$$

където Σ_1^{-1} и Σ_2^{-1} са обратните матрици на съответните ковариационни матрици Σ_1 и Σ_2.

Този вектор осигурява най-голямо отношение F за обектите на обучаващата матрица.

Тъй като в числителя на F имаме квадрат, то и векторът (–**w**) дава същото F, т.е. същото най-голямо разделяне на двата класа!

Тъй като отношението F е безразмерно, то и векторът (a**w**), където a е произволно число, дава същото F, т.е. същото най-голямо разделяне на двата класа!

Затова векторът **w** се нормира да има дължина единица:

$$\mathbf{w}' = \mathbf{w} \, / \, |\mathbf{w}|$$

По новата променлива d може да се избере праг (threshold), d_{thr}, за който да имаме

Ако **x** е от 1 клас $d = \mathbf{x}\,\mathbf{w} = x_1 w_1 + x_2 w_2 + \ldots + x_N w_N = \Sigma\, w_n x_n < d_{thr}$

Ако **x** е от 2 клас $d = \mathbf{x}\,\mathbf{w} = x_1 w_1 + x_2 w_2 + \ldots + x_N w_N = \Sigma\, w_n x_n > d_{thr}$

Разбира се, ако данните са линейно разделими, горните две отношения могат да са изпълнени за всички обекти от обучаващата извадка.

Има случаи на линейно разделими данни, които не се разделят с ЛДА!