Üzleti Elemzések Módszertana

2. Előadás: Osztályozás

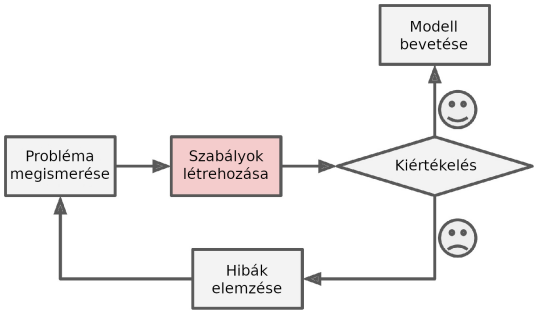
Kuknyó Dániel
Budapesti Gazdasági Egyetem

2023/24
2.félév

A determinisztikus szemléletmód

A hagyományos szoftverfejlesztési folyamatmodell eljárása:

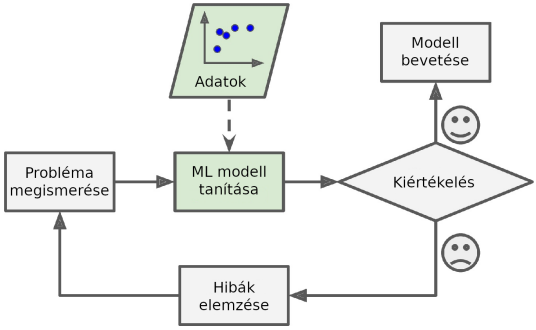
1. Az adott jelenség megfigyelése és adatok rögzítése
2. A megfigyelésekre olyan szabályok kidolgozása, amelyek jól leírják azt
3. A létrejött szabályrendszer kiértékelése
4. Rendszer fejlesztése a hibák alapján
5. Iteráció



A gépi tanulás szemléletmód

A gépi tanulás szemléletének folyamatmodellje:

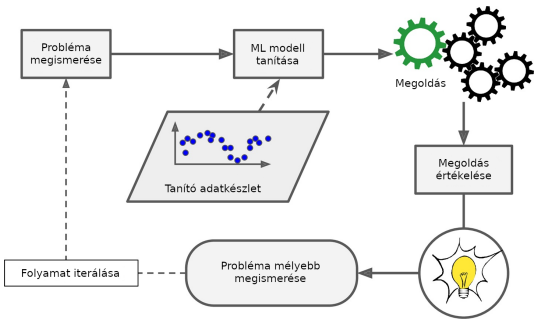
1. Adott jelenség megfigyelése és adatok rögzítése
2. Gépi tanulási modell tanítása az adatokon a szakterületi tudás segítségével
3. Modell kiértékelése
4. Hibák elemzése és kiértékelése
5. Iteráció



Tanítás automatizálása adatalapúan

Az gépi tanuló modellek tanítása és kiértékelése hosszú távon egy iteratív folyamat már létező keretrendszerekkel, mint az MLOps. Ennek számos területen vannak előnyei:

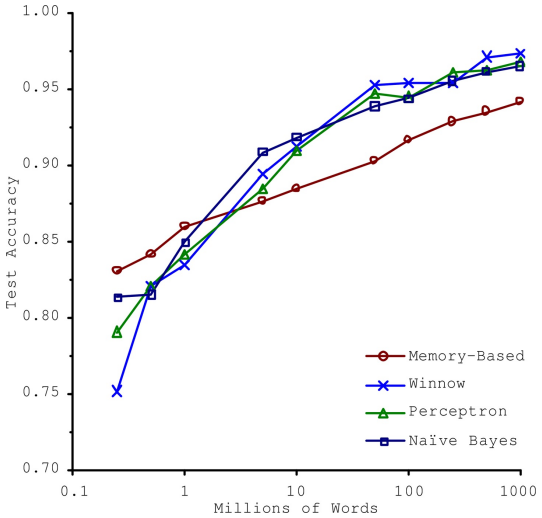
- Adaptáció az új adatokhoz
- Javuló modell teljesítmény
- Hibák és problémák azonosítása
- Új technológiai fejlődés integrálása
- Skálázhatóság és rugalmasság
- Szakterületi következtetések az elemzések által



Az adatok észszerűtlen hatékonysága

2001-es kutatásukban Michele Blanko és Eric Brill kimutatták, hogy a különböző ML algoritmusok **hasonlóan jól teljesítenek a természetes nyelvfelismerés területén mint a hagyományos algoritmusok**, ha elég sok adaton tanítják a modelleket. Ahogy ők fogalmaztak:

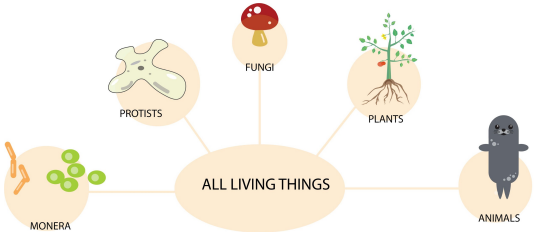
„Az eredmények azt mutatják, hogy újra kell gondolnunk, mire fordítjuk a pénzünket és erőforrásainkat: algoritmusok fejlesztésére, vagy adatgyűjtésre.”



Osztályozás

Osztályozás

Az osztályozás a felügyelt gépi tanulás egyik alapvető feladata, amelynek célja, hogy megtanuljon egy modellt vagy szabályrendszert egy adott bemeneti adat alapján **annak besorolására előre meghatározott kategóriákba vagy csoportokba**.

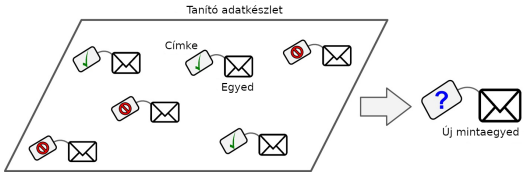


Five Kingdom system classification

Modellalapú osztályozás

Az osztályozó modell feladata, hogy a tanító adathalmaza alapján olyan szabályrendszert hozzon létre, ami **képes elszeparálni egymástól az egyedeket**.

Amennyiben érkezik egy új adatpont, a modell a saját szabályrendszere segítségével már **képes lesz becslést adni annak osztályára vonatkozóan**.

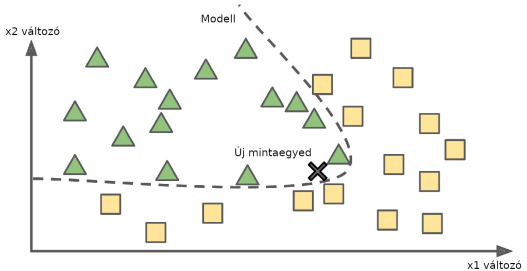


Modellalapú osztályozás

Döntési határ

Olyan **határérték, amelyet a modell állít be** az adatpontok különböző osztályokba való besorolásához.

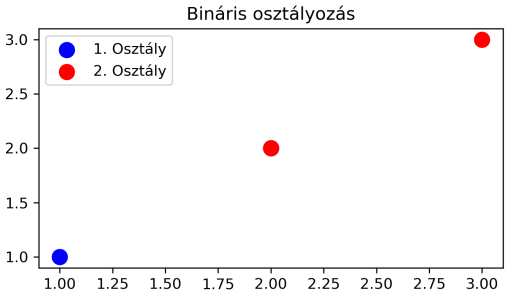
A határ **lehet egy vonal, egy sík vagy akár egy sokdimenziós felület**, attól függően, hogy milyen típusú osztályozó modellt használunk és milyen a bemeneti adatok dimenzionalitása.



Az osztályozás fajtái

Bináris osztályozás

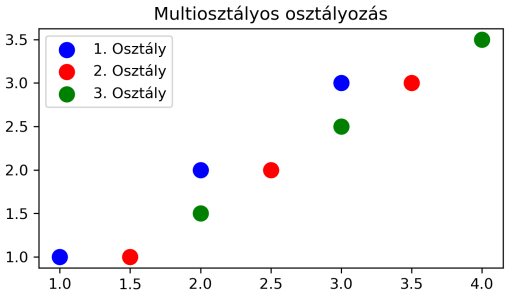
A modell két lehetséges osztály közül valamelyikbe sorolja be az egyedeket.
Minden egyedhez csakis 1 osztály tartozhat.



Az osztályozás fajtái

Multiosztályos osztályozás

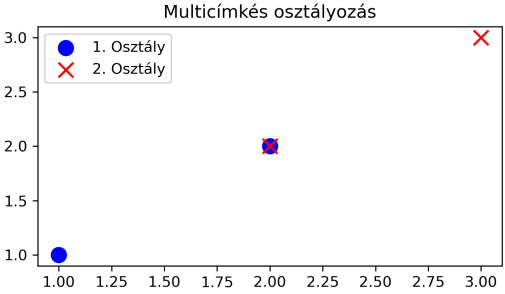
Több, mint két lehetséges kategória létezik, amibe az egyedek besorolhatók, ezek közül az egyikbe fog sorolódni az egyed. Minden egyedhez legalább és legfeljebb 1 osztály tartozik.



Az osztályozás fajtái

Multicímkés osztályozás

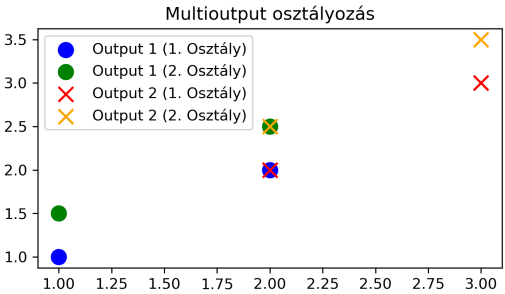
Minden mintaegyedhez több bináris vagy multicímkés címkekategóriából tartozhat osztály.



Az osztályozás fajtái

Multioutput osztályozás

A multicímkés osztályozás generalizált változata. Egy egyedhez egy multicímkés halmazból több elem is tartozhat.

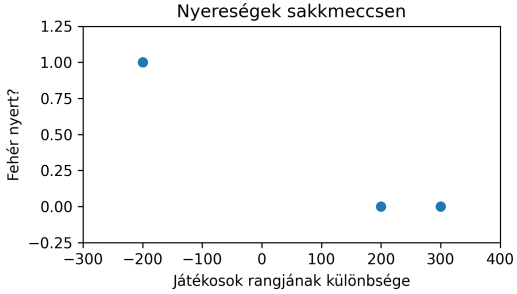


Példa: a probléma bemutatása

A következő kis adathalmaz három sakkjátszmának rögzítette az eredményét. Minden meccs esetén rögzítésre kerültek a következő rekordok:

Különbség	Nyertes
200	0
−200	1
300	0

Ebben az esetben az x változó, a **két játékos rangjának különbsége** a fehér és fekete játékos különbségét jelzi, az y célváltozó pedig egy azt a valószínűséget jelenti, hogy **a fehér nyert-e**.



Példa: lineáris predikció

Az adathalmazra egy lineáris regresszor modellt illesztve az eredmény a következő:

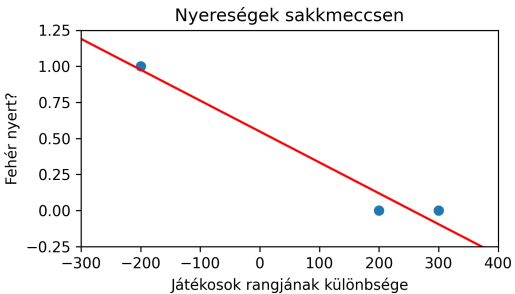
Különbség	Nyertes	Predikció
200	0	0.11
−200	1	0.97
300	0	−0.1

Ebben az estben a lineáris modell:

$$\hat{y} = \theta_0 + \theta_1 \cdot x = 0.54 - 0.0021 \cdot x$$

Ahol $\hat{y}$ a modell predikciója a nyertesre vonatkozóan, θ_0 a konstans torzítás, θ_1 a függvény meredeksége és x a két játékos rangjának különbsége.

Az adatpontokra egy lineáris regressziós függvényt illesztve az illesztett modell a következő lesz:

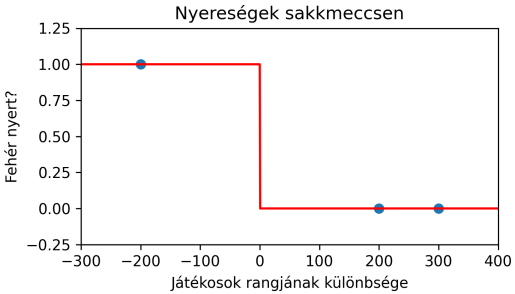


Példa: következtetések

A lineáris modell nem minden esetben ad racionális predikciót az adathalmazra vonatkozóan.

Negatív valószínűségek nem értelmezettek!

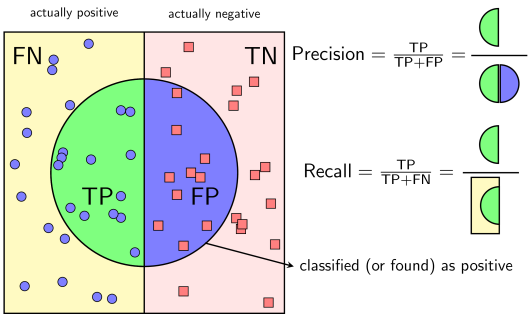
Éppen ezért ha a modellezés célváltozója egy valószínűség, szükség van arra, hogy az illesztett modell szélsőértéke 0 legyen ha a hely $-\infty$ és 1 ha a hely ∞.



Az osztályozás teljesítményének mérése

- **Valós pozitív (TP)**: Pozitív egyed, és annak is van osztályozva
- **Valós negatív (TN)**: Negatív egyed, és annak is van osztályozva
- **Hamis pozitív (FP)**: Negatív egyed, de pozitívnak van osztályozva
- **Hamis negatív (FN)**: Pozitív egyed, de negatívnak van osztályozva

Ennek alapján két fő mutatószám áll elő, amellyel egy osztályozó modellt lehetséges értékelni:



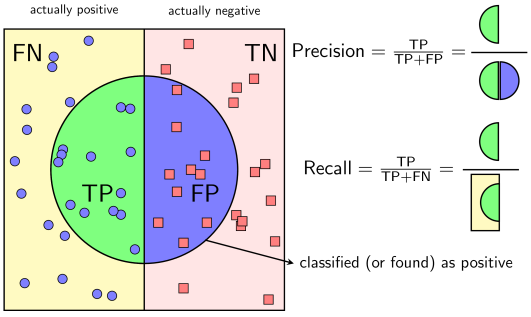
Az osztályozás teljesítményének mérése

Ennek alapján két fő mutatószám áll elő, amellyel egy osztályozó modellt lehetséges értékelni:

Pontosság

Megadja, hogy a pozitívnak osztályozott egyedek közül mekkora hányad volt ténylegesen pozitív:

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



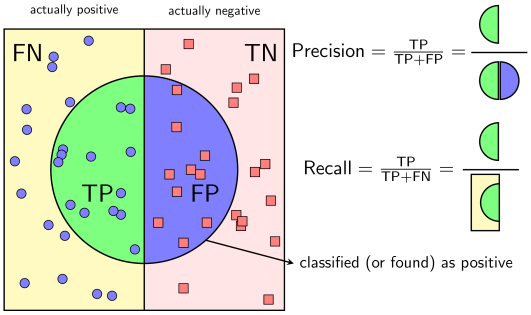
Az osztályozás teljesítményének mérése

Ennek alapján két fő mutatószám áll elő, amellyel egy osztályozó modellt lehetséges értékelni:

Visszahívás

Megadja, hogy az összes pozitív egyed mekkora hányadát osztályozta a modell pozitívnak:

$$R = \frac{TP}{TP + FN}$$

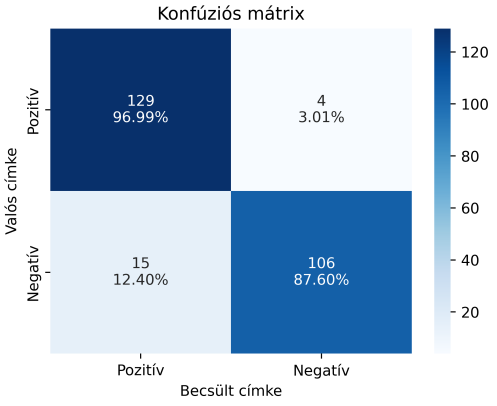


Konfúziós mátrix

A konfúziós mátrix vagy zavarmátrix a statisztikában és gépi tanulásban használatos egy gépi tanulási **algoritmus teljesítményének mérésére**.

A mátrix segít megérteni, hogy **milyen hibákat követett el a modell** és ezáltal segíti a modell finomhangolását és tovább tanítását.

A mátrix általánosítható tetszőleges címke számra.



Logisztikus regresszió

Gépi tanulási módszer kétosztályos (bináris) kimenetelek előrejelzésére, amely valószínűségek megbecslésére szolgál. A logisztikus regresszió eljárása:

1. Adott mintaegyedre annak a valószínűségnek a megbecslése, hogy a modell a pozitív osztályba tartozik-e.
2. Ha a becsült valószínűség magasabb mint egy küszöbérték, a becsült osztály pozitív, egyébként negatív.

$$\hat{y} \begin{cases} 0 & ha\ \hat{p} > \theta \\ 1 & ha\ \hat{p} \leq \theta \end{cases}$$

Ahol $\hat{p}$ a modell által becsült valószínűség, $\hat{y}$ a becsült osztály és θ a küszöbérték.

A logisztikus (szigmoid) függvény

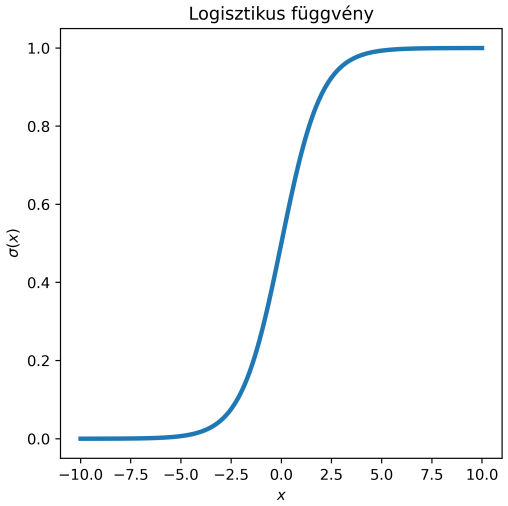
A logisztikus függvény a valószínűségek megbecslésére használt modell típus. A predikció előállításához először az eljárás előállítja z lineáris predikciót:

$$z = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \ldots + \theta_r x_r$$

Majd ezt behelyettesíti a logisztikus függvénybe:

$$\sigma(z) = \frac{1}{1 + e^z}$$

Ahol σ a logisztikus függvény és e a természetes logaritmus értéke.



A logisztikus regresszió költségfüggvénye

A logisztikus regresszió célja, hogy **magas valószínűséggel osztályozzon pozitív egyedeket és alacsony valószínűséggel osztályozzon negatív egyedeket**.

A költségfüggvény egy mintaegyedre:

$$J(\theta) = \begin{cases} -log(\hat{p}) & ha\ \hat{y}=1 \\ -log(1-\hat{p}) & ha\ \hat{y}=0 \end{cases}$$

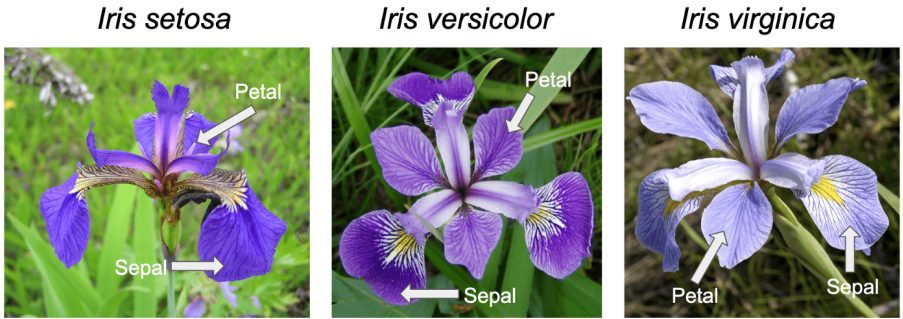
Az összes mintaegyedre kiszámított költségfüggvény az egyedi költségfüggvények összege:

$$J(\theta) = -\frac{1}{n}\sum_{i=1}^{n}\left[y_i \cdot log\left(\hat{p}_i\right) + (1-y_i) \cdot log\left(1-\hat{p}_i\right)\right]$$

A költségfüggvény konvex, de nem létezik a minimum megtalálására zárt formájú számítás. Ennek megfelelően a minimum közelítése iteratív algoritmusokkal lehetséges.

Írisz adathalmaz

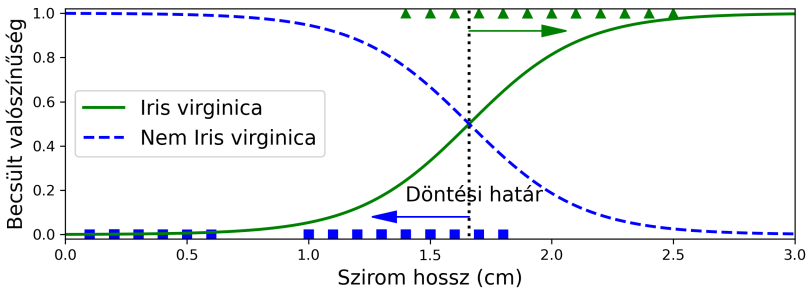
A következő példában a minta adathalmaz Írisz virágokról tartalmaz információkat. Az adathalmazban található oszlopok a virág fajtája (Setosa, Versicolor, Virginica) a csészelevelek hossza és a sziromlevelek hossza.



Logisztikus regresszió az Írisz adathalmazon

A következő példában egy bináris osztályozás a feladat. A **logisztikus regresszió eredménye egy 1D döntési határ**.

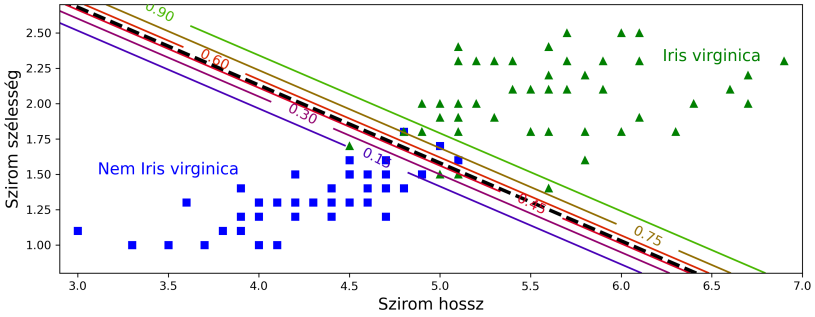
Az Iris Virginica sziromszélességei 1.4-től 2.5cm-ig terjednek, míg a többi Iris virág szirmai 0.1 és 1.8cm közöttiek. A döntési határ 1.65cm körül húzódik. 2cm fölött a modell egészen biztos benne, hogy Virginicáról van szó, 1cm alatt szinte biztos benne, hogy nem tartozik az osztályba.



Logisztikus regresszió több változóval

Ha több x változó alapján történik a modellezés, a döntési határ is több dimenziós lesz. Az alábbi példában a szirom szélesség és a szirom hossz alapján készült a becslés.

Ebben az esetben a becsült valószínűség a 3. dimenzió és a határ ott húzódik, ahol a becsült valószínűség megegyezik a küszöbértékkel, tehát $\hat{p} = \theta$.



Softmax regresszió

A logisztikus regresszió általánosítható tetszőleges számú (k) osztályra. **Ebben az esetben a modell azt becsüli meg, hogy mekkora valószínűséggel tartozik az egyed az adott osztályokba**.

Adott k osztályra számított beletartozási valószínűség:

$$\hat{p}_k = \sigma\left(s\left(x\right)\right)_k = \frac{e^{x^T \theta_k}}{\sum_{j=1}^{k} e^{x^T \theta_j}}$$

Ahol σ a logisztikus függvény és θ_k pedig k osztály tanítható paraméter vektora.

Miután a modell kiszámolta, hogy x mintaegyed mekkora valószínűséggel tartozik minden osztályba, **kiválasztja ezek közül a legnagyobb becsült valószínűséghez tartozót, és ez lesz a becsült érték**:

$$\hat{y} = \underset{k}{argmax}\left(\sigma\left(s\left(x\right)\right)_k\right)$$

Az $argmax$ operátor a változónak azt az értékét téríti vissza, amelyik maximalizálja az adott kritériumot. Ebben az esetben a kritérium a legnagyobb valószínűség.

Softmax regresszió az Írisz adathalmazon

A kép a létrejövő döntési határokat mutatja. Érdemes megfigyelni, hogy az osztályok között létrejövő döntési határok lineárisak. A görbe vonal az ábrán a Versicolor osztályhoz tartozó valószínűség.