**САНХҮҮ ЭДИЙН ЗАСГИЙН ИХ СУРГУУЛЬ** 

**ӨДРИЙН ХӨТӨЛБӨР**

ЭКОНОМИКСИЙН ТЭНХИМ

**ДИПЛОМЫН АЖИЛ**

**Сэдэв:** *Эдийн засгийн таамаглалд машин сургалтын аргыг хэрэглэх нь*

Гүйцэтгэсэн: . . . . . . . . . . . . . . . . . . . . . О.Лхагвасүрэн /FA16B319/

Удирдагч багш : . . . . . . . . . . . . . . . . . . . . . . Д.Хашбаатар /Мастер/

Огноо: 2020.04.19



Улаанбаатар. 2020

# УДИРТГАЛ

**Сэдвийн нэр: Машин сургалтыг эдийн засгийн таамаглалд ашиглах нь**

**Товч танилцуулга:**

**Эдийн засгийн бүтээлийн сангийн индекс:**

**Түлхүүр үгс:**

**АГУУЛГА**

[УДИРТГАЛ i](#_Toc38262041)

[ОРШИЛ iii](#_Toc38262042)

[I БҮЛЭГ. СУДЛАГДСАН БАЙДАЛ 1](#_Toc38262043)

[II БҮЛЭГ. ОНОЛЫН УХАГДАХУУН БА ЗАГВАР 1](#_Toc38262044)

[2.1 Шийдвэрийн модны аргазүй 1](#_Toc38262045)

[2.1.1 Регресийн мод 1](#_Toc38262046)

[2.1.2 Ангиллын мод 3](#_Toc38262047)

[2.2 Баггинг, санамсаргүй ойжуулалт, бүүстинг 5](#_Toc38262048)

[III БҮЛЭГ. ЭМПИРИК СУДАЛГААНЫ АРГА, АРГАЗҮЙ 6](#_Toc38262049)

[IV БҮЛЭГ. ЭМПИРИК СУДАЛГАА 6](#_Toc38262050)

[V БҮЛЭГ. ДҮГНЭЛТ, САНАЛ 6](#_Toc38262051)

[ХАВСРАЛТ 6](#_Toc38262052)

[АШИГЛАСАН МАТЕРИАЛ 6](#_Toc38262053)

**Хүснэгтэн мэдээллийн жагсаалт**

**ЗУРГАН МЭДЭЭЛЛИЙН ЖАГСААЛТ**

**ХАВСРАЛТЫН ЖАГСААЛТ**

No table of figures entries found.

**ТОВЧИЛСОН ҮГС, НЭР ТОМЪЁОНЫ ТАЙЛБАР**

ДНБ Дотоодын нийт бүтээгдэхүүн

ХҮИ Хэрэглээний үнийн индекс

ҮСХ Үндэсний статистикийн хороо

# ОРШИЛ

**Сэдвийг сонгосон үндэслэл буюу яагаад энэ сэдвээр судалгаа хийж байгаагаа тодруулах -** Таамаглалын модел сонглтын

**Судлагдсан байдлын тойм -**

**Судалгааны ажлын зорилго –** Монголын макро эдийн засгийн голлох үзүүлэлтүүдийг машин сургалтын баггинг аргаар үнэлж гүйцэтгэлийг уламжлалт таамаглалын аргуудтай харьцуулах

**Судалгааны ажлын зорилтууд -**

**Судалгааны обьект –** Макроэдийн засгийн голлох үзүүлэлтүүд – ДНБ-ийн өсөлт, инфляци, ажил эрхлэлтийн түвшин.

**Судалгааны ажлын онолын ухагдахуун ба загвар -**

**Эмпирик судалгааны арга, аргазүй -**

**Судалгааны таамаглалууд -**

**Судалгааны шинэлэг тал -**

**Судалгааны ажлын практикийн ач холбогдол -**

# СУДЛАГДСАН БАЙДАЛ

**Загварын тодорхой бус байдал:**

Таамаглалын загварыг симуляцийн аргаар үнэлж

# ОНОЛЫН УХАГДАХУУН БА ЗАГВАР

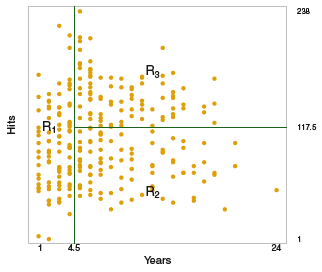
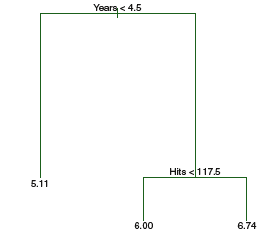
## Шийдвэрийн модны аргазүй

Өгөгдсөн ажиглалтын тусламжтай үзэгдлийг таамаглахдаа бид хуваасан бүлгүүдийн дундаж болон моодыг ашигладаг. Өгөгдлийг тодорхой хэсгүүдэд хуваах дүрмийг дээрээс доош салаалах модоор хураангуйлдаг бөгөөд үүнийг шийдвэрийн мод хэмээдэг. Шийдвэрийн модонд суурилсан аргазүй нь энгийн, тайлбарлахад хялбар юм. Энэхүү бүлгээр бид шийдвэрийн модонд суурилсан аргазүйн ерөнхий хэлбэрийг авч үзэх ба баггинг, санамсаргүй ойжуулалт болон бүүстинг гэх зэрэг хэд хэдэн модыг нэгтгэх замаар таамаглал хийдэг аргуудыг дурдана. Шийдвэрийн мод нь тоон болон чанарын шинжтэй асуудлыг шийддэг бөгөөд эхлээд регрессийн модыг дараа нь ангиллын номыг авч үзэх болно.

### Регресийн мод

Шийдвэрийн мод нь навч (эцсийн зангилаа), дотоод зангилаанаас бүрдэх бөгөөд дээрээс доош чиглэсэн буюу доошоо харсан модтой ижил тул навч (эцсийн зангилаа) нь хамгийн доор байрлана. Зангилаа бүр мөчирөөр холбогдоно. Эцсийн зангилаа нь эцсийн үр дүнг, дотоод зангилаа нь өгөгдлийг хуваах шалгуурыг илэрхийлдэг.

Хялбар байх үүднээс Бэйсболын тамирчиний цалинг регрессийн мод ашиглан дүрслэх жишээг авч үзье. Тамирчины цалинд лигт тоголж буй жил, өнгөрсөн жил гүйцэтгэсэн цохилтын тоо нөлөөлдөг гээд шийдвэрийн модыг дүрсэлвэл доорх мод үүснэ. Зурагт регрессийн мод өгөгдөлд хэрхэн тохирч байгааг болон модыг оройгоос салбарлуулан хуваах дүрмийг харж болно. нь өмнө дурдаж байсан эцсийн зангилаа буюу модны навчийг илэрхийлэх ба модны эцэст байрлана. Дотоод зангилаа болох “” болон “” нь өгөгдлийг гурван бүлэгт хувааж байгаа юм. Ингээд таамаглалын загвар бэлэн боллоо. Тухайлбал лигт 4.5 болон түүнээс дээш жил тоголсон, өнгөрсөн улирал 117.5 – оос бага цохилт хийсэн тамирчиний цалин 3 дугаар навч буюу 6 хэмээн таамаглагдаж байна.

Энгийн регрессийн модыг байгуулах аргыг хураангуй авч үзвэл таамаглагч хувьсагчидийг үл давхцах бүлгүүдэд хуваах ба нэг бүлэгт хамаарагдаж буй ажиглалтын хувьд ижил таамаглал хийх буюу хамааран хувьсагчийн дундаж ижил байхаар бүлэг бүрийг тусад нь авч үзнэ. Нэгдүгээр алхамыг хэрэгжүүлэхэд бид таамаглалыг алдааны квадратуудын нийлбэр (RSS) буюу дараах тэгшитгэлийг хамгийн бага байлгах шаардлагыг авч үздэг.

Энд нь дүгээр бүлэг дэхь дундаж хариу үйлдэл юм. Бид дээрээс доош чиглэсэн рекурсив хоёртын хуваалтыг ашигладаг бөгөөд модны орой буюу бүх таамаглагчдыг бүлгүүдэд хуваах замаар мөчир үүсгэж модыг доош салбарлуулдаг юм. Рекурсив хоёртын хуваалтыг бид эхлээд таамаглагч болон түүнийг болон гэсэн 2 хэсэгт хуваах хуваалтын цэгийг сонгох шаардлагатай. Өөрөөр өгөгдлийг бүлэг бүрийн хувьд RSS буюу таамаглалын алдааны квадратуудын нийлбэрийг хамгийн бага байхаар хуваах таамаглагч болон хуваалтын цэгийг олох явдал юм. Үр дүнд нь нэг бүлэгт 1 хүртэлх ажиглалттай байх хүртэл үргэлжилж болно.

#### Мод тайрах

Дээр дурдсан шийдвэрийн мод ургуулах үйл явц нь сургалтын өгөгдлийн хувьд сайн үр дүнг өгч болох ч өгөгдлийг хэт үнэлэх, тестийн өгөгдлийн хувьд муу гүйцэтгэлтэй байх эрсдэлтэй юм. Учир нь ургуулсан мод маань хэт цогц бүтэцтэй байх магадлалтай. Цөөн хуваалттай намхан мод багахан хэмжээний гажуудалтай ч бага хэлбэлзэлтэй, тайлбарлахад хялбархан байх хандлагатай. RSS ийг байж болох хамгийн ихээр бууруулах хуваалтыг агуулсан модыг ургуулах нь намхан мод үүсгэх ч хэт богиныг харсан хуваалт нь ирээдүйн үнэ цэнэ бүхий хуваалтыг үгүйсгэх эрсдэлтэй.

Маш том мод ургуулаад түүнийгээ тайрч засах замаар дэд модуудыг гаргаж авах нь илүү сайн арга юм. Мэдээж дэд модуудыг сонгож авахад ч гэсэн шалгуур байх бөгөөд энэ нь шалгуурын алдааны түвшин бага байх явдал юм. Үүнийг бид cross- validation- валидэйшн болон валидэйшн аргуудын тусламжтай гүйцэтгэнэ. Гэхдээ бүх боломжит дэд мод дээр үүнийг тооцох нь утгагүй бөгөөд цөөн хэдэн дэд модыг сонгож авах хэрэгтэй. Үүнийг cost Complexity тайралт эсвэл хамгийн сул холбоос тайралт хэмээн нэрлэдэг. Боломжит бүх модыг сонгохын оронд бид сөрөг биш утгатай параметрээр индексжүүлсэн дэд моднуудыг авч үзнэ. дэд модонд харгазах нь дараах тэгшитгэлийг хамгийн бага утгатай байхаар сонгогддог.

Энд нь модны эцсийн зангилаа, нь дүгээр эцсийн зангилаанд харгалзах тэгш өнцөгт, нь тэгш өнцөгт буюу дүгээр сургалтын өгөгдлийн дундаж буюу таамаглагдсан утга. Чухамдаа параметр нь сургалтын өгөгдөлийн тохирсон байдал болон дэд модны Complexity хоёрын ацан шалааг зохицуулдаг. Хэрвээ бол болох буюу дээрх тэгшитгэл нь сургалтын өгөгдлийн алдааг хэмжих юм. Бид -ийн утыг cross- validation болон валидэйшн аргуудын тусламжтай олох ба энэ утгад харгазах дэд моднуудыг өгөгдлөөс сонгон авна. Алхамуудыг Алгоримт 8.1-д харуулав.

Алгоритм 8.1 Регрессийн мод ургуулах

1. Рекурсив хоёртын хуваалтын тусламжтай сургалтын өгөгдөл дээр том мод ургуулах. Энэ үйл явц нь эцсийн зангилаад хамгийн бага ажиглалтын тооны хязгаар хүртэл үргэлжилнэ.
2. Ургуулсан том модоо функцын тусламжтай тайрч засан хамгийн шилдэг дэд модны цувааг олж авна.
3. K-дахин cross validation – г ашиглан – ийн утгыг олно. Энэ нь сургалтын өгөгдлийг K ширхэг бүлэгт хуваана гэсэн үг. байх бүлэгийн хувьд:
   1. 1 болон 2 дугаар алхамыг К дугаар бүлгээс бусал хэсэгт давтан гүйцэтгэнэ
   2. функцийн хувьд өгөгдлөөс үлдсэн K дугаар бүлгийг ашиглан таамаглалын алдааны квадратыг үнэлнэ

функцын тусламжтай үр дүнг дундажлах замаар дундаж алдааг хамгийн бага байлгах – г сонгон авна.

1. Сонгогдсон – д харгалзах дэд моднуудыг олж авна.

### Ангиллын мод

Ангиллын мод нь регрессийн модтой бараг л ижилхэн боловч тоон биш чанарын хувьсагчийг таамаглах болно. Регрессийн модыг санавал бид нэг ижил эцсийн зангилаад хамаарах ажиглалтын хувьд хамааран хувьсагчийн тооцож, дундажлаж замаар таамаглал хийж байсан. Харин ангиллын модны хувьд навч буюу эцсийн зангилаа бүрийн хувьд хамгийн их давтагдсан ангилалын тусламжтай таамаглал хийдэг.

Регрессийн модыг ургуулахад RSS – ийн тусламжтай рекурсив хоёртын хуваалтыг ашигладаг бол ангиллын модонд үүнийг ашиглах боломжгүй юм. Учир нь RSS -ийг чанарын хувьсагчийн хувьд тооцох боломжгүй тул *ангиллын алдааны түвшинг* ашигладаг юм. Тиймээс тухайн нэг навчинд буй хамгийн их давтагдсан ангилалыг хувуурилах болно, харин ангиллын алдааны түвшин нь тухайн бүлгийн ажиглалтын тоо болон хамгийн их давтагдсан ангилалд хамаарахгүй ажиглалтуудын тооны харьцаа юм.

Энд нь дүгээр бүлэгт дугаар ангилалын эзлэх хувийг илэрхийлнэ. Гэвч энэхүү ангиллын алдаа хэмээх нь мод ургуулахад хангалттай мэдрэмтгий биш тул практикт 2 аргыг өргөнөөр ашигладаг.

Жинийн индекс нь дараах байдлаар тодорхойлогдох бөгөөд K ангиллуудын нийт хэлбэлзлийг хэмждэг. Хэрвээ нь 1 эсвэл 0 тэй тэнцүү бол Жинийн индекс бага гарах нь илэрхий байна. Индексийн бага утга нь тухайн зангилаанд нэг анги зонхилсон байгааг итгэх агаад энэ нь зангилааны цэвэр байдлыг хэмжих боломж бүхий Жинийн индексийн давуу тал юм.

Дараагийн боломжит арга бол хөндлөн энтропи гэж нэрлэгдэх хэмжүүр :

Энд болон нөхцөл биелэнэ. Хэрвээ бүх – ийн утга 1 эсвэл 0 – тэй ойрхон бол хөндлөн энтропи бага утгатай байна. Жинийн индекстэй ойролцоо гарах энэхүү хэмжүүр нь мөн л дүгээр зангилааны цэвэр байдлыг хэмжинэ.

Ангиллын модыг ургуулахдаа Жинийн индекс эсвэл Кроссентропи нь ихэвчлэн тодорхой хуваагдлын чанарыг үнэлэхэд ашиглагддаг ба аль аль нь зангилааны цэвэр байдалд ангиллын алдааны түвшингээс илүү мэдрэмтгий байдаг. Модыг тайрахад эдгээр гурван аргуудын аль нэгийг нь ашиглаж болно, гэхдээ эцсийн тайрсан модны зорилго нь таамаглалын нарийвчлал бол ангиллын алдааны түвшинг ашиглах нь илүү юм.

#### Шийдвэрийн модны сул ба давуу тал

* Давуу тал
  + Бусад хүмүүст тайлбарлахад хялбар, ойлгомжтой. Шугаман регрессээс илүү хялбар гэж үздэг.
  + Зарим хүмүүс шийдвэрийн мод нь бодит амьдрал дээрх хүний шийдвэр гаргалттай ижил хэмээн үздэг.
  + Зураглахад хялбар бөгөөд мэргэжлийн бус хүмүүс ч төвөггүй ойлгоно.
  + Чанарын шинжтэй таамаглалыг дамми үүсгэхгүйгээр хялбархан шийддэг.
* Сул тал
  + Бусад машин сургалтын аргуудтай харьцуулахад нарийвчлал бага, вариац өндөр

## Баггинг, санамсаргүй ойжуулалт, бүүстинг

Баггинг, санамсаргүй ойжуулалт болон бүүстингийн тусламжтай илүү хүчирхэг загварыг гарган авах боломжтой байдаг.

### Баггинг

Өмнөх бүлэгт авч үзсэн шийдвэрийн мод нь харьцангуй хэлбэлзэл ихтэй байдаг. Өөрөөр хэлбэл сургалтын өгөгдлийг бид санамсаргүйгээр хуваагаад хоёуланд нь шийдвэрийн модыг ургуулвал үр дүн нь маш их зөрүүтэй гарна. Эсрэгээрээ ялгаатай өгөгдөл дээр энэхүү арыг давтвал хэлбэлзэл багатай ижил үр дүн гарна, мөн болон -ийн харьцаа их үед шугаман регресс бага хэлбэлзэлтэй байх хандлагатай. Бүүсттрап нэгтгэл буюу баггинг аргазүй нь статистик сургалтын аргуудын вариацыг бууруулах зорилготой ба шийдвэрийн модтой холбоотой асуудалд өргөнөөр ашигладаг.

Бие биеэсээ үл хамаарах, вариацтай ширхэг үл хамаарах ажиглалт өгсөн гэвэл дундаж нь болох – ийн вариац болно. Өөрөөр хэлбэл ажиглалтуудыг дундажлах нь вариацыг бууруулдаг. Таамаглалын вариацыг бууруулах, нарийвчлалыг сайжруулах уламжлалт арга нь эх олонлогоос маш олон сургалтын өгөгдөл салган авч, тус бүрд нь таамаглалын загвар боловсруулж, үр дүнг дундажлах явдал билээ. Өөрөөр хэлбэл гэсэн ширхэг бие даасан сургалтын өгөгдөл гаргаж авсан гэвэл бага вариацтай статистик сургалтын загварыг дараах байдлаар олно:

Мэдээж бодит байдал дээр олон сургалтын өгөгдөл байх боломжгүй тул бүүстрап ашиглан ганц сургалтын өгөгдлөөс олон тооны түүврийг гарган авдаг юм. ширхэг ялгаатай бүүстрап хийгдсэн сургалтын өгөгдөл бий болгон, түүнийгээ нэгтгэдэг. Тухайлбал ширхэг сургалтын өгөгдлийн тусламжтай таамаглал хийж нэгтгэнэ.

Үүнийг л баггинг гэх бөгөөд шийдвэрийн модыг гүн ургуулах боломж олгон, тайрч засах шаардлагагүй болгодог. Мод бүр дангаараа гажуудал багатай, вариац өндөртэй байх боловч үүнийг дундажлах замаар бууруулдаг. Регрессийн модны хувьд дундажлах нь илэрхий юм. Харин ангиллын модны хувьд *олонхын санал* буюу B ширрэг таамаглалд хамгийн их тохиолдох ангилалыг тооцох аргыг ашигладаг.

**Цүнхнээс гадна буй алдааны тооцоолол**

Баггинг ашигласан загвырын хувьд тестийн өгөгдлийн алдааг шууд тооцох арга байдаг буюу cross-validation, validation хийх шаардлагагүй. Баггинг арга мод нь ойролцоогоор нийт ажиглалтын гуравны хоёртой тэнцүү хэмжээний өгөгдлийг ашигладаг ба үлдсэн хэсгийн цүнхнээс гадна буй хэсэг гэж нэрлэдэг. Тестийн өгөгдлийн алдааг тооцохдоо цүнхний гадна байх өгөгдлийг мод бүрээр таамаглаж дундажлах замаар 1 таамаглал гарган авч бодит өгөгдөлтэй харьцуулан үздэг.

# ЭМПИРИК СУДАЛГААНЫ АРГА, АРГАЗҮЙ

# ЭМПИРИК СУДАЛГАА

# ДҮГНЭЛТ, САНАЛ

# ХАВСРАЛТ

# АШИГЛАСАН МАТЕРИАЛ