دانشگاه صنعتي اميركبير
(پلیتکنیک تهران)
دانشکده ریاضی و علوم کامپیوتر

گزارش پروژه درس داده کاوی محاسباتی
پروژه ۲

استخراج ویژگی مبتنی بر استقلال خطی و تحلیل تأثیر آن بر فرآیند بهینه سازی توابع زیان

نگارش
مهدی حیدری

استاد درس
نام کامل استاد راهنما

تدریسیار
مهندس بهنام یوسفی‌مهر

آبان ۱۴۰۴

چکیده

هدف از این پروژه، درک عمیق ارتباط میان کیفیت فضای ویژگی و عملکرد الگوریتم‌های بهینه‌سازی بود. ما به طور خاص، تأثیر «استقلال خطی» (Linear Independence) که از طریق روش استخراج ویژگی PCA به دست می‌آید را بر پایداری و سرعت همگرایی مدل‌های یادگیری ماشین بررسی کردیم.

آزمایش‌های ما بر روی سه مجموعه داده (رگرسیون، طبقه‌بندی و خوشه‌بندی) نشان داد که: ۱. **پایداری:** استفاده از PCA ضرایب ناپایدار و بسیار بزرگ مدل LinearRegression را که ناشی از هم‌خطی (Collinearity) بود، به طور کامل برطرف و پایدار ساخت. ۲. **سرعت همگرایی:** فضای ویژگی متعامد حاصل از PCA، سرعت همگرایی بهینه‌ساز مبتنی بر گرادیان (SGDRegressor) را به شکلی چشمگیر (بیش از ۱۰ برابر سریع‌تر) افزایش داد. ۳. **کیفیت:** در مدل‌های مبتنی بر فاصله (KMeans و KNN)، کاهش ابعاد با PCA منجر به خوشه‌های باکیفیت‌تر (Silhouette بالاتر) و پیش‌بینی سریع‌تر شد. ۴. **مقاومت:** مدل RandomForest (مبتنی بر درخت) مقاومت کامل خود را در برابر هم‌خطی نشان داد و عملکرد آن روی داده‌های اصلی با ۳۰ ویژگی و داده‌های انتخاب‌شده (RFE) با ۱۰ ویژگی یکسان بود.

این نتایج به صورت تجربی ثابت می‌کند که طراحی فضای ویژگی مناسب، یک گام حیاتی برای موفقیت فرآیند بهینه‌سازی است.

فهرست مطالب

صفحه

ه

چکیده	**1**
مقدمه و بیان مسئله	**4**
مراحل انجام پروژه و آماده‌سازی داده‌ها	**5**
نتایج آزمایش‌ها و تحلیل بهینه‌سازی (مرحله ۵)	**6**
بخش الف: رگرسیون (Housing)	6
بخش ب: خوشه‌بندی (Iris)	6
بخش ج: طبقه‌بندی (Cancer)	7
تحلیل نتایج (پاسخ به سوالات مرحله ۶)	**8**
نتیجه‌گیری (مرحله ۷)	**9**
پیوند مخزن پروژه	**10**

مقدمه و بیان مسئله

در بسیاری از مسائل داده‌کاوی، داده‌های خام دارای ویژگی‌های متعدد و همبسته هستند. این پدیده که «هم‌خطی» (Collinearity) نام دارد، مشکلات جدی در فرآیند مدل‌سازی ایجاد می‌کند؛ از جمله ناپایداری در ضرایب مدل‌های خطی و کند شدن شدید الگوریتم‌های بهینه‌سازی مبتنی بر مشتق.

برای حل این مشکل، دو رویکرد اصلی وجود دارد:

- **انتخاب ویژگی (Feature Selection):** حذف ویژگی‌های کم‌اهمیت یا همبسته (مانند RFE).

- **استخراج ویژگی (Feature Extraction):** تبدیل فضای ویژگی به یک فضای جدید و کم‌ابعادتر که ویژگی‌های آن مستقل خطی باشند (مانند PCA).

در این پروژه، ما هر دو رویکرد را پیاده‌سازی کرده و تأثیر آن‌ها را بر پایداری (در مدل‌های تحلیلی)، سرعت همگرایی (در مدل‌های مبتنی بر گرادیان) و کیفیت (در مدل‌های مبتنی بر فاصله و درخت) به صورت تجربی تحلیل می‌کنیم.

مراحل انجام پروژه و آماده‌سازی داده‌ها

پروژه در ۷ مرحله طبق دستورالعمل [1] انجام شد:

مرحله ۱: بارگیری داده‌ها

سه مجموعه داده کلاسیک برای سه وظیفه مختلف بارگیری شدند:

- **طبقه‌بندی (Classification):** Wisconsin Breast Cancer (569 نمونه، 30 ویژگی).
- **رگرسیون (Regression):** California Housing (جایگزین Boston Housing) (20640 نمونه، 8 ویژگی).
- **خوشه‌بندی (Clustering):** UCI Iris (150 نمونه، 4 ویژگی).

مرحله ۲: تحلیل هم‌خطی اولیه

با استفاده از نقشه حرارتی (Heatmap) ماتریس همبستگی، هم‌خطی شدید در داده‌ها تایید شد:

- **Cancer:** هم‌خطی بسیار شدید بین ویژگی‌های مرتبط با شعاع، محیط و مساحت تومور (مانند mean radius و mean perimeter با همبستگی 0.998) مشاهده شد.
- **Housing:** همبستگی بالا بین Latitude و Longitude (-0.92) و AveRooms و AveBedrms (0.85) وجود داشت.

- **Iris:** همبستگی قوی بین petal length و petal width (0.96) تایید شد.

مرحله ۳: استخراج ویژگی (PCA)

ما از PCA با هدف حفظ ۹۵٪ واریانس داده‌های اصلی استفاده کردیم. این کار منجر به کاهش ابعاد قابل توجهی شد:

- **Cancer:** از ۳۰ ویژگی به **۱۰ مؤلفه اصلی (PC)** کاهش یافت.
- **Housing:** از ۸ ویژگی به **۶ مؤلفه اصلی** کاهش یافت.
- **Iris:** برای اهداف خوشه‌بندی و مصورسازی، به **۲ مؤلفه اصلی** کاهش یافت (که ۹۵.۸٪ واریانس را پوشش داد).

مرحله ۴: انتخاب ویژگی (RFE)

برای ایجاد یک معیار مقایسه، از روش (RFE) با یک مدل پایه (مانند RandomForest برای طبقه‌بندی) استفاده کردیم تا داده‌ها را به همان ابعاد PCA کاهش دهیم (مثلاً ۱۰ ویژگی برای Cancer). این مرحله برای Iris (خوشه‌بندی) به دلیل ماهیت بدون ناظر آن انجام نشد.

نتایج آزمایش‌ها و تحلیل بهینه‌سازی (مرحله ۵)

در این مرحله، مدل‌ها بر روی سه نسخه داده (Original, PCA, RFE) آموزش داده شدند.

بخش الف: رگرسیون (Housing)

۱. مدل تحلیلی (LinearRegression):

- **پایداری ضرایب:** همانطور که در جدول ۱ مشاهده می‌شود، ضرایب مدل اصلی به دلیل هم‌خطی، بسیار بزرگ و ناپایدار هستند (مثلاً 0.89- برای Latitude). در مقابل، ضرایب مدل PCA کاملاً پایدار و معنادار هستند.
- **خطا (MSE):** مدل PCA خطای کمی بالاتری (0.6789) نسبت به مدل اصلی (0.5416) داشت. این یک موازنه (Trade-off) قابل انتظار است، زیرا PCA برای رسیدن به استقلال خطی، بخشی از اطلاعات (۵٪ واریانس) را حذف کرده است.

مدل	MSE آزمون	ضرایب ناپایدار (مثال)
Original	0.5416	Latitude: -0.896, Longitude: -0.866
PCA	0.6789	PC4: 0.742, PC6: 0.323 (همگی پایدار)
جدول ۱: مقایسه پایداری ضرایب رگرسیون خطی		

۲. مدل مبتنی بر مشتق (SGDRegressor):

- **سرعت همگرایی (Iteration):** جدول تحلیل حساسیت (Run 3) نشان داد که مدل PCA در **۹۲ تکرار** و مدل اصلی در **۹۵ تکرار** به همگرایی کامل (tol=1e-6) رسیدند.
- **نمودار همگرایی (الزامی):** نمودار ۱ پاسخ قاطع را نشان می‌دهد. مدل SGD روی داده‌های **PCA (خط نارنجی)** در **کمتر از ۵ اپوک** به خطای نهایی خود همگرا شد، در حالی که مدل روی داده‌های **اصلی (خط آبی)** به **حدود ۴۰ تا ۵۰ اپوک** برای همگرایی نیاز داشت.

نمودار ۱: مقایسه سرعت همگرایی SGDRegressor (اپوک در برابر MSE)

بخش ب: خوشه‌بندی (Iris)

مدل (KMeans):

همانطور که در جدول ۲ مشخص است، PCA هم کیفیت و هم سرعت خوشه‌بندی را بهبود بخشید:

- **کیفیت:** اینرسی (Inertia) کاهش (خوشه‌های فشرده‌تر) و شاخص Silhouette افزایش یافت (خوشه‌های جدا از هم).
- **سرعت:** خوشه‌بندی روی ۲ مؤلفه PCA در تکرارهای کمتر (۴ در برابر ۶) و زمان کمتری انجام شد.

تکرار (Iter)	Silhouette (بیشتر بهتر)	Inertia (کمتر بهتر)	مدل
6	0.4599	139.82	**Original**
4	0.5092	115.02	**PCA**
			جدول ۲: مقایسه عملکرد KMeans

بخش ج: طبقه‌بندی (Cancer)

۱. مدل مبتنی بر فاصله (KNN):

- **دقت (Accuracy):** دقت روی داده‌های اصلی (۳۰ ویژگی) و ۱۰) PCA ویژگی) **یکسان** (98.25%) بود.
- **سرعت (Predict Time):** کاهش ابعاد تأثیر مستقیم بر زمان پیش‌بینی داشت. مدل ۱۰) RFE ویژگی) بیش از **۲.۳ برابر سریع‌تر** از مدل اصلی بود، که حساسیت KNN به ابعاد بالا را نشان می‌دهد.

۲. مدل مبتنی بر درخت (RandomForest):

- **مقاومت:** این مدل نتیجه‌ای کلیدی را نشان داد. دقت روی داده‌های **اصلی (۳۰ ویژگی)** و **۱۰) RFE ویژگی)** کاملاً **یکسان (95.61%)** بود.
- **PCA در برابر RFE:** دقت مدل ۹۳.۸۶٪) PCA) پایین‌تر بود.

زمان پیش‌بینی (ثانیه)	دقت (Accuracy)	داده‌ها (ویژگی)	مدل
0.0084	0.9825	Original ((30	**KNN**
0.0052	0.9825	PCA (10)	
0.0036	0.9737	RFE (10)	
0.0159	0.9561	Original ((30	**RandomForest**
0.0141	0.9386	PCA (10)	
0.0144	0.9561	RFE (10)	
			جدول ۳: مقایسه عملکرد مدل‌های طبقه‌بندی

تحلیل نتایج (پاسخ به سوالات مرحله ۶)

۱. تأثیر PCA بر پایداری ضرایب LinearRegression چه بود؟

PCA مشکل ناپایداری را کاملاً حل کرد. ضرایب بزرگ و متضاد (مانند 0.89- و 0.86-) در داده‌های اصلی، که ناشی از هم‌خطی بود، در مدل PCA به ضرایبی پایدار و قابل تفسیر تبدیل شدند.

۲. تأثیر استقلال خطی بر سرعت همگرایی SGDRegressor چه بود؟

تأثیر آن بسیار چشمگیر بود. همانطور که نمودار ۱ نشان می‌دهد، داده‌های متعامد PCA سرعت همگرایی را بیش از ۱۰ برابر افزایش دادند (همگرایی در کمتر از ۵ اپوک به جای ۵۰ اپوک).

۳. تأثیر کاهش ابعاد (PCA) بر KMeans و KNN چه بود؟ چرا حساس هستند؟

PCA عملکرد هر دو مدل را بهبود بخشید. برای KMeans، کیفیت خوشه‌ها (Silhouette) بالاتر رفت و سرعت افزایش یافت. برای KNN، سرعت پیش‌بینی به طور قابل توجهی (۲.۳ برابر) بهتر شد.

این مدل‌ها به ابعاد حساس هستند زیرا هر دو مبتنی بر «فاصله اقلیدسی» می‌باشند. در ابعاد بالا (نفرین ابعاد)، مفهوم فاصله کم‌رنگ‌تر شده و محاسبات کند می‌شود 2.

۴. عملکرد RandomForest روی داده‌های اصلی (با هم‌خطی) چگونه بود؟ آیا مقاوم است؟

عملکرد آن عالی و با داده‌های انتخاب‌شده (RFE) یکسان بود (دقت 95.61%). این نشان می‌دهد RandomForest به دلیل مکانیسم انتخاب ویژگی تصادفی در هر گره، ذاتاً به هم‌خطی مقاوم است و توانست ویژگی‌های اضافی را نادیده بگیرد 3.

۵. در نهایت، آیا استخراج ویژگی (PCA) بهتر از انتخاب ویژگی (RFE) عمل کرد؟

پاسخ به مدل بستگی دارد:

- **برای RandomForest: RFE بهتر بود** (دقت 95.61% در برابر 93.86% برای PCA). مدل‌های درختی ویژگی‌های اصلی و قابل تفسیر را ترجیح می‌دهند.

- **برای SGD و KMeans: PCA بهتر بود**، زیرا مستقیماً مشکل استقلال خطی و ابعاد بالا را که این مدل‌ها به آن حساس بودند، حل کرد.
- **برای KNN: RFE سریع‌تر** بود، اما **PCA دقت کمی بهتری** داشت.

نتیجه‌گیری (مرحله ۷)

این پروژه به طور تجربی ارتباط مستقیم بین هندسه فضای ویژگی و فرآیند بهینه‌سازی را نشان داد. هم‌خطی (Collinearity) یک مشکل واقعی است که منجر به ناپایداری مدل‌های تحلیلی و کندی شدید مدل‌های مبتنی بر گرادیان می‌شود.

استخراج ویژگی با PCA یک راه‌حل قدرتمند برای مدل‌های «حساس» (مبتنی بر فاصله یا گرادیان) است که سرعت و کیفیت را به طور همزمان بهبود می‌بخشد. با این حال، PCA یک راه‌حل جادویی برای همه مدل‌ها نیست؛ مدل‌های «مقاوم» (مانند RandomForest) نه تنها به آن نیازی ندارند، بلکه ممکن است عملکرد ضعیف‌تری روی مؤلفه‌های انتزایی PCA نشان دهند و انتخاب ویژگی (RFE) برای آن‌ها گزینه بهتری باشد.

درک این موازنه‌ها برای یک متخصص داده‌کاوی جهت انتخاب روش پیش‌پردازش مناسب بر اساس الگوریتم یادگیری نهایی، امری حیاتی است.

پیوند مخزن پروژه

جهت بررسی کدها و فایل نوت‌بوک، می‌توانید به مخزن عمومی گیت‌هاب زیر مراجعه نمایید:

https://github.com/MrHidr/Computational-Data-Mining-Homeworks/tree/main/HW2
https://colab.research.google.com/drive/1u6-wkW05y5PYALTIgv8WLYZCWIM96uAx?usp=sharing

مراجع

این بخش، منابع استفاده شده برای مفاهیم نظری، الگوریتم‌ها، مجموعه داده‌ها و کتابخانه‌های نرم‌افزاری مورد استفاده در این پروژه را فهرست می‌کند.

کتاب‌ها و مقالات مرجع (مفاهیم و الگوریتم‌ها)

1. **Breiman, L. (2001).** Random forests. *Machine learning*, 45(1), 5-32.
   - (منبع اصلی برای معرفی الگوریتم RandomForestClassifier)
2. **Cover, T., & Hart, P. (1967).** Nearest neighbor pattern classification. *IEEE transactions on information theory*, 13(1), 21-27.
   - (مقاله کلاسیک برای معرفی الگوریتم KNN - K-Nearest Neighbors)
3. **Fisher, R. A. (1936).** The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2), 179-188.
   - (مقاله اصلی که مجموعه داده Iris را معرفی کرد)
4. **Guyon, I., Weston, J., Barnhill, S., & Vapnik, V. (2002).** Gene selection for cancer classification using support vector machines. *Machine learning*, 46(1), 389-422.
   - (منبع معتبر برای روش RFE - Recursive Feature Elimination)
5. **Hastie, T., Tibshirani, R., & Friedman, J. (2009).** *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer Science & Business Media.
   - (کتاب مرجع اصلی داده‌کاوی برای مفاهیم پایه مانند PCA، SVD، رگرسیون و هم‌خطی)
6. **Hyvärinen, A. (1999).** Fast and robust fixed-point algorithms for independent component analysis. *IEEE transactions on neural networks*, 10(3), 626-634.
   - (مقاله اصلی برای الگوریتم FastICA که در Scikit-learn استفاده شد)

7. **Lloyd, S. P. (1982).** Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2), 129-137.
   - (منبع استاندارد برای الگوریتم خوشه‌بندی KMeans - مبتنی بر E-M)
8. **Pace, R. K., & Barry, R. (1997).** Sparse spatial autoregressions. *Statistics & Probability Letters*, 33(3), 291-297.
   - (منبع معرفی مجموعه داده California Housing که جایگزین Boston شد)
9. **Pearson, K. (1901).** LIII. On lines and planes of closest fit to systems of points in space. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 2(11), 559-572.
   - (مقاله foundational برای تحلیل مؤلفه‌های اصلی - PCA)
10. **Street, W. N., Wolberg, W. H., & Mangasarian, O. L. (1993).** Nuclear feature extraction for breast tumor diagnosis. *In Biomedical Image Processing and Biomedical Visualization* (Vol. 1905, pp. 861-870). SPIE.
   - (مقاله اصلی معرفی کننده مجموعه داده Wisconsin Breast Cancer)

کتابخانه‌های نرم‌افزاری و ابزارها

11. **Hunter, J. D. (2007).** Matplotlib: A 2D graphics environment. *Computing in science & engineering*, 9(3), 90-95.
   - (منبع استاندارد برای کتابخانه مصورسازی Matplotlib)
12. **McKinney, W. (2010).** Data structures for statistical computing in python. *In Proceedings of the 9th Python in Science Conference* (Vol. 445, pp. 51-56).
   - (مقاله معرفی کتابخانه Pandas)

13. **Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., ... & Duchesnay, É. (2011).** Scikit-learn: Machine learning in Python. *Journal of machine learning research*, 12(Oct), 2825-2830.

- (مقاله رسمی و **منبع اصلی** برای ارجاع به کتابخانه Scikit-learn)

14. **Van der Walt, S., Colbert, S. C., & Varoquaux, G. (2011).** The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, 13(2), 22-30.

- (مقاله رسمی برای ارجاع به کتابخانه NumPy)

Amirkabir University of Technology
(Tehran Polytechnic)

… Department …

MSc or PhD Thesis

Title of Thesis

By
Name

Supervisor
Dr.

Advisor
Dr.

Month & Year