Орлогын Таамаглалын Статистик Шинжилгээ / Income Prediction

Javkhlan

2025-12-02

Table of contents

1 Оршил (Introduction)

Энэ төслийн зорилго нь Adult Income өгөгдлийн санг ашиглан хувь хүн жилийн орлого >50K эсэхийг таамаглах юм. Бид өөрсдийн хэрэгжүүлсэн логистик регресс загварыг sklearn Pipeline-той хослуулан ашигласан. Загвар нь L2 регуляризацитай, batch градиент бууруулалтаар суралцаж, сурах хурдны бууралт хэрэглэдэг.

Логистик магадлалын загвар:

$$P(Y = 1 \mid X = x) \;=\; \sigma\!\left(w^{\top} x + b\right) \;=\; \frac{1}{1 + e^{-\left(w^{\top} x + b\right)}} \tag{1}$$

Сургалтын алдагдал (хоёртын cross-entropy + L2):

$$J(w, b) \;=\; -\frac{1}{m} \sum_{i=1}^{m} \Big[y^{(i)} \log \hat{y}^{(i)} + (1 - y^{(i)}) \log(1 - \hat{y}^{(i)}) \Big] \;+\; \frac{\lambda}{2m} \, \|w\|^2$$

Сурах хурдны бууралт:

$$\alpha_t \;=\; \frac{\alpha_0}{1 + \text{decay} \cdot t}$$

Pipeline: - Тоон шинжүүд → StandardScaler - Категори шинжүүд → OneHotEncoder(handle_unknown=``ignore'' - Классын жин → class_weight=``balanced''

2 Өгөгдөл (Data)

Эх сурвалж: Adult Income (Census Bureau). Зорилтот хувьсагч: `income_>50K` (0/1). Өгөгдөл нь тоон ба категори шинжүүдийн холимог бөгөөд ангиудын тэнцвэргүй байдал ажиглагдана (>50K нь цөөн).

1) Ашигласан ба унагаасан хувьсагчид (сплит хийх шатанд цөөлсөн)

- Ашигласан (kept)
  - Тоон: `age`, `educational-num`, `capital-gain`, `capital-loss`, `hours-per-week`
  - Категори: `education`, `marital-status`, `occupation`, `gender`
- Унагаасан (dropped) --- давхардал/тайлбарлах ач холбогдол бага, эсвэл дуу чимээ ихтэй багануд:
  - Жишээлбэл: `workclass`, `fnlwgt`, `relationship`, `race`, `native-country`, бусад үл ашигласан бүх багануд

2) Хуваалт

- 80% сургалт, 20% баталгаажуулалт (stratify=`income_>50K`)
- Хуваасны дараа зөвхөн дээрх kept багануд болон зорилтот багана бүхий `train_split.csv`, `val_split.csv` файлд хадгалсан

3) Урьдчилсан боловсруулалт

- Тоон шинжүүдийг стандартчилж, категори шинжүүдийг one-hot кодчилно
- Үүссэн онцлог вектор дээр логистик регресс сургав:
  - learning_rate ☒ 0.1, max_iter ☒ 800, reg_lambda = 1e-4, lr_decay = 1e-4
  - class_weight = ``balanced'', threshold = 0.5

4) Ангиудын тэнцвэргүй байдал

- ☒50K анги нь давамгай (ихэнх), >50K анги нь цөөн тул F1/Recall зэрэг тэнцвэртэй үзүүлэлтүүдийг давхар харна. Cross-entropy алдагдал болон жинлэлт нь цөөн ангийг үл тоомсорлохоос сэргийлдэг.

3 Логистик загвар / Logistic Model

$$P(Y = 1 midX = x) = sigma(w^T x + b) = frac11 + e^{-(w^T x + b)} \tag{2}$$

Дээрх Equation 2 тэгшитгэл нь магадлалыг сигмоид функцээр дүрслэнэ.

4 Кирилл үсгийн тест / Cyrillic Smoke Test

Доорх мөрүүд зөв гарч буй эсэхийг PDF дээр шалгана:

Монгол хэлний кирилл: Үүнд ө, ү, ч, ж, й, h, ё тэмдэгтүүд орно.
Туршилт өгөгдөл: ``Сургалтын загвар амжилттай ажиллалаа.''
English fallback: Model trained successfully.

5 Үзүүлэлт / Metrics

Metric	Value
Accuracy	0.85
Precision	0.79
Recall	0.72
F1 Score	0.75

6 Дүгнэлт / Conclusion

Энэхүү тайлан нь Adult Income өгөгдлийн сан дээр логистик регресс ашиглан орлого >50K эсэхийг таамаглахад чиглэсэн. Логистик регресс нь хоёр ангиллын хоорондох магадлалыг загварчлахад ашиглагддаг статистикийн үндсэн арга юм. Манай загвар нь 85% үнэн зөв, 79% нарийвчлал, 72% сануулга, 75% F1 оноо үзүүлсэн. Эдгээр үзүүлэлтүүд нь ангиллын тэнцвэргүй байдлыг харгалзан үзэхэд чухал ач холбогдолтой.

7 Ном зүй / References