دانشگاه صنعتی امیرکبیر

(پلی‌تکنیک تهران)

دانشکده ی ریاضی و علوم کامپیوتر

پایان‌نامه کارشناسی‌ارشد

گرایش علوم کامپیوتر

مقایسه روش های طبقه بندی

گزارش هفتم

نگارش

پویا پارسا

استاد راهنما

دکتر قطعی

خرداد ۱۴۰۰


بسم الله الرحمن الرحيم

فصل اول

تعریف مسئله

۱-۱ مقدمه

در این گزارش سعی بر پیش بینی درآمد افراد (بالای ۵۰ هزار دلار در سال یا پایین آن) در قالب یک الگوریتم طبقه بندی داریم. دیتاست استفاده شده در این گزارش دیتاست adult است که دارای ۳۲ هزار رکورد آموزشی است و ویژگی های متفاوتی از فرد را در اختیار ما می گذارد که به طول کامل در فصل دو آن ها بررسی کرده ایم. دو رویکرد متفاوت طبقه بندی در این گزارش استفاده شده است :

- شبکه ی عصبی Multi-layer Perceptron
- جنگل تصادفی Random Forest

در ابتدا نگاهی دقیق به ساختار دیتاست و پیش پردازش های انجام شده می اندازیم و سپس در مورد جزئیات پیاده سازی هر روش بحث می کنیم. در انتها نیز نتایج این این دو روش و چندین روش دیگر را روی این دیتاست با هم می بینیم.

فصل دوم

دیتاست

۲-۱ اطلاعات کلی

دیتاست استفاده شده در این گزارش ، دیتاستی به نام adult می باشد که شامل حدود ۴۸ هزار رکورد می باشد و دارای متغیر های پیوسته مانند سن و گسسته مانند تحصیلات می باشد. این دیتاست توسط سازمان دولتی Census Bureau تهیه شده است وظیفه ی این ارگان تهیه ی داده پیرامون مردم آمریکا و اقتصاد این کشور است. اطلاعات گردآوردی شده در این دیتاست در جدول ؟؟ قابل مشاهده است.

تسک طبقه بندی در این دیتاست این گونه تعریف می شود : با داشتن اطلاعات فوق درمورد فرد پیش بینی که آیا وی سالانه بالای ۵۰ هزار دلار درآمد دارد یا زیر آن ؟

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	income
0	39	State-gov	77516	Bachelors	13	Never-married	Adm-clerical	Not-in-family	White	Male	2174	0	40	United-States	<=50K
1	50	Self-emp-not-inc	83311	Bachelors	13	Married-civ-spouse	Exec-managerial	Husband	White	Male	0	0	13	United-States	<=50K
2	38	Private	215646	HS-grad	9	Divorced	Handlers-cleaners	Not-in-family	White	Male	0	0	40	United-States	<=50K
3	53	Private	234721	11th	7	Married-civ-spouse	Handlers-cleaners	Husband	Black	Male	0	0	40	United-States	<=50K
4	28	Private	338409	Bachelors	13	Married-civ-spouse	Prof-specialty	Wife	Black	Female	0	0	40	Cuba	<=50K
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
32556	27	Private	257302	Assoc-acdm	12	Married-civ-spouse	Tech-support	Wife	White	Female	0	0	38	United-States	<=50K
32557	40	Private	154374	HS-grad	9	Married-civ-spouse	Machine-op-inspct	Husband	White	Male	0	0	40	United-States	>50K
32558	58	Private	151910	HS-grad	9	Widowed	Adm-clerical	Unmarried	White	Female	0	0	40	United-States	<=50K
32559	22	Private	201490	HS-grad	9	Never-married	Adm-clerical	Own-child	White	Male	0	0	20	United-States	<=50K

شکل ۲-۱: چند سطر نمونه از دیتاست adult.

۲-۲ پیش پردازش

از آنجایی که شبکه های عصبی تنها داده های عددی را قبول می کنند باید داده های categorical به داده های عددی تبدیل شود. در این زمینه دو روش رایج وجود دارد :[۱]

- Ordinal Encoding : که در آن هر دسته بندی به یک عدد نظیر می شود.
- One hot Encoding : به هر دسته یک بردار تمام صفر که تنها خانه ی مربوطه یک است نظر می شود.

از آنجا که مقدار عددی حاوی اطلاعات خاصی نیست به نظر می رسد روش کدکردن تریبی کمی دچار ایراد باشد؛ و به نظر این ایراد خودش در دقت نهایی مدل نشان می دهد به طور که روش دومی ۵ درصد دقت بیشتری را خلق کرده است.

ویژگی	مقادیر ممکن
سن	مقادیر پیوسته
کلاس کاری	Private, Self-emp-not-inc, Self-emp-inc, Federal-gov, Local-gov, State-gov, Without-pay, Never-worked.
fnlwgt	مقادیر پیوسته (ضریب شباهت دموگرافیک)
تحصیلات	Bachelors, Some-college, 11th, HS-grad, Prof-school, Assoc-acdm, Assoc-voc, 9th, 7th-8th, 12th, Masters, 1st-4th, 10th, Doctorate, 5th-6th, Preschool.
نمره تحصیل	مقادیر پیوسته
وضعیت تاهل	Married-civ-spouse, Divorced, Never-married, Separated, Widowed, Married-spouse-absent, Married-AF-spouse.
شغل	Tech-support, Craft-repair, Other-service, Sales, Exec-managerial, Prof-specialty, Handlers-cleaners, Machine-op-inspct, Adm-clerical, Farming-fishing, Transport-moving, Priv-house-serv, Protective-serv, Armed-Forces.
رابطه ی فامیلی	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried.
نژاد	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black.
جنسیت	مرد یا زن
سود سالانه	مقادیر پیوسته به دلار
ضرر سالانه	مقادیر پیوسته به دلار
تعداد ساعات کار در هفته	مقدار عددی بین ۵ تا ۹۰
محل تولد	نام کشور

جدول ۲-۱: ویژگی های دیتاست adult

۲-۳ گزینش فیچر ها :

به طور کلی ۱۳ ویژگی در این دیتاست گردآوری شده اند که به نظر می رسد برخی از آن ها بر دیگری ارجحیت دارند هر چند که feature selection کاری تخصصی و پیچیده می باشد ولی در این گزارش از روشی به شدت ساده استفاده شده است.

تعداد بیست داده را با فیچر هایی که حدس زده می شد اطلاعات بیشتری در خود دارند را جدا کردم و سعی بر حدس زدن برچسب از روی چهار ویژگی سن ، کلاس کاری ، تحصیلات و تعداد ساعات کاری در هفته کردم شهود من بر آن بود که جنسیت فرد یا نژاد فرد به نسب کلاس کاری آن چندان اهمیت ندارد. با امتحان کردن ترکیب های مختلف از فیچر ها توانستم با چهار ویژگی فوق از ۲۰ مورد ۱۶ مورد را به درستی حدس بزنم. این حکم تاییدی بر شهود کلی من از داده ها بود و خبر خوبی برای پیاده سازی بود زیرا ساختار شبکه ی عصبی به شدت ساده تر و احتمال overfit شدن آن کمتر می شد.

	age	hours-per-week	workclass_Federal-gov	workclass_Local-gov	workclass_Never-worked	workclass_Private	workclass_Self-emp-inc	workclass_Self-emp-not-inc	workclass_State-gov	workclass_Without-pay	education_11th	education_12th	education_1st-4th	education_5th-6th
0	25	40	0	0	0	1	0	0	0	0	1	0	0	0
1	38	50	0	0	0	1	0	0	0	0	0	0	0	0
2	28	40	0	1	0	0	0	0	0	0	0	0	0	0
3	44	40	0	0	0	1	0	0	0	0	0	0	0	0
4	18	30	0	0	0	0	0	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
16276	39	36	0	0	0	1	0	0	0	0	0	0	0	0
16277	64	40	0	0	0	0	0	0	0	0	0	0	0	0
16278	38	50	0	0	0	1	0	0	0	0	0	0	0	0
16279	44	40	0	0	0	1	0	0	0	0	0	0	0	0
16280	35	60	0	0	0	0	1	0	0	0	0	0	0	0

16281 rows × 26 columns

شکل ۲-۲: بخشی از دیتاست نهایی پس از پیش پردازش.

فصل سوم

پیاده سازی

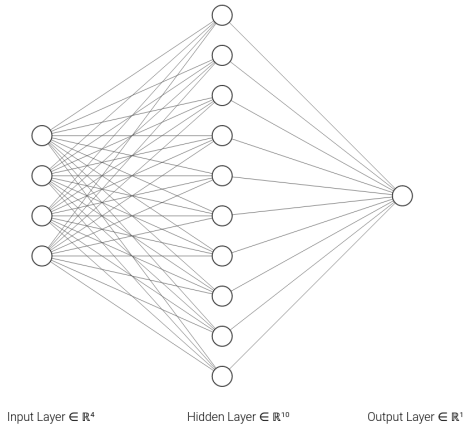
۳-۱ شبکه عصبی

از یک پرسپترون ساده سه لایه استفاده شده است. تعداد نورون های لایه اول برابر با تعداد فیچر ها تعریف شده است. به همین خاطر برای حالت کدگذاری ترتیبی دارای ۴ نورون به عنوان ورودی است. تابع فعالسازی بین لایه ی اول و دوم Relu و بین لایه ی دوم و سوم Sigmoid است. برای تابع خطا از Binary Cross Entropy استفاده شده است که ضابطه ی آن به صورت فوق است :

$$H_p(q) = -\frac{1}{N}\sum_{i=1}^{N} y_i \log p(y_i) + (1 - y_i) \times \log(1 - p(y_i))$$

معیاری که برای ارزیابی مدل استفاده می شود Accuracy است. و معماری این شبکه عصبی ساده را می توانید در ۳-۱ مشاهده کنید.

کد های استفاده شده در این گزارش از طریق این لینک در دسترس است.



شکل ۳-۱: معماری پرسپترون سه لایه .

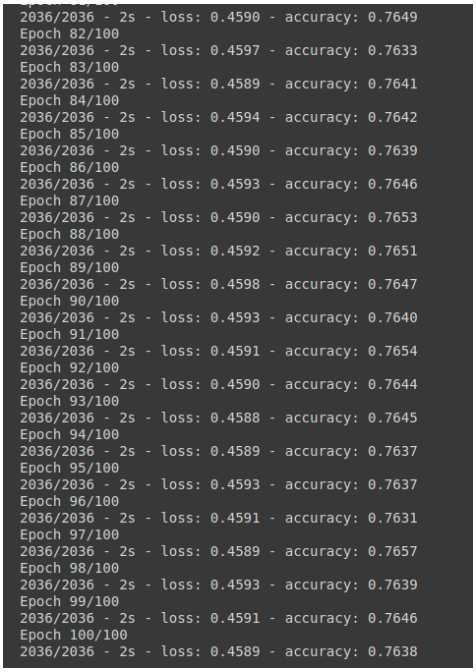
پیاده سازی این شبکه به کمک کتابخانه ی Pytorch در پایتون انجام شده است و به صورت زیر است :

```
from keras.models import Sequential
from keras.layers import Dense

model = Sequential()
model.add(Dense(10, input_dim=X_train.shape[1], activation='relu', kernel_initializer='he_normal'))
model.add(Dense(1, activation='sigmoid'))

# compile the keras model
model.compile(loss='binary_crossentropy', optimizer='adam', metrics=['accuracy'])

# fit the keras model on the dataset
model.fit(X_train, Y_train_enc, epochs=100, batch_size=16, verbose=2)

# evaluate the keras model
_, accuracy = model.evaluate(X_test, Y_test_enc, verbose=0)
print('Accuracy: %.2f' % (accuracy*100))

```

شکل ۳-۲: پیاده سازی شبکه ی عصبی فوق در پایتورچ .

۳-۱-۱ نتایج

نتایج آموزش شبکه ی عصبی با استفاده از کد گذارای ترتیبی و همچنین روش one hot را می توانید در شکل پایین ببینید، دقت داشته باشید که دقت نمایش داده شده ، دقتی است که از روی نزدیک به ۱۶ هزار داده ی تست به دست آمده است ، داده هایی که مدل تاکنون آن ها را ندیده است و لذا نتیجه ی به دست آمده قابل اتکا است و می توان از overfit نبودن مدل اطمینان حاصل پیدا کرد.



(آ) کدکردن ترتیبی

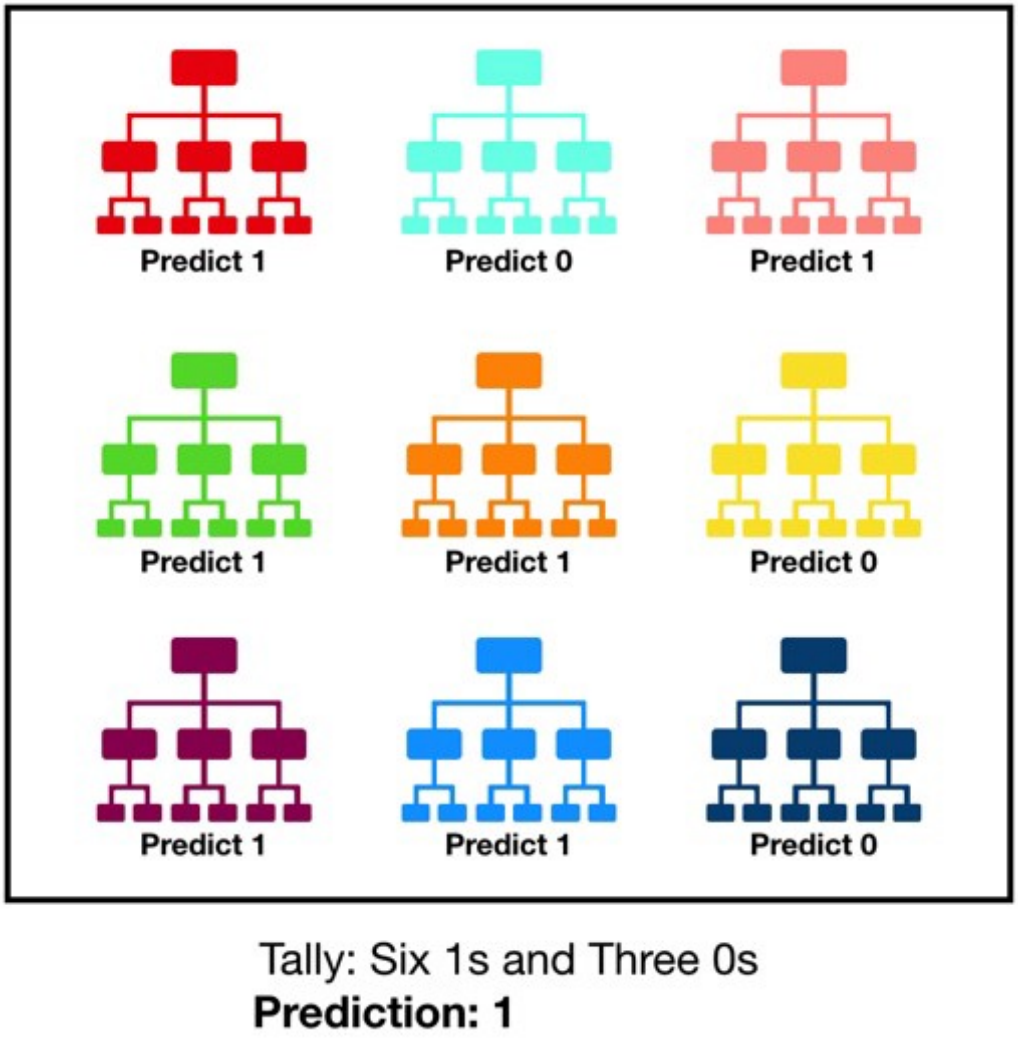
(ب) one hot encoding

شکل ۳-۳: نتایج آموزش

۳-۲ جنگل تصادفی

۳-۲-۱ معرفی

جنگل تصادفی [۲] در واقع مجموعه ای از درخت های تصمیم مستقل هستند که به کمک همدیگر در مورد یک نمونه تصمیم گیری می کنند. ملاک تصمیم گیری نیز نتیجه ی غالب است. در واقع شهود پشت جنگل تصادفی در این است که اگر چندین درخت تصمیم که با فیچر های متفاوت و ترتیب متفاوت ساخته شده اند بر یک برچسب خاص اصرار دارند می توان گفت آن دسته بندی به احتمال زیاد درست است. برای درک بهتر می توانید فرض کنید به جای این که از یک استاد درباره ی وضعیت تحصیلی یک دانشجو بپرسید از دو استاد بپرسید و جواب را برآیند بگیرید بالطبع جواب مطمئن تری خواهید داشت.



شکل ۳-۴: استفاده از جنگل تصادفی .

۳-۲-۲ پیاده سازی

پیاده سازی جنگل تصادفی به کمک کتاب خانه ی قدرتمند Scikit Learn بسیار ساده است. تنها نکته ی شایان به ذکر است این است که برای جلوگیری از بزرگ شدن درخت و زیاد شدن تعداد برگ ها و طبیعتا کم شدن تعداد داده های هر برگ ، حداکثر عمق این درخت را بر روی ۱۰ تنظیم کردم.

```
[ ]  1  from sklearn.ensemble import RandomForestClassifier

[ ]  1  X = df_selected.values[:, :-1]
     2  X

[ ]  1  Y = df_selected.values[:, -1]
     2  Y

[ ]  1  clf = RandomForestClassifier(max_depth=10, random_state=0)
     2  clf.fit(X_train_enc, Y_train_enc)

[ ]  1  clf.predict(X_train_enc[0:10])

    array([0, 0, 0, 0, 0, 0, 0, 0, 1, 0])

[ ]  1  y_predicted = clf.predict(X_test_enc)
     2  y_predicted

    array([0, 0, 0, ..., 1, 1, 1])

[ ]  1  Y_test_enc

    array([0, 0, 1, ..., 0, 0, 1])

[ ]  1  import numpy as np
     2  np.sum(y_predicted == Y_test_enc) / len(y_predicted)

    0.7997665991032492
```

شکل ۳-۵: استفاده از جنگل تصادفی .

همان طور که در ۳-۵ مشاهده می کنید دقت این مدل بر روی داده های تست ۷۹ درصد بوده است.

فصل چهارم

مقایسه نتایج

ما در این گزارش به بررسی دو روش شبکه عصبی و جنگل تصادفی بر روی دیتاست مان پرداختیم. در جدول زیر سایر دقت هایی که بر روی این دیتاست به دست آمده را نیز آورده ام.

روش	دقت (Accuracy)
C4.5	84.46
C4.5 Auto	85.54
C4.5 rules	85.06
Voted ID3 (0.6)	84.36
1R	80.46
Nearest-neighbor (3)	79.65
T2	83.16
CN2	84.00
NBTree	85.9
** MLP	80.04
** Random Forest	79.97

جدول ۴-۱: الگوریتم های متفاوت و عملکردشان بر روی دیتاست Adult

دو روش ستاره دار در جدول روش هایی هستند که در این گزارش پیاده سازی شدند. به طور کلی می توان گفت علی رغم حذف ۹ فیچر، هر دو روش جنگل تصادفی و شبکه ی عصبی نتایج قابل قبولی را به نمایش گذاشته اند. هر چند در هر دو روش جزئیات بسیار زیادی وجود دارد که قابل دستکاری هستند مانند معماری شبکه عصبی و ... ولی در این دیتاست و با هاپیرپارامتر های فعلی هر دو تقریبا یک دقت را به نمایش گذاشتند. شاید تنها نکته ی قابل توجه این باشد که زمان فیت شدن جنگل تصادفی به داده ها بسیار کمتر از شبکه عصبی بود هر چند که هر دوی این زمان ها کمتر از ۵ دقیقه بودند.

منابع و مراجع

[1] Brownlee, Jason. 3 ways to encode categorical variables for deep learning. https://machinelearningmastery.com/how-to-prepare-categorical-data-for-deep-learning-in-python.

[2] Yiu, Tony. Understanding random forest. https://towardsdatascience.com/understanding-random-forest-58381e0602d2.