Intelligens adatelemzés ea. vázlat 1. rész

A tematika

1.ea. a tárgy tematikájának áttekintése. Egy mintapélda M-S adatok elemzése (PA).
2.ea. HF-ok jellegének megbeszélése, a HF témák választásához szempontok
3.ea. Statisztikai próbák
4.ea. Statisztikai próbák (folyt)
5. ea. Statisztikai próbák (folyt)
6.ea. Lineáris regressziós eljárások
7.ea. Bayes lineáris regressziós módszerek
8.ea. Osztályozás
9.ea. Lineáris osztályozási eljárások
10.ea. Lineáris osztályozási eljárások folytatás
11.ea. Kernel módszerek
12.ea. Kernel módszerek (folyt), SVM.

A tárgy tematikájának kialakításánál feltételeztük, hogy a klasszikus adatelemző módszereket és a tanuló rendszerek alapjait mindenki ismeri. (Háttéranyagként szolgálhat pl. az Altrichter…: Neurális hálózatok c. könyv)

1. Statisztikai próbák

A statisztikai próbák célja hogy egy vagy több statisztikai sokaság, mintakészlet eloszlásával kapcsolatban hipotéziseket állítson fel és ezen hipotéziseket ellenőrizze. A hipotézisek általában vonatkozhatnak a sokaság eloszlására, ismert eloszlás feltételezésével az eloszlás ismeretlen paramétereire, két vagy több sokaság eloszlásnak hasonlóságára (az eloszlások paramétereinek egyezésére vagy különbözőségére), két vagy több mintakészlet függetlenségére, stb. A tárgy keretében csak az alapfogalmakat és néhány egyszerű statisztikai próbát tárgyalunk. Ezek és további eljárások a bőséges irodalomban megtalálhatók: pl. [1], [2], [3]. A statisztikai próbák gyakorlati alkalmazását jól segítik a különböző statisztikai programcsomagok. Ezek közül néhány fontosabb: SPSS [4], Statistica [5], [6], WEKA [7], MATLAB [8], RapidMiner [9], KNIME [10] és újabban az igen intenzíven fejlődő R [11].

A statisztikai próbák során a mintakészletből ún. próbastatisztikát csinálunk, melynek alapján eldönthető, hogy a fenti típusú kérdésekre milyen válasz adható. A próbastatisztika eloszlása ismert – megfelelő feltevések teljesülése esetén – így az eloszlás alapján meg lehet határozni, hogy a feltevés (hipotézis) milyen valószínűséggel fogadható el, vagy el kell-e vetnünk.

A statisztikai próbák során az eljárás:

- felállítunk egy nullhipotézist és azt ellenőrizzük, hogy ez a hipotézis megfelelő megbízhatósággal teljesül-e.

- ehhez ismernünk kell vagy meg kell határoznunk a próbastatisztika eloszlását (sűrűségfüggvényét) és ennek alapján meghatározható, hogy mi annak a valószínűsége, hogy a próbastatisztika alátámasztja a hipotézisünket. A sűrűségfüggvény alapján megadható az ún. elfogadási tartomány.

- az elfogadási tartományba esés valószínűsége 1-p, tehát annak valószínűsége, hogy H_0 hipotézist elfogadjuk 1-p. A próbát jellemzi p értéke, mely általában kicsi. Tipikus értékek: p=0,05; 0,01; 0,001.

Statisztikai hipotézisek:

paraméteres (ismert az eloszlás)
nemparaméteres (az eloszlás nem ismert)

További csoportosítása a statisztikai próbáknak:

- illeszkedésvizsgálat

  itt a kérdés, hogy a mintakészlet eloszlása a H_0 hipotézisnek megfelelő-e
  paraméteres illeszkedésvizsgálatok: u, t, F próbák
  nemparaméteres illeszkedésvizsgálat χ^2 teszt

- függetlenségvizsgálat
  itt két vagy több mintakészlet függetlenségét vizsgáljuk. A legfontosabb próba a χ^2 teszt

- homogenitásvizsgálat
  itt két vagy több mintakészlet eloszlásának azonosságát ellenőrizzük.
  Fontosabb homogenitáspróbák:Wilcoxon-próba, Kolmogorov-Szmirnov-próba, …

A hipotézisteszteknél fontos fogalmak még az *elsőfajú* és *másodfajú* hibák, melyek a próba alapján hozott hibás döntések lehetőségeit veszik számba.

Egy hipotézistesztelésnél a döntések az alábbiak lehetnek:

	Ha a H_0 hipotézist	
	elfogadjuk	elutasítjuk
H_0 fennáll	helyes döntés	elsőfajú hiba
H_0 nem áll fenn	másodfajú hiba	helyes döntés

1.1 Paraméteres illeszkedésvizsgálatok

Feltevés: ismert a mintakészlet eloszlása, de az eloszlás paramétere(i) nem ismert(ek).

1.1.1 Az u-próba

Az egymintás u-próba: normális eloszlású független mintáink vannak, ahol ismert a szórás σ_0>0, de nem ismert a várható érték, μ.

- A H_0 hipotézis: $\mu=\mu_0$, vagyis a mintakészletünk eloszlása H_0: $\mathcal{N}\left(\mu_0, \sigma_0^2\right)$
- A próbastatisztika az n-elemű mintakészlet átlaga. Ekkor $\bar{x}_n = \mathcal{N}\left(\mu_0, \frac{\sigma_0^2}{n}\right)$
- Standardizálás után $u(x_1, x_2, \ldots, x_n) = \frac{\bar{x}_n - \mu_0}{\sigma_0}\sqrt{n} \in \mathcal{N}(0,1)$

A teszt elvégzéséhez két küszöbértéket kell meghatározni, de a szimmetria miatt elegendő egy is u_p. Így a standard normális eloszlás eloszlásfüggvénye alapján:

$$\Phi(u_p) = 1 - \frac{p}{2}$$

Az elfogadási tartomány határai: $K_1(p) = -u_p$; $K_2(p) = u_p$; vagyis ha $|u| < u_p$ akkor p szignifikancia-szinten elfogadjuk a H_0 hipotézist.

A kétmintás *u*-próba: adott két független Gauss mintakészlet $x_1, x_2, ..., x_n$ és $y_1, y_2, ..., y_m$ ahol ismert $\sigma_1 > 0$ és $\sigma_2 > 0$, de a két várhatóérték μ_1 és μ_2 ismeretlen. A H_0 hipotézis szerint $\mu_1 = \mu_2$
A próbastatisztika a két mintaátlag különbsége. Ha a minták Gauss eloszlásúak, akkor az átlaguk is az és az átlagok különbsége is az. A próbastatisztika standardizálás után:

$$\frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \in \mathcal{N}(0,1)$$

tehát, ha

$$\left| \frac{\bar{x}_n - \bar{y}_m}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}} \right| < u_p$$

akkor 1-p szignifikanciaszinten elfogadjuk a H_0 hipotézist, vagyis, hogy a két várhatóérték megegyezik.

1.1.2 A t-próba

Egymintás *t*-próba
A t-próba annyiban tér el az u-próbáktól, hogy itt a szórást sem ismerjük. Egyébként most is Gauss eloszlású valószínűségi változó értékeiből áll a mintakészlet.
Adott $x_1, x_2, ..., x_n$ Gauss eloszlású mintakészlet ismeretlen $\sigma > 0$ szórással és μ várható értékkel.
A H_0 hipotézis: $\mu = \mu_0$, vagyis a mintakészletünk eloszlása H_0: $\mathcal{N}(\mu_0, \sigma^2)$

A próbastisztika-jelölt a standardizált átlag lehetne: $u(x_1, x_2, ..., x_n) = \frac{\bar{x}_n - \mu_0}{\sigma}\sqrt{n} \in \mathcal{N}(0,1)$

de a szórás ismeretének hiányában mégsem lehet. Helyette a standardizálásnál a szórás becslését használjuk, ahol a szórás becslésére a korrigált tapasztalati szórást alkalmazzuk:

$s^{*2} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Így a próbastatisztika $t(x_1, x_2, ..., x_n) = \frac{\bar{x}_n - \mu_0}{s^*}\sqrt{n}$

A szórás becslése miatt ez a próbastatisztika már nem Gauss, hanem (n-1) szabadságfokú Student eloszlású valószínűségi változó.
Az elfogadási tartomány határait ennek az eloszlásnak a táblázatai alapján határozhatjuk meg. A p szignifikancia-szinthez tartozó t_p küszöbérték alapján. Ha $|t| < t_p$, vagyis, ha $P(|t| < t_p) = 1-p$ akkor p szignifikancia-szinten fogadjuk el a H_0 hipotézist, egyébként vessük el.

Kétmintás t-próba

Van két párosított mintakészletünk: $x_1, x_2, ..., x_n$ és $y_1, y_2, ..., y_n$ párosítva: $(x_1, y_1)(x_2, y_2)\ldots(x_n, y_n)$
$\sigma_1>0$ és $\sigma_2>0$ ismeretlenek és μ_1 és μ_2 is ismeretlen.
A H_0 hipotézis szerint $\mu_1= \mu_2$
Próbastatisztika lehetne megint a $z_i=x_i-y_i$ átlagának standardizált változata:

$$\bar{z}_n \in \mathcal{N}\left(\mu_1-\mu_2; \sqrt{\frac{\sigma_1^2+\sigma_2^2}{n}}\right); \quad \bar{z}_n \in \mathcal{N}\left(0; \sqrt{\frac{\sigma_1^2+\sigma_2^2}{n}}\right)$$

és standardizálás után

$$\frac{\bar{z}_n}{\sqrt{\frac{\sigma_1^2+\sigma_2^2}{n}}} \in \mathcal{N}(0;1)$$

Mivel a szórásokat most sem ismerjük helyettük az empirikus szórásokkal lehet standardizálni:

Ekkor a

F-próba

ANOVA

2. Regressziós eljárások

Regressziós feladat: mintakészlet egyes csoportjai $\{\mathbf{x}_i\}$ és $\{d_i\}$ (i=1,2,...,*P*) közötti kapcsolat $y_i=y(\mathbf{x}_i)=f(\mathbf{x}_i)$ függvénykapcsolat meghatározása vagy közelítése, ahol általában a d_i értékek az y_i értékek zajos megfigyelései: $d_i = d(\mathbf{x}_i) = y(\mathbf{x}_i)+n_i$.

A regressziós feladat *rosszul definiált* feladat. A véges számú mintapontra végtelen különböző függvény illeszthető. Valamilyen irányú elfogultságra (bias) van szükség, hogy a lehetőségek közül bizonyos típusú megoldásokat preferáljunk.

Modell választás (struktúra, biased modell), modell paraméterek meghatározása, modell-"hangolás" valamilyen kritérium, hibafüggvény alapján.

Hibafüggvény (veszteségfüggvény, loss function), eredő hiba, kockázat, risk. A veszteségnek a teljes mintakészletre vett várható értéke: meghatározásához szükség lenne a mintapontok sűrűség- v. eloszlásfüggvényére.

2.1 Lineáris regresszió

Feltételezzük, hogy $\mathbf{x}_i$-k és y_i-k között lineáris a kapcsolat: $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x}+b$. Általánosított lineáris kapcsolat esetén előbb egy nemlineáris leképezés $\mathbf{x}_i \rightarrow \boldsymbol{\varphi}(\mathbf{x}_i)$, majd a lineáris kapcsolat $y(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x})+b$ jön. A feladat a $\mathbf{w}$ paramétervektor meghatározása (becslése).

Megoldási lehetőségek:

- least squares (LS) becslés: csak a mintapontok állnak rendelkezésre, a veszteségfüggvény a négyzetes hiba

- az egyes megfigyelések hibái eltérő súlyozással is figyelembe vehetők: súlyozott LS becslés
- a megfigyelési zaj valószínűségi jellemzése ismert: likelihood függvény felírható: maximum likelihood (ML) becslés
- a keresett paramétervektor is valószínűségi változó, melynek prior eloszlása (sűrűségfüggvénye) ismert: Bayes becslés.

LS-becslés: Négyzetes hiba alkalmazásakor az LS megoldás a $C(\mathbf{w}) = \frac{1}{2}\boldsymbol{\varepsilon}^T(\mathbf{w})\,\boldsymbol{\varepsilon}(\mathbf{w}) = \frac{1}{2}(\mathbf{d}-\mathbf{X}\mathbf{w})^T(\mathbf{d}-\mathbf{X}\mathbf{w})$ hibafüggvény minimumát biztosítja, ahol **X** a bemeneti $\mathbf{x}_i$ vektorokból, mint sorvektorokból képezett bemeneti mátrix és **d** a d_i megfigyelésekből képezett oszlopvektor (i=1,2,...P). A cél a $\mathbf{w}_{LS}$ meghatározása. Elvégezve a szélsőérték-keresést, a megoldás $\mathbf{w}_{LS}^* = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d}$.

Regularizált LS becslés: A hibafüggvény a regularizációs taggal bővül (járulékos feltétel belefogalmazása): $C(\mathbf{w}) = \frac{1}{2}\boldsymbol{\varepsilon}^T(\mathbf{w})\,\boldsymbol{\varepsilon}(\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w}$, ahol λ a regularizációs együttható, vagy a numerikus instabilitás elkerülése érdekében. A megoldás: $\mathbf{w}_{LS}^* = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{d}$

Súlyozott LS becslés: a $C(\mathbf{w}) = \frac{1}{2}\boldsymbol{\varepsilon}^T(\mathbf{w})\,\boldsymbol{\varepsilon}(\mathbf{w}) = \frac{1}{2}(\mathbf{d}-\mathbf{X}\mathbf{w})^T\mathbf{Q}(\mathbf{d}-\mathbf{X}\mathbf{w})$ hibafüggvény minimumát biztosítja, ahol **Q** a súlyozó mátrix. A megoldás: $\mathbf{w}_{QLS}^* = (\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}\mathbf{d}$.

Az általánosított megoldásnál **X** helyére mindenhol **Φ** kerül.

Regularizált súlyozott LS becslés: $\mathbf{w}_{QLS}^* = (\mathbf{X}^T\mathbf{Q}\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{Q}\mathbf{d}$

Maximum likelihood becslés:

- A megfigyelési zajról feltesszük: $\mathcal{N}\left(0,\sigma^2\right)$
- A zajos megfigyelés $d = d(\mathbf{x}) = y(\mathbf{x}) + n = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x}) + n$. szintén Gauss eloszlású: $\mathcal{N}\left(y(\mathbf{x},\mathbf{w}),\sigma^2\right)$
- P elemű megfigyeléskészlet esetén, a megfigyelések feltételes sűrűségfüggvénye, ha a minták függetlenek: $p\left(\mathbf{d}\middle|\mathbf{X},\mathbf{w},\sigma\right) = \prod_{i=1}^{P}\mathcal{N}\left(y_i(\mathbf{x}_i,\mathbf{w}),\sigma^2\right)$, amit likelihood függvénynek is neveznek. Szokásos jelölés még $\beta = \frac{1}{\sigma^2}$, ezzel a likelihood függvény

  $$p\left(\mathbf{d}\middle|\mathbf{X},\mathbf{w},\sigma\right) = \prod_{i=1}^{P}p(d_i\,|\,\mathbf{x}_i\mathbf{w},\beta^{-1}) = \prod_{i=1}^{P}\mathcal{N}\left(y_i(\mathbf{x}_i,\mathbf{w}),\beta^{-1}\right)$$

- A maximum likelihood becslés a likelihood függvény maximumához tartozó paraméter. A likelihood függvény helyett annak logaritmusára végezzük el a szélsőérték-keresést:
  A log-likelihood függvény:

$$\mathcal{L} = \ln p\left(\mathbf{d}\middle|\mathbf{x},\mathbf{w},\beta^{-1}\right) = \ln \frac{1}{\left(2\pi\sigma^2\right)} \prod_{i=1}^{P} \exp\left(-\frac{\left(d_i - \mathbf{w}^T\boldsymbol{\varphi}(x_i)\right)^2}{2\sigma^2}\right)$$

$$= -\frac{P}{2}\ln(2\pi) - \frac{P}{2}\ln\sigma^2 - \sum_{i=1}^{P}\frac{\left(d_i - \mathbf{w}^T\boldsymbol{\varphi}(x_i)\right)^2}{2\sigma^2}$$

$$= -\frac{P}{2}\ln(2\pi) + \frac{P}{2}\ln\beta - \frac{\beta}{2}\sum_{i=1}^{P}\left(d_i - \mathbf{w}^T\boldsymbol{\varphi}(x_i)\right)^2$$

A szélsőérték-keresés Gauss megfigyelési zaj esetén az LS becslés eredményével azonos eredményre vezet.

$$\mathbf{w}_{ML}^{*} = (\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{d}$$

Az ML becslésnél a megfigyelési zaj szórására is adható becslés

$$\frac{1}{\beta_{ML}} = \sigma_{ML}^2 = \frac{1}{P}\sum_{i=1}^{P}\left(d_i - \mathbf{w}_{ML}^T\boldsymbol{\varphi}(x_i)\right)^2 .$$

Maximum likelihood becslés korrelált zajminták mellett:

Ha a zajminták korreláltak, akkor kicsit módosul az ML becslés.

A feltételek: $E\{\mathbf{n}\}=\mathbf{0}$, $\mathrm{cov}[\mathbf{n}]=\boldsymbol{\Sigma}_{nn}$ és a megfigyelések most is $\mathbf{d}=\boldsymbol{\Phi}\mathbf{w}+\mathbf{n}$

A megfigyelési Gauss zaj sűrűségfüggvénye:

$$p(\mathbf{n}) = \frac{1}{(2\pi)^{P/2}\left|\boldsymbol{\Sigma}_{nn}\right|^{1/2}}\exp\left(-\frac{1}{2}\mathbf{n}^T\boldsymbol{\Sigma}_{nn}{}^{-1}\mathbf{n}\right)$$

A megfigyelések feltételes sűrűségfüggvénye, a likelihood függvény:

$$p\left(\mathbf{d}\middle|\boldsymbol{\Phi},\mathbf{w},\boldsymbol{\Sigma}_{nn}\right) = \frac{1}{(2\pi)^{P/2}\left|\boldsymbol{\Sigma}_{nn}\right|^{1/2}}\exp\left(-\frac{1}{2}\left(\mathbf{d}-\boldsymbol{\Phi}\mathbf{w}\right)^T\boldsymbol{\Sigma}_{nn}{}^{-1}\left(\mathbf{d}-\boldsymbol{\Phi}\mathbf{w}\right)\right)$$

A log-likelihood függvény és ennek alapján az ML becslő:

$$\mathcal{L} = \ln p\left(\mathbf{d}\middle|\boldsymbol{\Phi},\mathbf{w},\boldsymbol{\Sigma}_{nn}\right) = \ldots.$$

$$\mathbf{w}_{ML}^{*} = (\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_{nn}^{-1}\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_{nn}^{-1}\mathbf{d}$$

amit szokás Gauss-Markov (GM) becslőnek is hívni. Láthatóan a GM becslő súlyozott LS becslő, ha a súlymátrix a zaj kovarianciamátrixának inverze.

Az ML és GM becslők (mivel a zaj valószínűségi változó) maguk is valószínűségi változók, meghatározható a várható értékük és a varianciájuk (kovariancia mátrixuk).

$$E\left\{\mathbf{w}_{GM}\right\} = E\left\{(\boldsymbol{\Phi}^T\boldsymbol{\Phi})^{-1}\boldsymbol{\Phi}^T\mathbf{d}\right\} = \mathbf{w}_0 \quad \text{ha} \quad \mathbf{d} = \boldsymbol{\Phi}\mathbf{w}_0 + \mathbf{n}$$

és

$$\mathrm{var}\left[\mathbf{w}_{ML}^{*}\right] = \left[\beta\boldsymbol{\Phi}^T\boldsymbol{\Phi}\right]^{-1}$$

és

$$\mathrm{var}\left[\mathbf{w}_{GM}^{*}\right] = \left[\boldsymbol{\Phi}^T\boldsymbol{\Sigma}_{nn}^{-1}\boldsymbol{\Phi}\right]^{-1}$$

3. Osztályozás lineáris modellek alapján.

A feladat mintapontok két osztályba sorolása azzal a megkötéssel, hogy az elválasztó felület lineáris. A feladat általánosabb megfogalmazásakor az elválasztó felületet nem a bemeneti $\mathbf{x}$ térben, hanem a jellemzőtérben értelmezzük, így a lineáris elválasztófelület a $\boldsymbol{\varphi}(\mathbf{x})$ térben értendő.

Az osztályozó modell ennek megfelelően az $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x} + b$ illetve az $y(\mathbf{x}) = \sum_{i=0}^{M} w_i\varphi_i(\mathbf{x}) = \mathbf{w}^T\boldsymbol{\varphi}(\mathbf{x})$ leképezéssel írható le. Ez utóbbi esetben a bias értéket w_0-ként értelmezzük, úgy hogy $\boldsymbol{\varphi}(\mathbf{x})$ $(M+1)$-dimenziós, és $\varphi_0(\mathbf{x}) \equiv 1$. Az osztályozó konstrukciója most is az $\{\mathbf{x}_i, d_i\}_{i=1}^{P}$ tanítópont-készlet alapján történik.

A lineáris osztályozók között a következő megközelítéseket kell megemlíteni.

A perceptron,
mely a perceptron tanuló eljárással lineárisan szeparálható minták egy lehetséges elválasztó felületének véges tanító lépésben való megtalálását biztosítja. Részletesen ld. az NNkönv 3. fejezetében.

Legkisebb négyzetes hibájú megoldás.
Ez olyan lineáris osztályozót jelent, melynél az osztályozó tényleges válaszai és a kívánt válaszok közötti négyzetes eltérés összegének minimumát biztosító megoldást keressük. Négyzetes hiba alkalmazásakor az LS megoldás a $C(\mathbf{w}) = \frac{1}{2}\boldsymbol{\varepsilon}^T(\mathbf{w})\,\boldsymbol{\varepsilon}(\mathbf{w}) = \frac{1}{2}(\mathbf{d}-\mathbf{X}\mathbf{w})^T(\mathbf{d}-\mathbf{X}\mathbf{w})$

hibafüggvény minimumát biztosítja. A cél tehát a $\mathbf{w}_{LS}$ meghatározása. Elvégezve a szélsőérték-keresést, a megoldás $\mathbf{w}_{LS}^{*} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{d}$. Ez a megoldás formailag megegyezik a lineáris regressziós probléma megoldásával azzal a különbséggel, hogy most $\mathbf{d} = [d_1, d_2, ..., d_P]^T$ olyan megfigyelésvektor, melynek elemei 0 vagy 1 értéket vehetnek fel $d_i \in \{0,1\}$.

A négyzetes hibafüggvény alkalmazása a kilógó adatok hatását felnagyítja, ezért az *LS* osztályozó a kilógó adatokra (outliers) nagyon érzékeny.

Fisher diszkrimináns
A Fisher diszkrimináns a többdimenziós adatok olyan egydimenziós vetületét keresi, mely vetület mentén a szeparálás a lehető legkönnyebben megtehető. A vetület $y(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$, a vetítés pedig a megfelelő $\mathbf{w}$ irányba történik. Ha a két osztályba tartozó mintapontok átlaga $\mathbf{m}_1$ illetve $\mathbf{m}_2$, akkor a legjobb szeparálást úgy biztosítjuk, ha a mintapontok osztályok közötti és osztályon belüli átlagos négyzetes eltéréseinek aránya a lehető legnagyobb.

Jelölje ezt a mennyiséget $J(\mathbf{w}) = = \frac{\mathbf{w}^T \mathbf{S}_b \mathbf{w}}{\mathbf{w}^T \mathbf{S}_w \mathbf{w}}$, melyben $\mathbf{S}_b$ a két osztály között négyzetes eltérést jellemzi (between classes), míg $\mathbf{S}_w$ a mintapontok osztályon belüli szóródására jellemző (within classes). A szélsőérték-keresés eredménye $\mathbf{w} \propto \mathbf{S}_w^{-1}(\mathbf{m}_2 - \mathbf{m}_1)$.

Valószínűségi értelmezés

Az osztályozás valószínűségi értelmezése (Bayes megközelítés) során a kiindulás feltételezi az egyes osztályok előfordulását jellemző a priori valószínűségek ismeretét. Egy kétosztályos osztályozási problémánál a két osztály legyen C_1 és C_2. Az a priori valószínűségek ennek megfelelően $p(C_1)$ és $p(C_2) = 1 - p(C_1)$.

A megfigyelések felhasználását követően az egyes osztályok a posteriori valószínűségei a Bayes-tétel segítségével felírhatók:

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)\,p(C_1)}{p(\mathbf{x}|C_1)\,p(C_1) + p(\mathbf{x}|C_2)\,p(C_2)} = \frac{1}{1+\exp(-a)}, \text{ ahol } \quad a = \ln \frac{p(\mathbf{x}|C_1)\,p(C_1)}{p(\mathbf{x}|C_2)\,p(C_2)}.$$

A posterior tehát az a kifejezés logisztikus szigmoid függvénye.

Az osztályozási feladat megoldása ebben a megközelítésben a feltételes sűrűségfüggvények ismeretét igényli. Feltételezve, hogy $p(\mathbf{x}|C_1)$ és $p(\mathbf{x}|C_1)$ azonos $\mathbf{\Sigma}$ kovariancia-mátrixú és $\boldsymbol{\mu}_1$, illetve $\boldsymbol{\mu}_2$ várható értékű Gauss eloszlás, a sűrűségfüggvény alakja a következő (feltételezve, hogy a bemenet *N*-dimenziós):

$$p(\mathbf{x}|C_i) = \frac{1}{(2\pi)^{N/2}\,|\mathbf{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_i)^T\,\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_i)\right) \quad i=1,2$$

ezzel felírva a -t, ami az együttes valószínűségek hányadosának logaritmusa

$$a = \ln p(\mathbf{x}|C_1) - \ln p(\mathbf{x}|C_2) + \ln \frac{p(C_1)}{p(C_2)}$$
$$= -\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_1)^T\,\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) + \frac{1}{2}(\mathbf{x}-\boldsymbol{\mu}_2)^T\,\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu}_2) + \ln \frac{p(C_1)}{p(C_2)}$$

elvégezve a műveleteket látható, hogy az $\mathbf{x}$ -ben másodfokú tag a közös $\mathbf{\Sigma}$ miatt kiesik, így $\mathbf{x}$ -ben lineáris összefüggést kapunk.

$$a = \mathbf{w}^T\mathbf{x} + w_0, \text{ ahol } \mathbf{w} = \mathbf{\Sigma}^{-1}(\boldsymbol{\mu}_1 - \boldsymbol{\mu}_2) \text{ és } w_0 = -\frac{1}{2}\boldsymbol{\mu}_1{}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_1 + \frac{1}{2}\boldsymbol{\mu}_2{}^T\mathbf{\Sigma}^{-1}\boldsymbol{\mu}_2 + \ln\frac{p(C_1)}{p(C_2)}$$

A megoldás tehát a Gauss sűrűségfüggvények paraméterei $\boldsymbol{\mu}_1$ $\boldsymbol{\mu}_2$ és $\mathbf{\Sigma}$, valamint a priorok $p(C_1)$ és $p(C_2)$ becslése útján nyerhető. Ez a közvetett vagy ***indirekt*** módszer.

Ezek meghatározása ML eljárással lehetséges, ha felírjuk a likelihood függvényt.

A megfigyelések feltételes sűrűségfüggvénye, ha a feltétel a paraméterek értéke (*P* megfigyelés alapján):

$$p(\mathbf{X}, \mathbf{d}|p(C_1), \boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \mathbf{\Sigma}) = \prod_{i=1}^{P} \left\{p(C_1)\,N(\mathbf{x}_i|\boldsymbol{\mu}_1, \mathbf{\Sigma})\right\}^{d_i} \left\{(1-p(C_1))\,N(\mathbf{x}_i|\boldsymbol{\mu}_2, \mathbf{\Sigma})\right\}^{1-d_i}$$

A likelihood függvény negatív logaritmusát képezve a szélsőérték-kereső problémát külön fogalmazhatjuk meg $p(C_1)$-re és a Gauss eloszlás paramétereire. Ebből az ML-becslés:

$$p(C_1) = \frac{1}{P}\sum_{i=1}^{P} d_i = \frac{P_1}{P},$$

$$\boldsymbol{\mu}_1 = \frac{1}{P_1}\sum_{i=1}^{P} d_i \mathbf{x}_i,$$

hasonlóan

$$\boldsymbol{\mu}_2 = \frac{1}{P_2}\sum_{i=1}^{P} (1-d_i)\mathbf{x}_i,$$

míg a kovarianciamátrix az egyes osztályokra vonatkozó tapasztalati kovarianciamátrixokat eredményező $\mathbf{S}_1$ és $\mathbf{S}_2$ ML-becslés alapján:

$$\mathbf{S}_1 = \frac{1}{P_1}\sum_{i\in C_1}^{P} (\mathbf{x}_i - \boldsymbol{\mu}_1)(\mathbf{x}_i - \boldsymbol{\mu}_1)^T \quad \text{és} \quad \mathbf{S}_2 = \frac{1}{P_1}\sum_{i\in C_2}^{P} (\mathbf{x}_i - \boldsymbol{\mu}_2)(\mathbf{x}_i - \boldsymbol{\mu}_2)^T,$$

továbbá

$$\mathbf{S} = \frac{P_1}{P}\mathbf{S}_1 + \frac{P_2}{P}\mathbf{S}_2 = \boldsymbol{\Sigma}$$

Ez a megközelítés tehát a lineáris kapcsolat $a = \mathbf{w}^T\mathbf{x} + w_0$ paramétereit közvetve a Gauss sűrűség-függvény és a priorok mintapontokból való becslése útján határozza meg.

Logisztikus regresszió

Lehetséges a lineáris kapcsolat paramétervektorát közvetlenül is meghatározni az adatokból. A közvetlen vagy ***direkt*** módszer a következő lépésekből áll. A direkt módszer, melynél egy súlyozott lineáris kapcsolatra ható szigmoid függvény után kapunk eredményt. Ezt az eljárást ***logisztikus regressziónak*** is szokták nevezni, annak ellenére, hogy nem regresszióról, hanem osztályozásról van szó.

Itt is a

$$p(C_1|\mathbf{x}) = \frac{p(\mathbf{x}|C_1)\,p(C_1)}{p(\mathbf{x}|C_1)\,p(C_1) + p(\mathbf{x}|C_2)\,p(C_2)} = \frac{1}{1+\exp(-a)}$$

-ből indulunk ki. Ez tehát annak valószínűsége, hogy adott $\mathbf{x}$ bemenet a C_1 osztályba tartozik.

Mivel ez a súlyvektor nemlineáris függvénye, itt lesz egy kis nehézség. Analitikus eredményt nem kapunk, hanem iteratív megoldásunk lesz csak.

Írjuk fel a likelihood függvényt:

$$p(\mathbf{X}, \mathbf{d}|\mathbf{w}) = \prod_{i=1}^{P} y_i^{d_i}\,(1-y_i)^{1-d_i}$$

ahol a $\mathbf{w}$ -től való függés y -on keresztül valósul meg.

Vegyük ennek is a negatív logaritmusát, amit hibafüggvénynek tekinthetünk:

$$\mathcal{L} = \varepsilon(\mathbf{w}) = \ln p(\mathbf{X}, \mathbf{d}|\mathbf{w}) = -\sum_{i=1}^{P} d_i \ln y_i + (1-d_i)\ln(1-y_i)$$

Ennek a likelihood függvénynek a deriválása útján kapjuk a súlyvektor ML becslését a közvetlen úton. A nehézség, hogy most nemlineáris a kapcsolat. Ezért a részletek kihagyásával.

$\nabla\varepsilon(\mathbf{w}) = \frac{\partial \mathcal{L}}{\partial \mathbf{w}} = \frac{\partial \mathcal{L}}{\partial y}\frac{\partial y}{\partial a}\frac{\partial a}{\partial \mathbf{w}} = \sum_{i=1}^{P}(y_i - d_i)\mathbf{x}_i = \mathbf{X}^T(\mathbf{y}-\mathbf{d})$. Megjegyezzük, hogy ez a kifejezés nem tartalmazza a szigmoid deriváltját, mert a likelihood függvény valójában egy kereszt entrópia kritériumfüggvénynek felel meg, és a deriválásnál a szigmoid derivált kiesik és csak a hiba marad meg.

IRLS: Most egy olyan iteratív eljárás következik, mely segítségével a szélsőérték hatékony megtalálása lehetséges. Ez az Iteratív Újrasúlyozott Legkisebb Négyzetes hibájú (Iterative Reweighted Least Squares, IRLS) eljárás.

A gradiens alapú iteratív eljárás a következő

$$\mathbf{w}^{új} = \mathbf{w}^{régi} - \mathbf{H}^{-1}\nabla\varepsilon(\mathbf{w}),$$

ahol $\mathbf{H}$ a Hesse mátrix, a hibafelület második deriváltjaiból képezett mátrix (a másodfokú felület feltételezésével ez a leghatékonyabb gradiens alapú eljárás).

A Hesse mátrix négyzetes hibafelületnél $\mathbf{H} = \nabla\nabla\varepsilon(\mathbf{w}) = \mathbf{X}^T\mathbf{X}$

Ezzel elvégezve az iteratív eljárást:

$$\mathbf{w}^{új} = \mathbf{w}^{régi} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\nabla\varepsilon(\mathbf{w}) = \mathbf{w}^{régi} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T(\mathbf{X}\mathbf{w}-\mathbf{d}) = \mathbf{w}^{régi} - \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\left(\mathbf{X}^T\mathbf{X}\mathbf{w}^{régi} - \mathbf{X}^T\mathbf{d}\right),$$

ami négyzetes hibafelületnél triviálisan egy lépésben kiadja az LS megoldást.

A nemlineáris kapcsolat miatt itt nem lesz négyzetes a hibafelület, ezért egylépéses megoldás sem lesz. A Hesse mátrix a szigmoid függvény miatt: $\mathbf{H} = \nabla\nabla\varepsilon(\mathbf{w}) = \mathbf{X}^T\mathbf{R}\mathbf{X}$, ahol $R_{ii} = y_i(1-y_i)$, a szigmoid derivált.

Az iteratív eljárás ennek megfelelően

$$\begin{aligned}\mathbf{w}^{új} &= \mathbf{w}^{régi} - \left(\mathbf{X}^T\mathbf{R}\mathbf{X}\right)^{-1}\nabla\varepsilon(\mathbf{w}) = \mathbf{w}^{régi} - \left(\mathbf{X}^T\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^T\left(\mathbf{X}\mathbf{w}^{régi} - \mathbf{d}\right)\\ &= \mathbf{w}^{régi} - \left(\mathbf{X}^T\mathbf{R}\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{R}\mathbf{z}\end{aligned}.$$

ahol

$$\mathbf{z} = \mathbf{X}\mathbf{w}^{régi} - \mathbf{R}^{-1}(\mathbf{y}-\mathbf{d})$$

Ez jól láthatóan egy súlyozott LS becslés, ahol a súlymátrix $\mathbf{R}$**.** Szemben azonban a klasszikus súlyozott LS becsléssel, itt nem fix az $\mathbf{R}$ súlymátrix, hanem függ $\mathbf{w}$-től, tehát minden újabb $\mathbf{w}$ értéknél újra meg kell határozni. A súlymátrix tehát minden iterációs lépésben frissítendő.

Az összes eredmény értelemszerűen alkalmazható a jellemzőtérben is. Ebben az esetben minden $\mathbf{x}$ helyére $\boldsymbol{\varphi}(\mathbf{x})$, $\mathbf{X}$ helyére pedig $\boldsymbol{\Phi}$ kerül, egyébként minden változatlanul érvényes.

Logisztikus regresszió Bayes megközelítésben

A logisztikus regresszió Bayes megközelítésben is tárgyalható. Ebben az esetben a posterior meghatározására is szükség van, mely analitikusan nem lehetséges. Ezért csak közelítő módszerek alkalmazásával lehet eredményre jutni. A közelítés lényege, hogy a nem Gauss sűrűségfüggvényt Gauss-szal közelítjük a Laplace approximációt alkalmazva.

Irodalom

[1] Prékopa A.: Valószínűségelmélet, Műszaki Könyvkiadó, 1980.

[2] Vincze I. Matematikai statisztika ipari alkalmazásokkal, Műszaki Könyvkiadó, 1980.

[3] Kecskeméthy- SPSS