دوره

علم داده، یادگیری ماشین و هوش مصنوعی

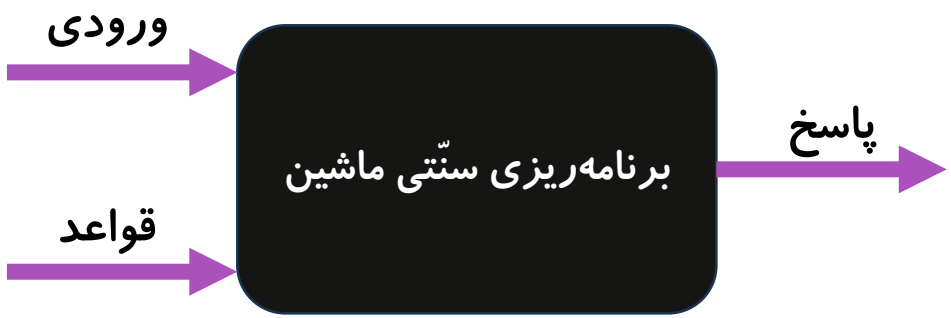
جلسه نهم:

مقدمه‌ای بر مفاهیم یادگیری ماشین

ارائه‌کننده:

دکتر فرزاد مینویی

برنامه‌ریزی سنّتی ماشین در برابر یادگیری ماشین



یادگیری ماشین برای حل چه مسائلی مناسب است؟

- مسائلی که راهکار فعلی نیازمند قواعد بسیار زیاد و بهبود دستی آن است.
- مسائل پیچیده‌ای که برنامه‌ریزی سنّتی ماشین راهکار مناسبی برای آن ندارد.
- محیط پویا که در آن یادگیری ماشین خود را می‌تواند بر اساس داده‌های جدید به‌روز نگه دارد.
- گرفتن شهود از حجم بالایی از داده‌ها در مسائل پیچیده

مثال‌هایی از کاربردهای یادگیری ماشین

- پیش‌بینی سطح دستمزد نیروی انسانی براساس تجربه، سطح تحصیلات، رشته تحصیلی و ...
- شناسایی محتواهای توهین‌آمیز در بخش نظرات یک وبسایت خبری
- پیش‌بینی اینکه آیا یک غده سرطانی خوشخیم است یا بدخیم
- تقسیم‌بندی مشتریان یک شرکت بر اساس رفتار خرید
- پیش‌بینی درآمد ماه‌های آتی یک شرکت بر اساس شاخص‌های عملکردی
- پیش‌بینی قیمت خودروهای دست دوم براساس کیلومتر کارکرد، عمر خودرو، مدل خودرو و ...

یادگیری ماشین چیست؟

Y	X_1	X_2	…	X_m
24000	10	17000	…	1
13000	60	98000	…	0
13500	75	89000	…	0
. . .	. . .	. . .	. . .	. . .
16500	1	1000	…	1

- مثال پیش‌بینی قیمت خودروهای دست دوم

$$Y = f(X) + \varepsilon$$

متغیر پاسخ
(Response Variable)

ویژگی‌ها
(Features)

خطای تصادفی
(Random Error)

$$X = (X_1, X_2 , \ldots, X_m)$$

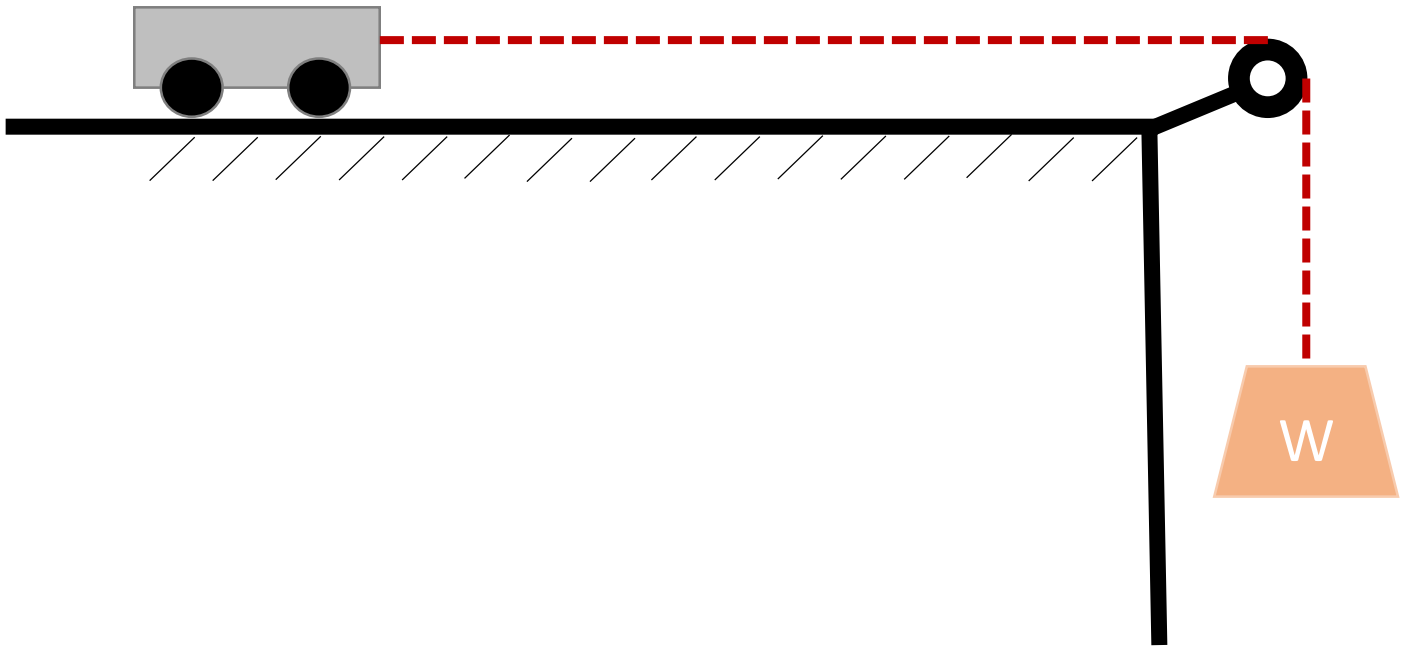
یادگیرنده (Learner) تلاش می‌کند تا f را برآورد کند.

رویکرد قطعی‌نگر
(Deterministic)

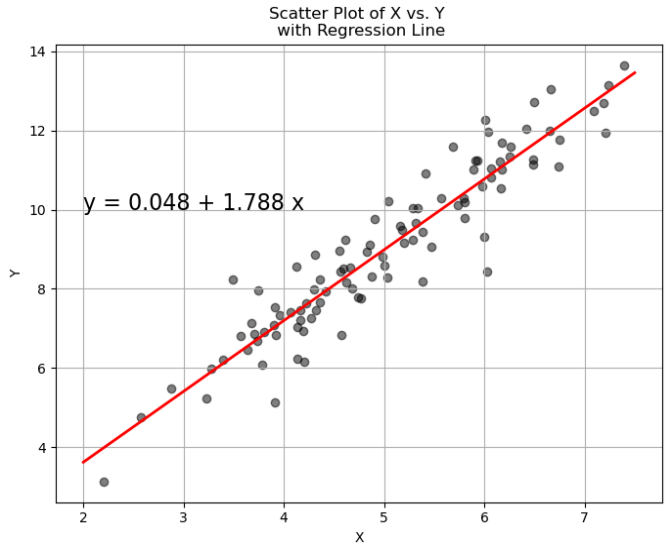
$$a = \frac{1}{m} F$$

رویکرد تصادفی
(Stochastic)

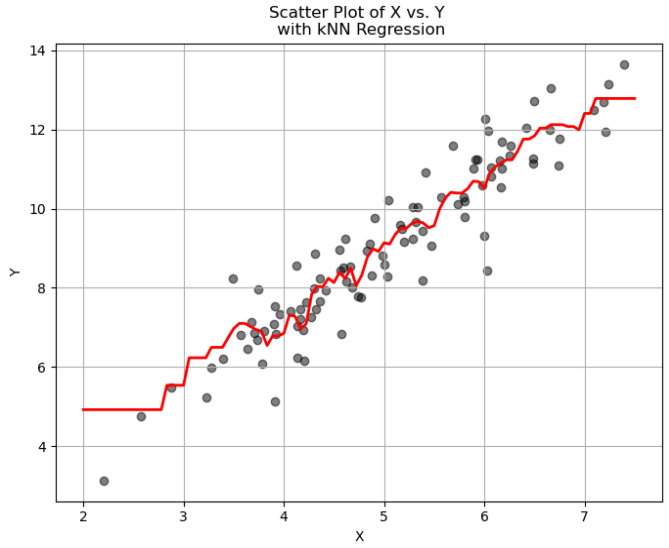
$$a = \frac{1}{m} F + \varepsilon$$



رویکردها به برآورد f



- رویکردهای پارامتری (Parametric Methods)
  - مانند الگوریتم رگرسیون خطی



- رویکردهای ناپارامتری (Non-parametric Methods)
  - مانند الگوریتم kNN

انواع مسائل یادگیری ماشین

- یادگیری نظارت‌شده (Supervised Learning)
- یادگیری نظارت‌نشده (Unsupervised Learning)
- یادگیری تقویتی (Reinforcement Learning)

یادگیری نظارت‌شده

- یادگیری تحت نظارت متغیر پاسخ صورت می‌گیرد.
- هدف اصلی پیش‌بینی است.

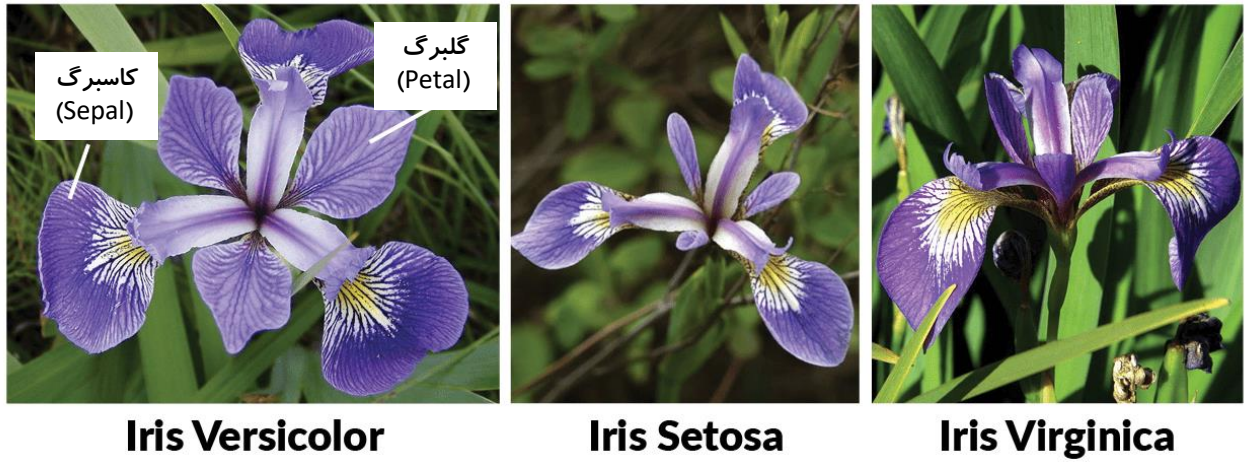
Y	X_1	X_2	...	X_m

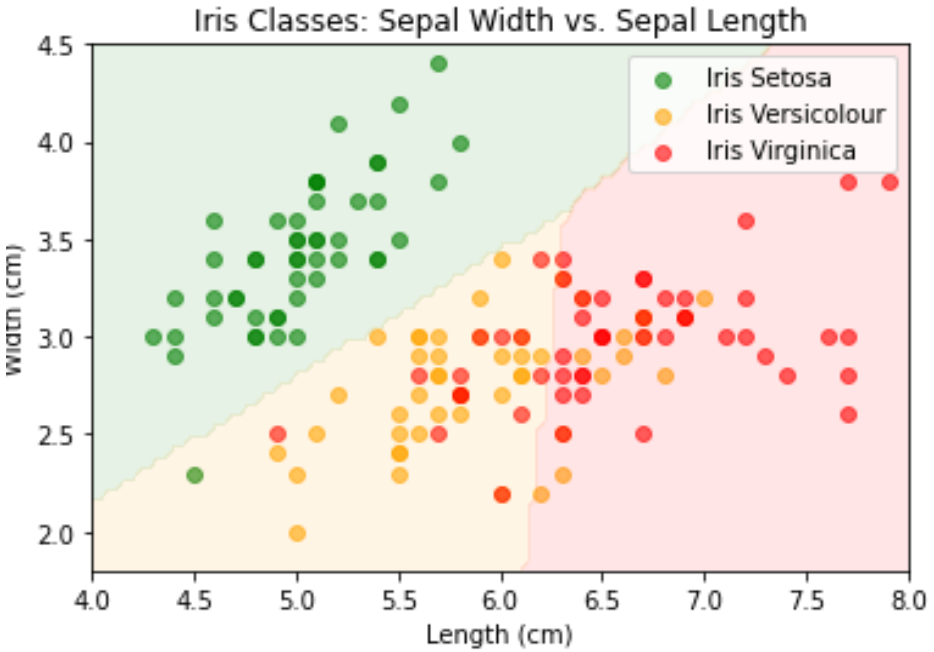
یادگیری نظارت‌شده

مسائل رگرسیون (Regression Problems)

یادگیری نظارت‌شده

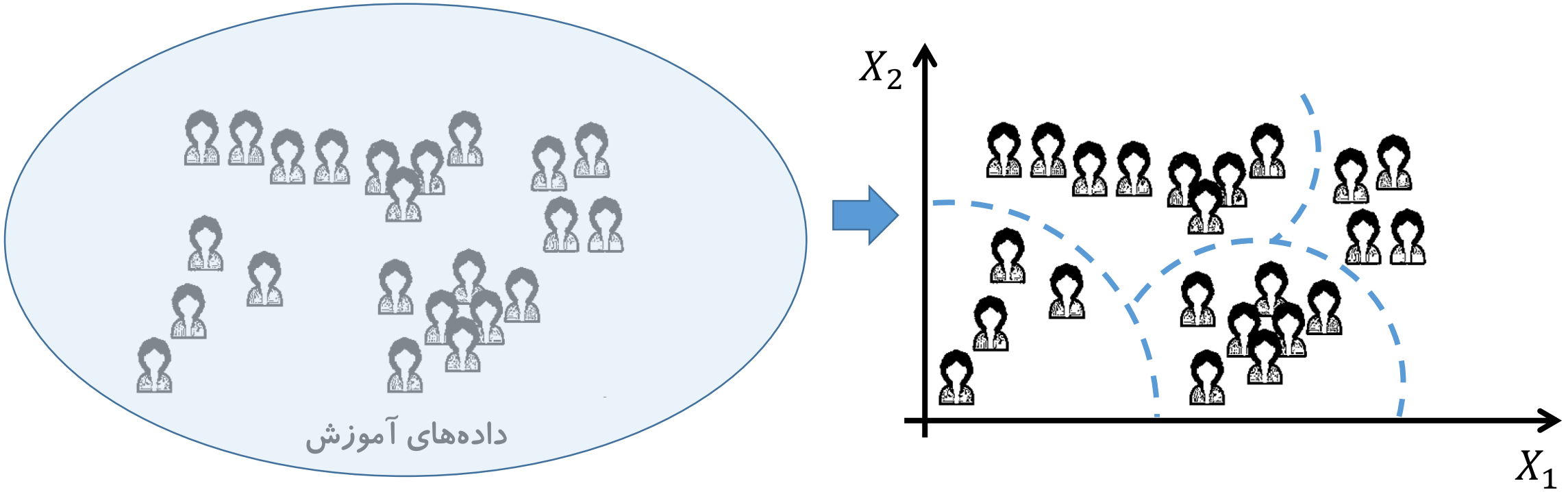
مدل‌های دسته‌بندی (Classification)





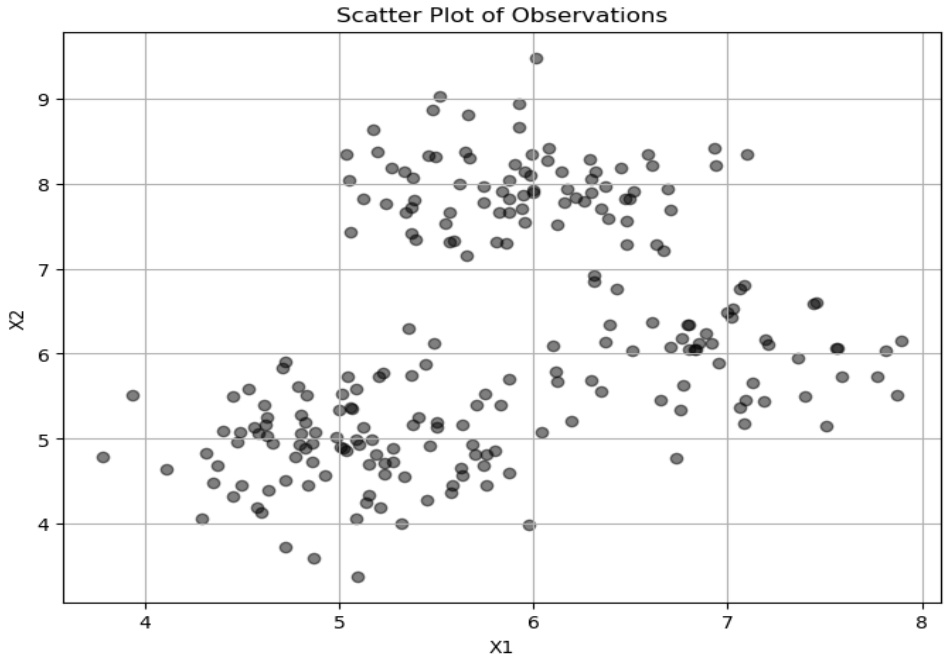
یادگیری نظارت‌نشده

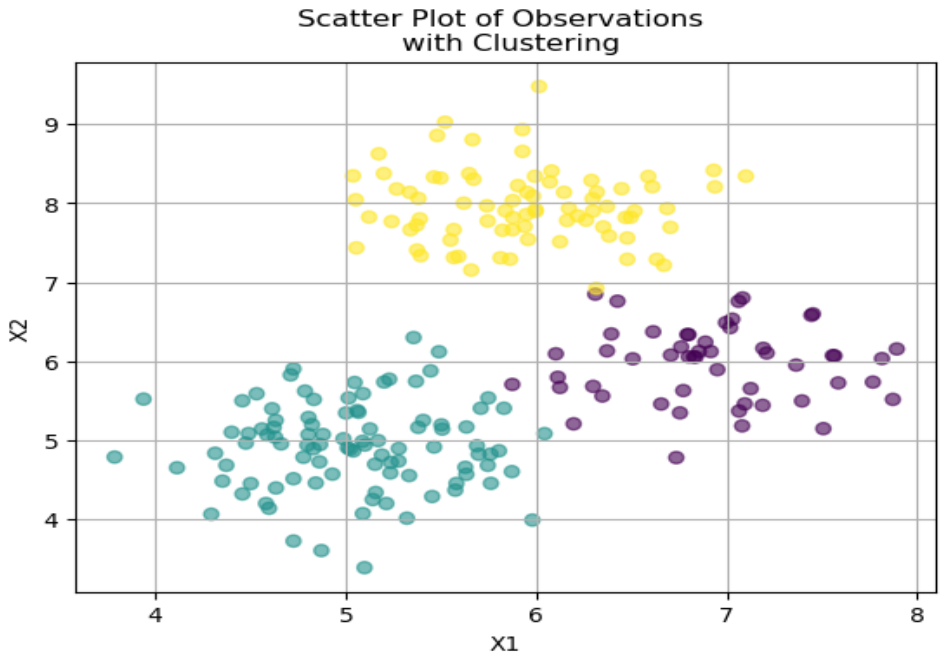
- یادگیری تحت نظارت متغیر پاسخ صورت نمی‌گیرد.
- هدف پیداکردن الگوهایی برای توصیف بهتر داده‌ها و گروه‌بندی آنها است.



یادگیری نظارت‌نشده

خوشه‌بندی (Clustering)

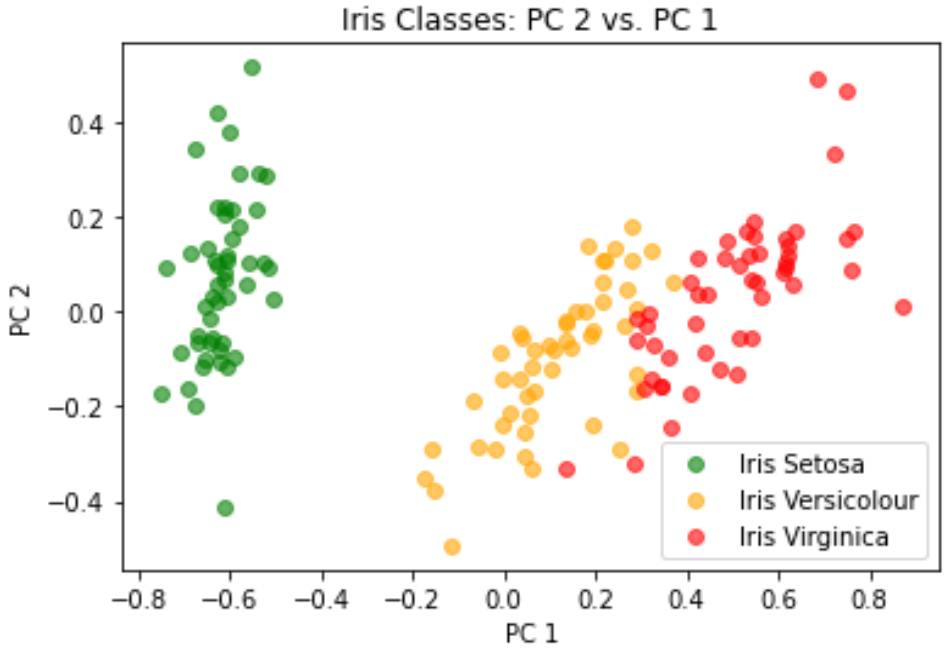




یادگیری نظارت‌نشده

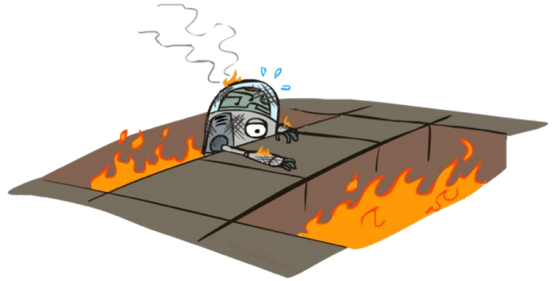
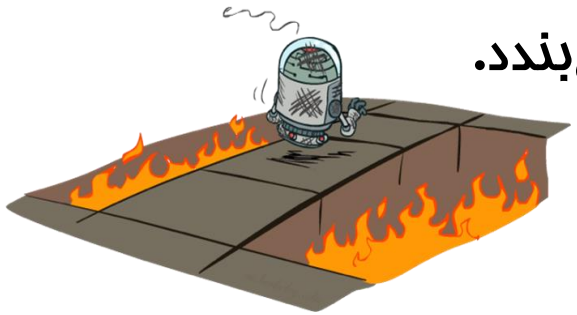
- کاهش بعد (Dimension Reduction)

	sepal length (cm)	sepal width (cm)	petal length (cm)	petal width (cm)
0	5.1	3.5	1.4	0.2
1	4.9	3.0	1.4	0.2
2	4.7	3.2	1.3	0.2
3	4.6	3.1	1.5	0.2
4	5.0	3.6	1.4	0.2



یادگیری تقویت‌شده

- یادگیرنده در اینجا عامل (Agent) نامیده می‌شود که می‌تواند محیط را مشاهده کند و عملی (Action) از خود بروز دهد و در عوض پاداش (Reward) دریافت کند.
- هدف آن است تا عامل از طریق تعامل با محیط یاد بگیرد چه سیاستی (Policy) بهینه است تا بتواند در طول زمان پاداش خود را بیشینه کند.
- منظور از سیاست مجموعه اقداماتی است که عامل با توجه به شرایط محیط (State) بکار می‌بندد.

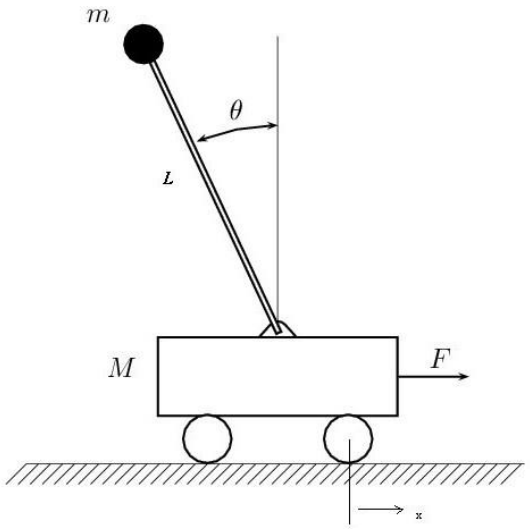




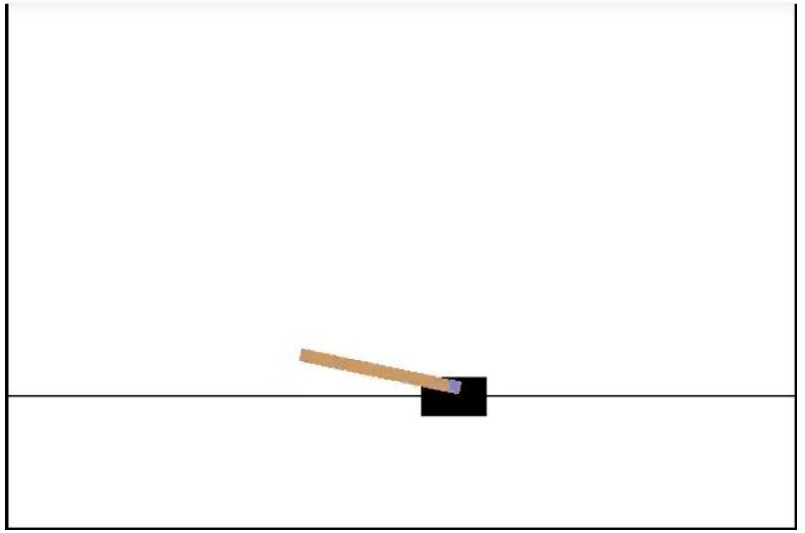
یادگیری تقویت‌شده

- بکارگیری یادگیری تقویت‌شده تحت شرایط زیر مناسب است:
  - تنها راه جمع‌آوری داده، تعامل با محیط است.
  - مدل محیط مشخص است، ولی راه حل تحلیلی وجود ندارد.
  - شبیه‌سازی محیط شدنی است.

یادگیری تقویت‌شده

- چطور به آونگ وارون یاد دهیم که خود را متعادل نگه دارد؟

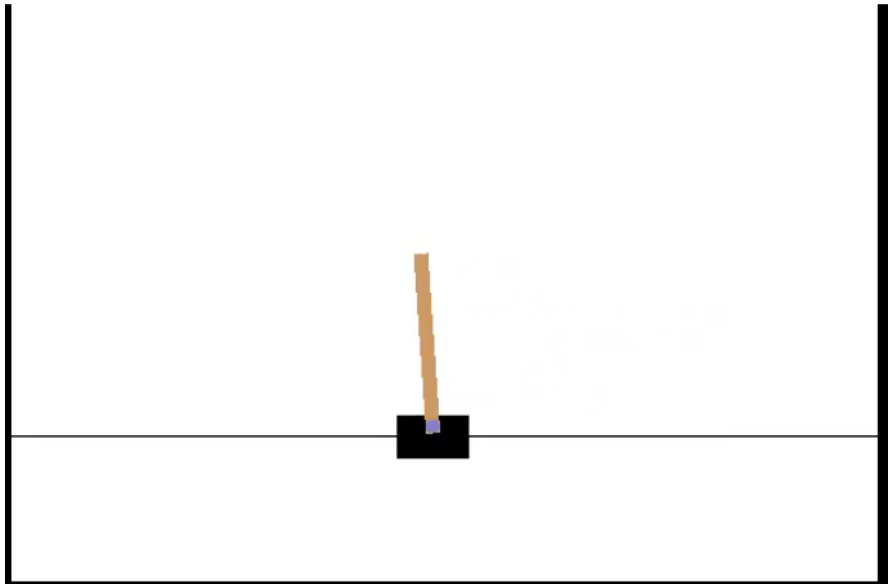




https://www.aparat.com/v/ZMsbu

یادگیری تقویت‌شده

- مسئله آونگ وارونه: آموزش با Q-Learning



https://www.aparat.com/v/hNkLA

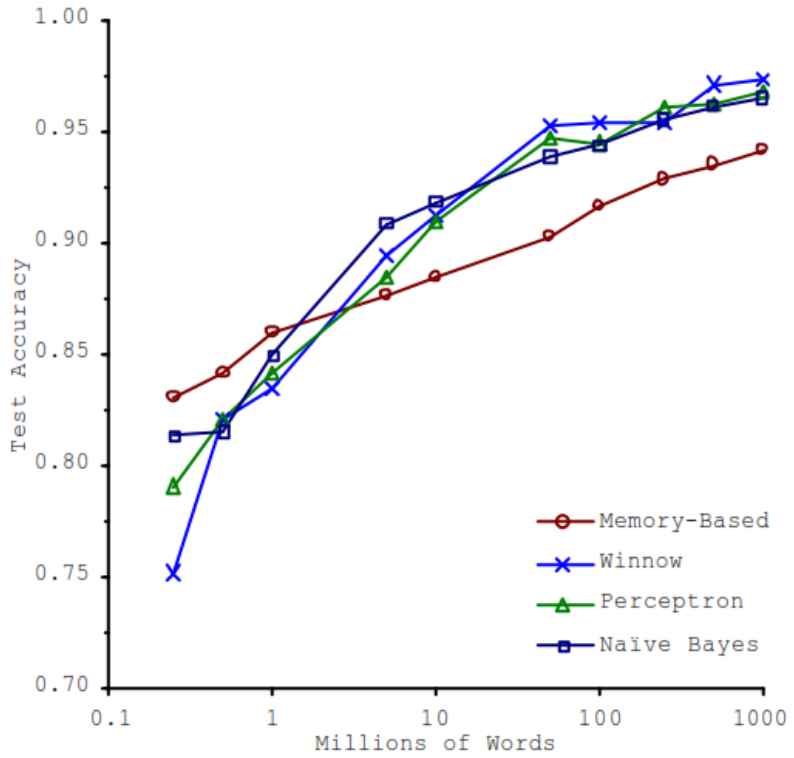
تمرین

- حدس بزنید برای هر مسئله، کدام رویکرد یادگیری ماشین مناسب‌تر است؟

مثال ۱	پیدا کردن گروه‌هایی از کاربران که رفتار مشابهی در جستجوی صفحات یک وبسایت سرگرمی دارند
مثال ۲	پیش‌بینی قیمت خانه براساس سال ساخت، متراژ، مکان جغرافیایی و ...
مثال ۳	آماده‌سازی یک دیتاست با ۸۰۰ متغیر برای تحلیل رگرسیون
مثال ۴	یافتن سلول‌های سرطانی مشابه بر اساس شاخص‌های ژنتیکی آنان
مثال ۵	پیش‌بینی جهت حرکت (صعودی/نزولی) شاخص بازار سهام در روز آینده
مثال ۶	محاسبه احتمال آنکه یک تراکنش بانکی متقلبانه است
مثال ۷	رباتی که می‌تواند بازی Pacman انجام دهد

چالش‌های یادگیری ماشین

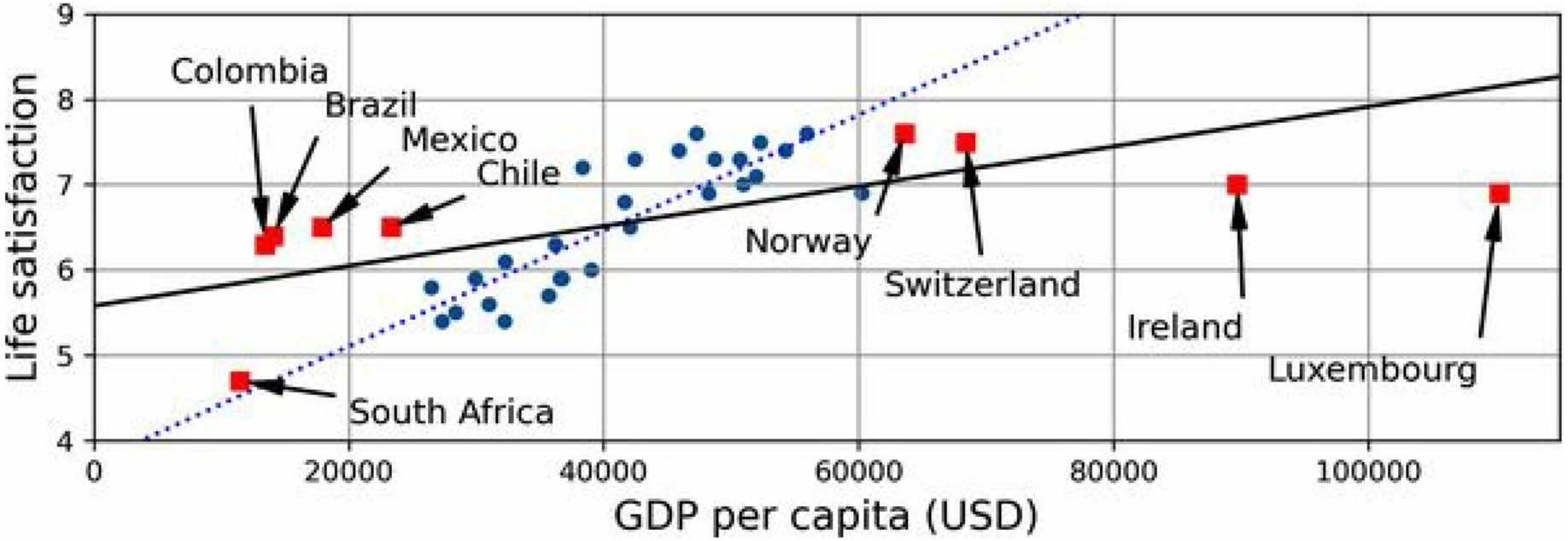
- ناکافی بودن داده آموزش



Banko, Michele, and Eric Brill. "Scaling to very very large corpora for natural language disambiguation." In *Proceedings of the 39th annual meeting of the Association for Computational Linguistics*, pp. 26-33. 2001.

چالش‌های یادگیری ماشین

- نماینده نبودن نمونه آموزش از جامعه



چالش‌های یادگیری ماشین

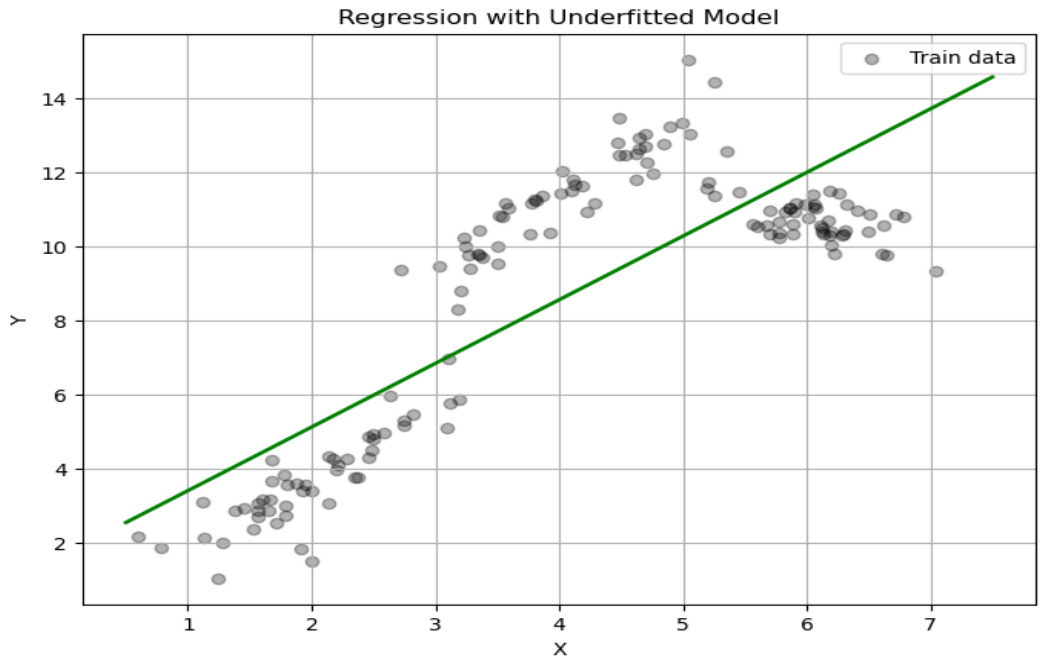
- کیفیت پایین داده‌ها
  - داده‌های گمشده (Missing values)
  - داده‌های پرت (Outliers)

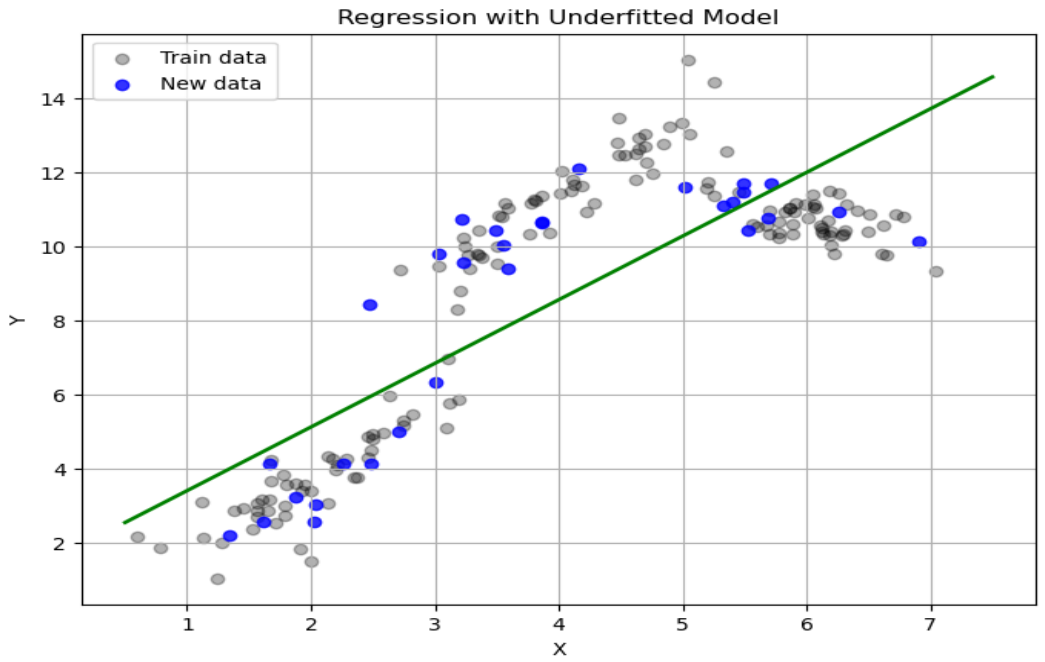
چالش‌های یادگیری ماشین

- ویژگی‌های نامرتبط
  - انتخاب ویژگی‌ها (Feature Selection)
  - ساخت ویژگی‌های جدید (Dimension Reduction)
  - جمع‌آوری داده‌های جدید با ویژگی‌های جدید

چالش‌های یادگیری ماشین

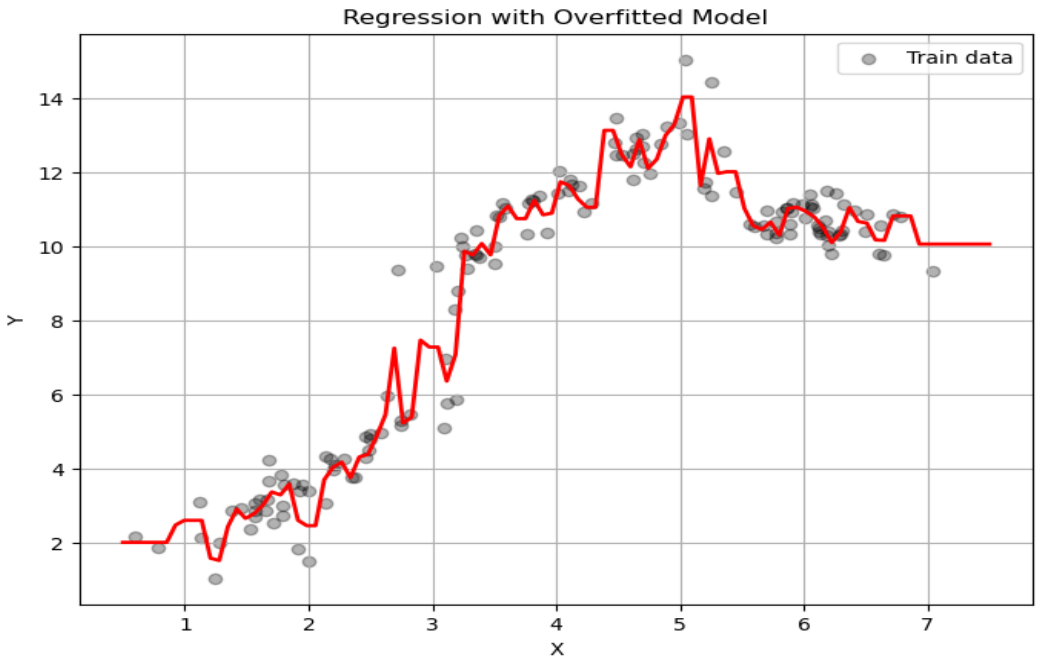
- خطای سوگیری (Bias) در مدل‌سازی
  - عدم توانایی الگوریتم یادگیری ماشین در برآورد صحیح رابطه (f) سوگیری نامیده می‌شود.

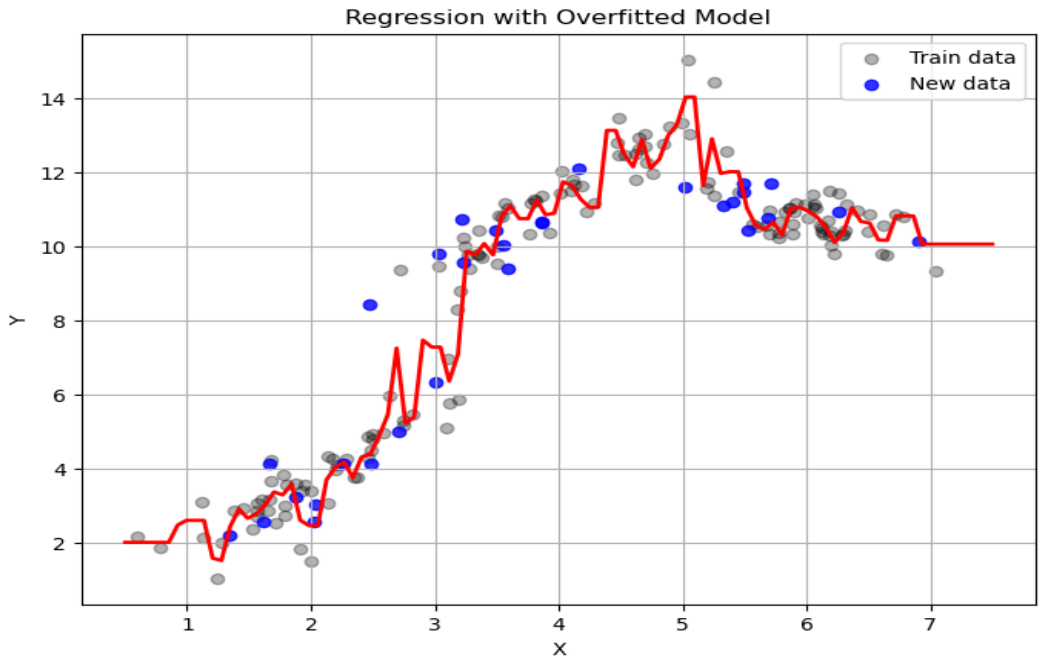




چالش‌های یادگیری ماشین

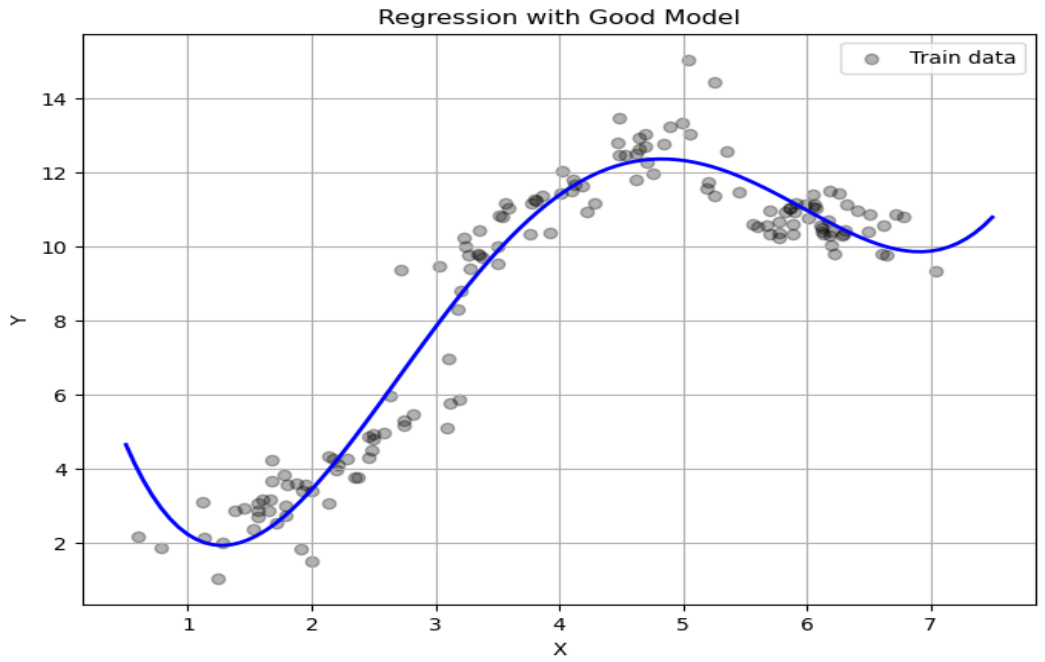
- خطای واریانس (Variance) در مدل‌سازی
  - خطای واریانس ناشی از بیش‌برازش مدل یادگیری ماشین روی داده‌های آموزش است.

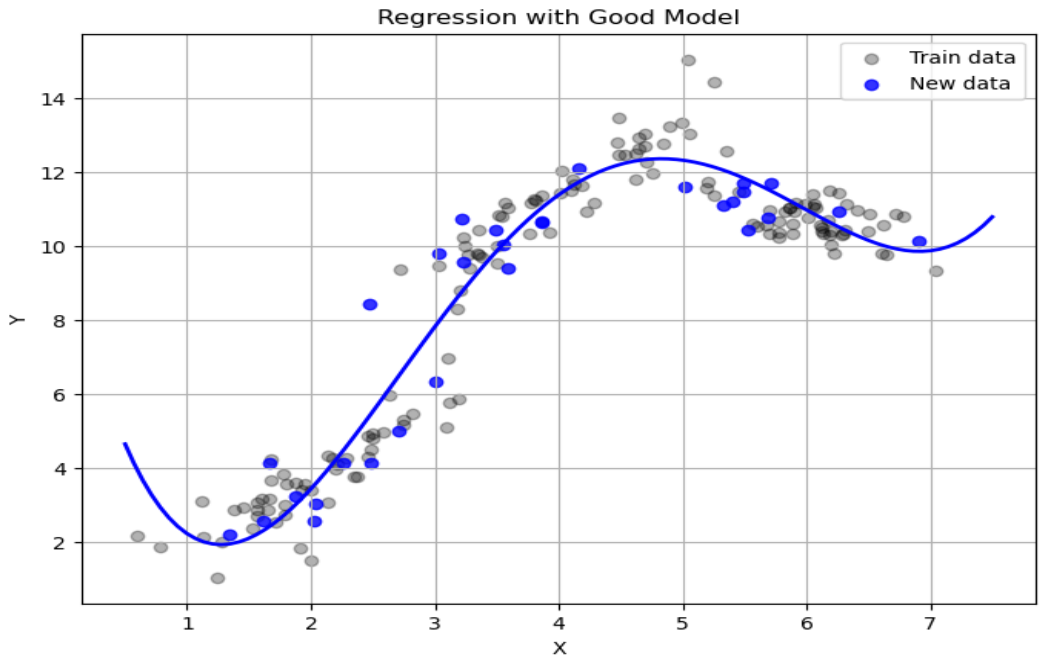




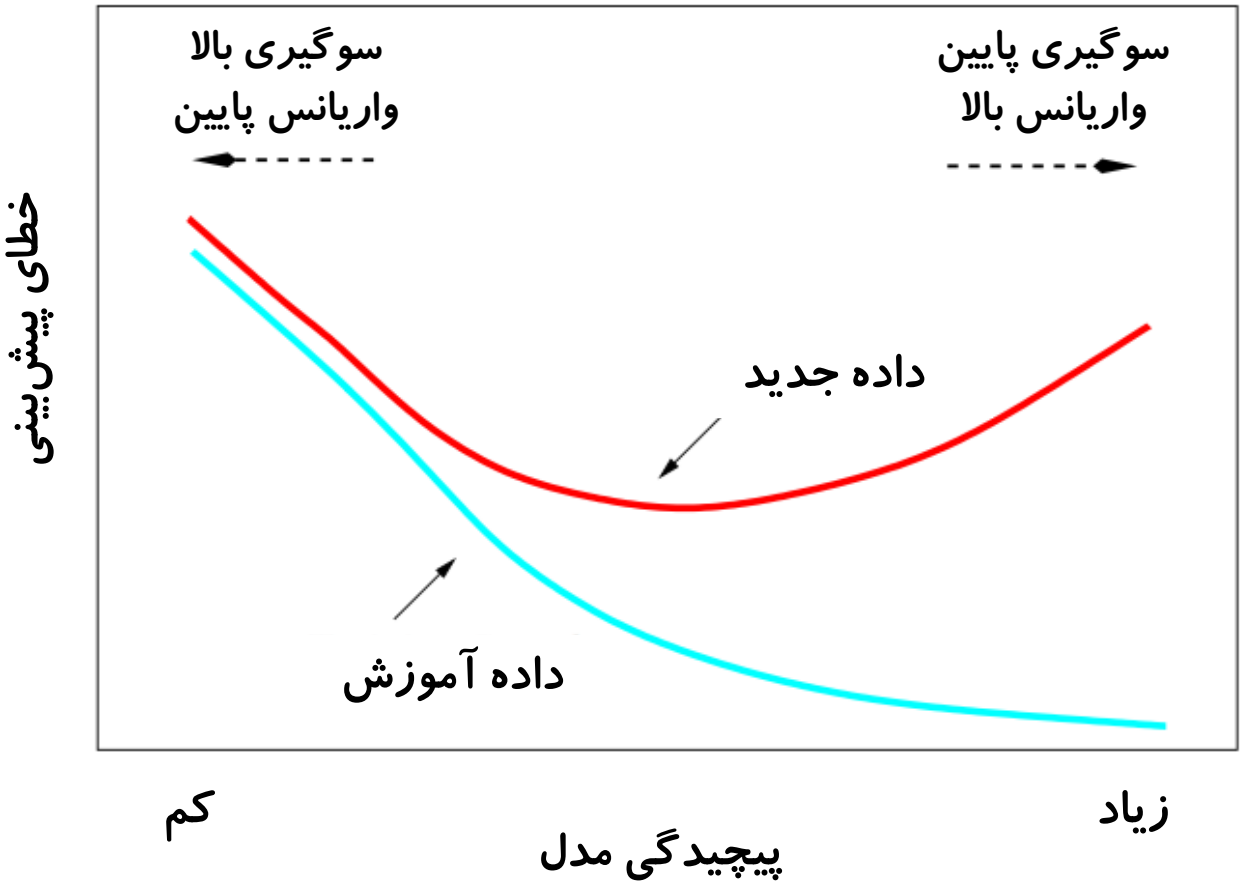
چالش‌های یادگیری ماشین

- تعادل بین سوگیری و واریانس





مفهوم تعادل بین سوگیری و واریانس



جداسازی داده جهت آموزش و آزمایش

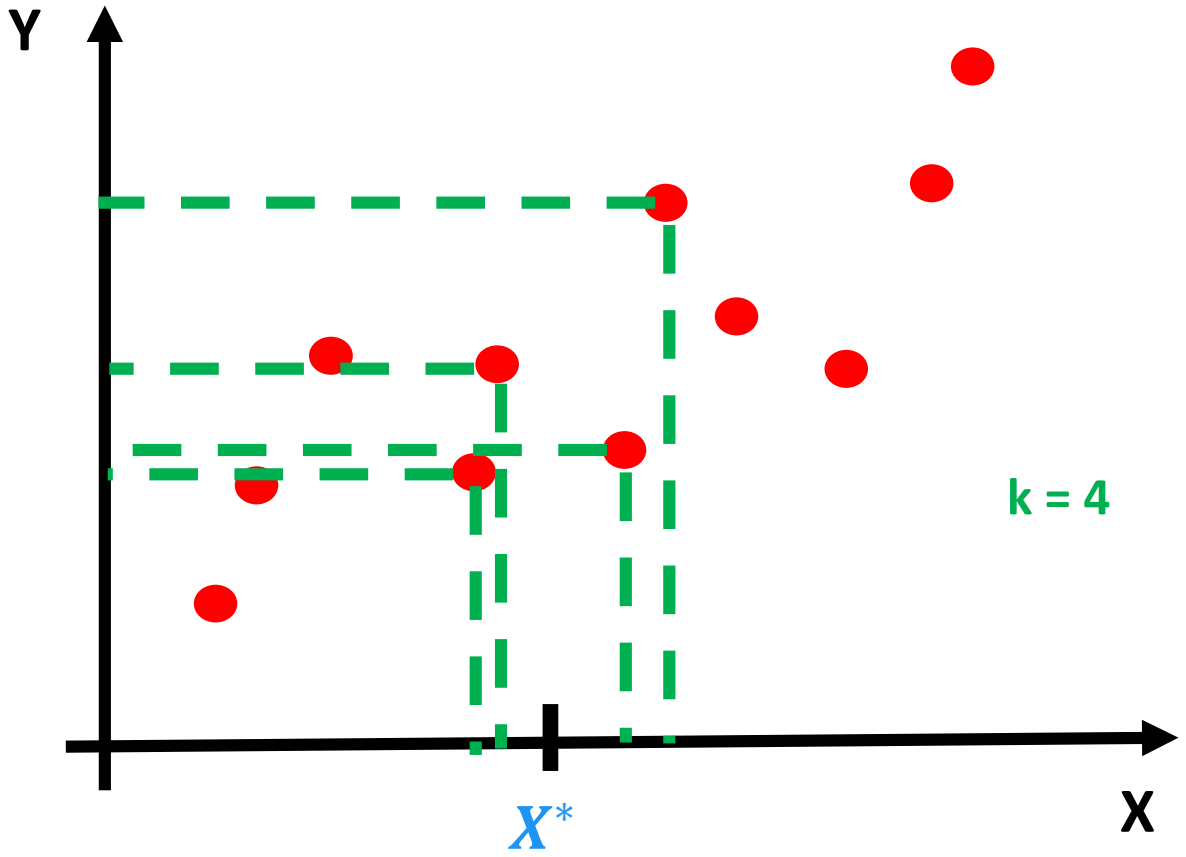
- مسئله تعمیم‌پذیری مدل (Generalizability)
- جداسازی داده (Data Splitting)
  - داده‌های آموزش (Train)
  - داده‌های آزمایش (Test)
  - قانون سرانگشتی:

60% (train) - 40% (test), 70% - 30%, or 80% - 20%

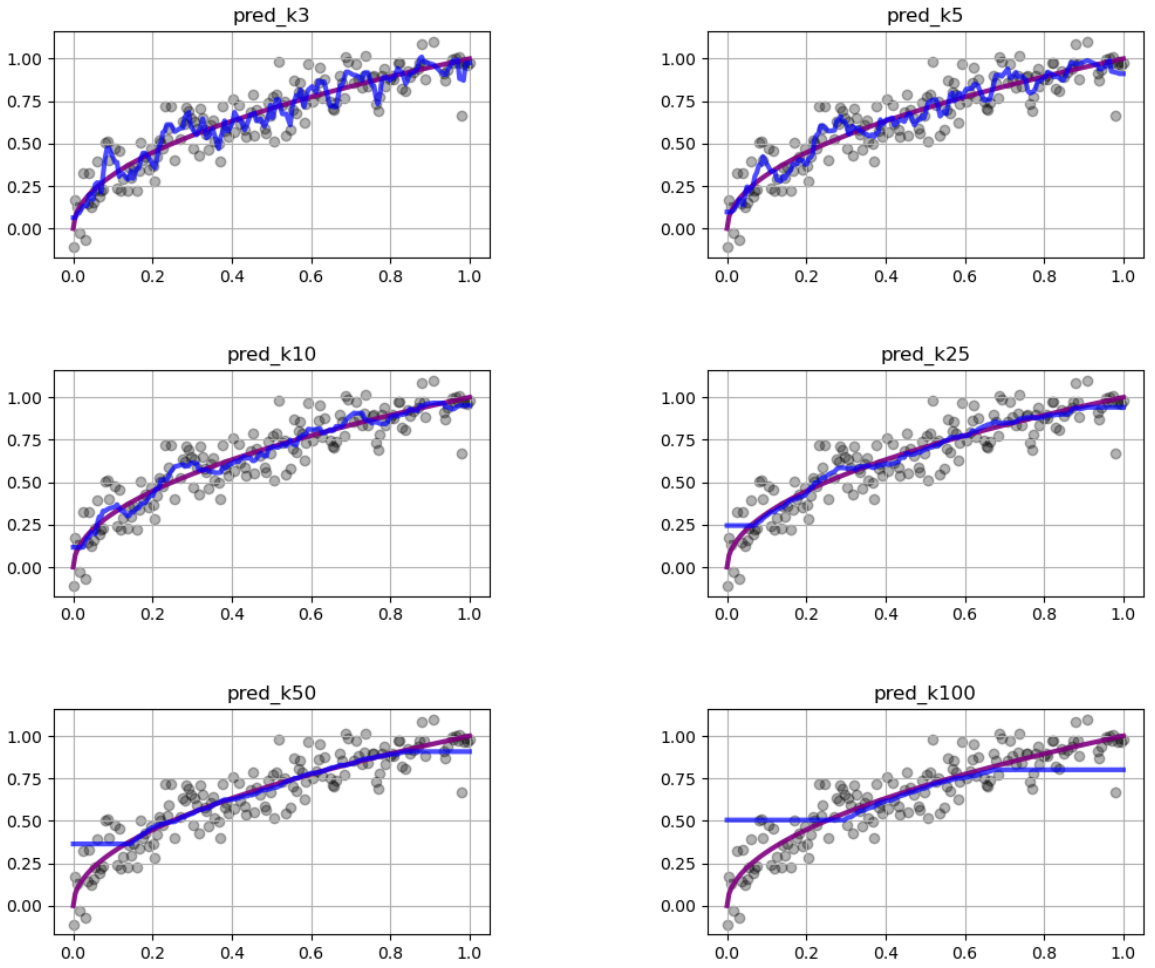
- لزوماً حجم داده بزرگ (۱۰۰هزار <) برای آموزش مدل منجر به بهبود چشمگیر مدل نمی‌شود. ممکن است سرعت محاسبات را کاهش دهد.
- اگر تعداد ویژگی‌ها > تعداد داده‌ها باشد، حجم داده بیشتر برای آموزش مدل بهتر است.

k-Nearest Neighbors Regression (k-NN Regression)

$$\hat{y} = \frac{1}{k} \sum_{x_i \epsilon N_k(x)} y_i$$



تعادل بین سوگیری و واریانس در الگوریتم kNN



ارزیابی مدل رگرسیون

MSE: Mean Squared Error → min

$$MSE = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}$$

RMSE: Root Mean Squared Error → min

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{n}}$$

MAE: Mean Absolute Error → min

$$MAE = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{n}$$

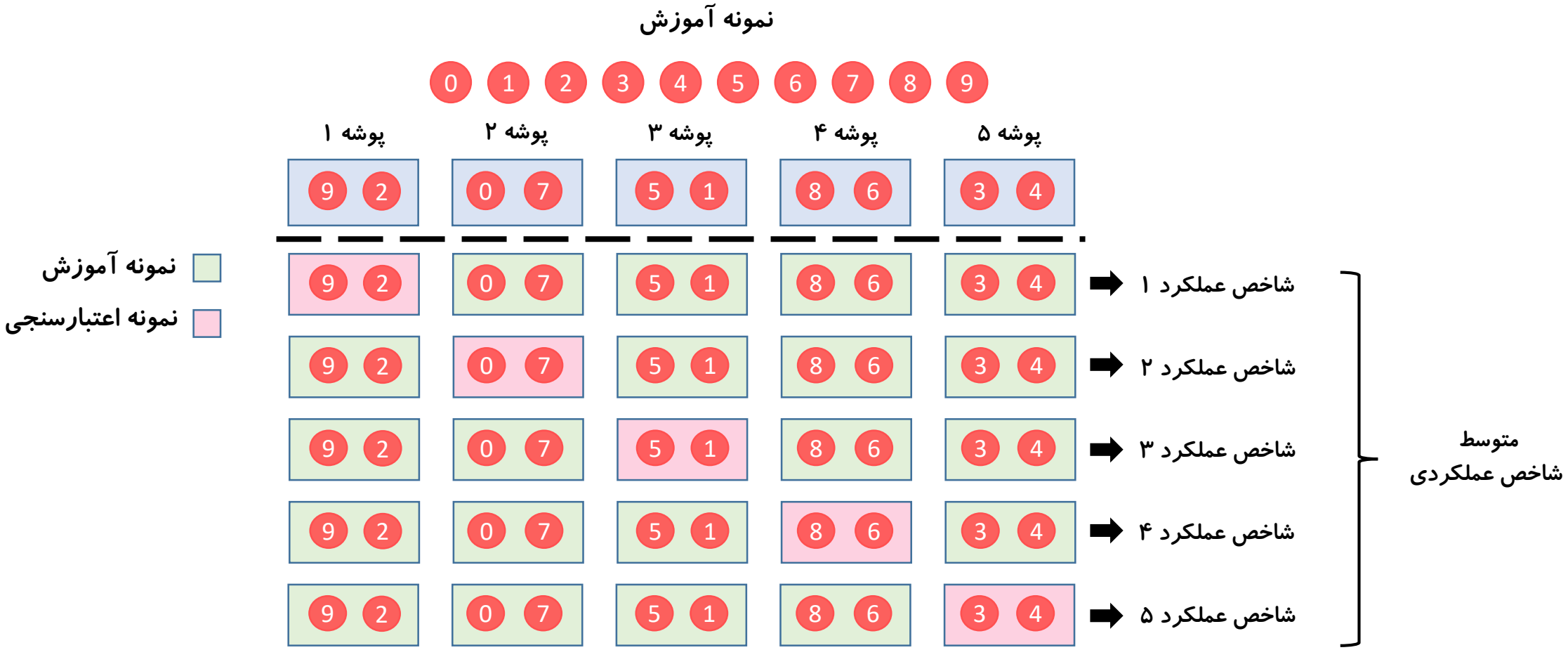
MAPE: Mean Absolute Percentage Error → min

$$MAPE = \frac{\sum_{i=1}^{n} |\frac{y_i - \hat{y}_i}{y_i}|}{n}$$

تنظیم فراپارمترهای مدل (Hyper-parameter Tuning)

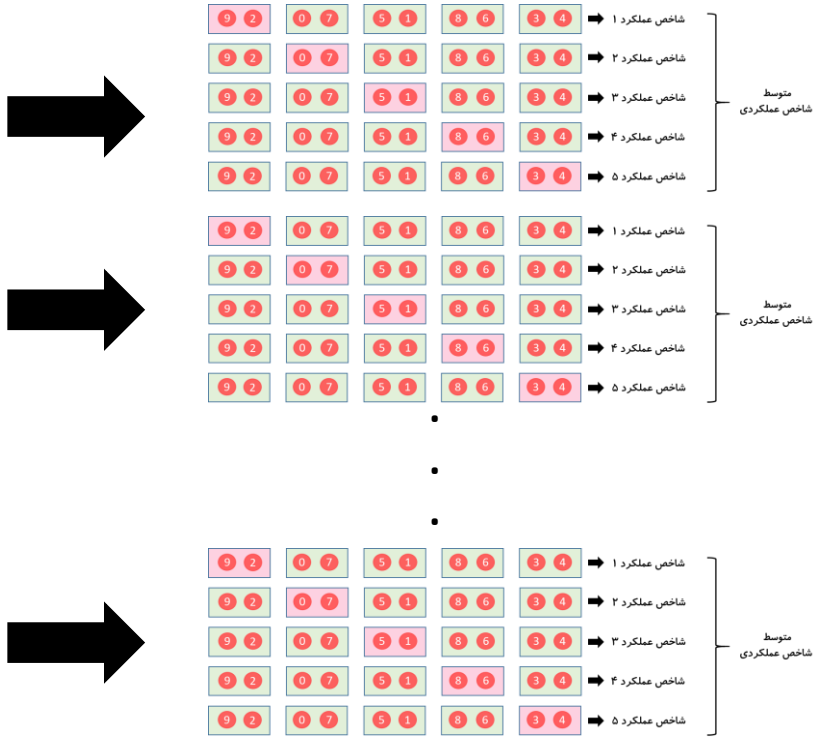
- از داده‌های آزمایش برای ارزیابی عملکرد مدل جهت تنظیم فراپارامترها استفاده نکنید.
- چگونه مدل آموزش دیده را در فاز آموزش ارزیابی کنیم؟
  - اعتبارسنجی مدل (Validation)
  - تقسیم داده‌های آموزش به دو بخش
    - Training set
    - Validation set

الگوریتم k-Fold Cross-Validation

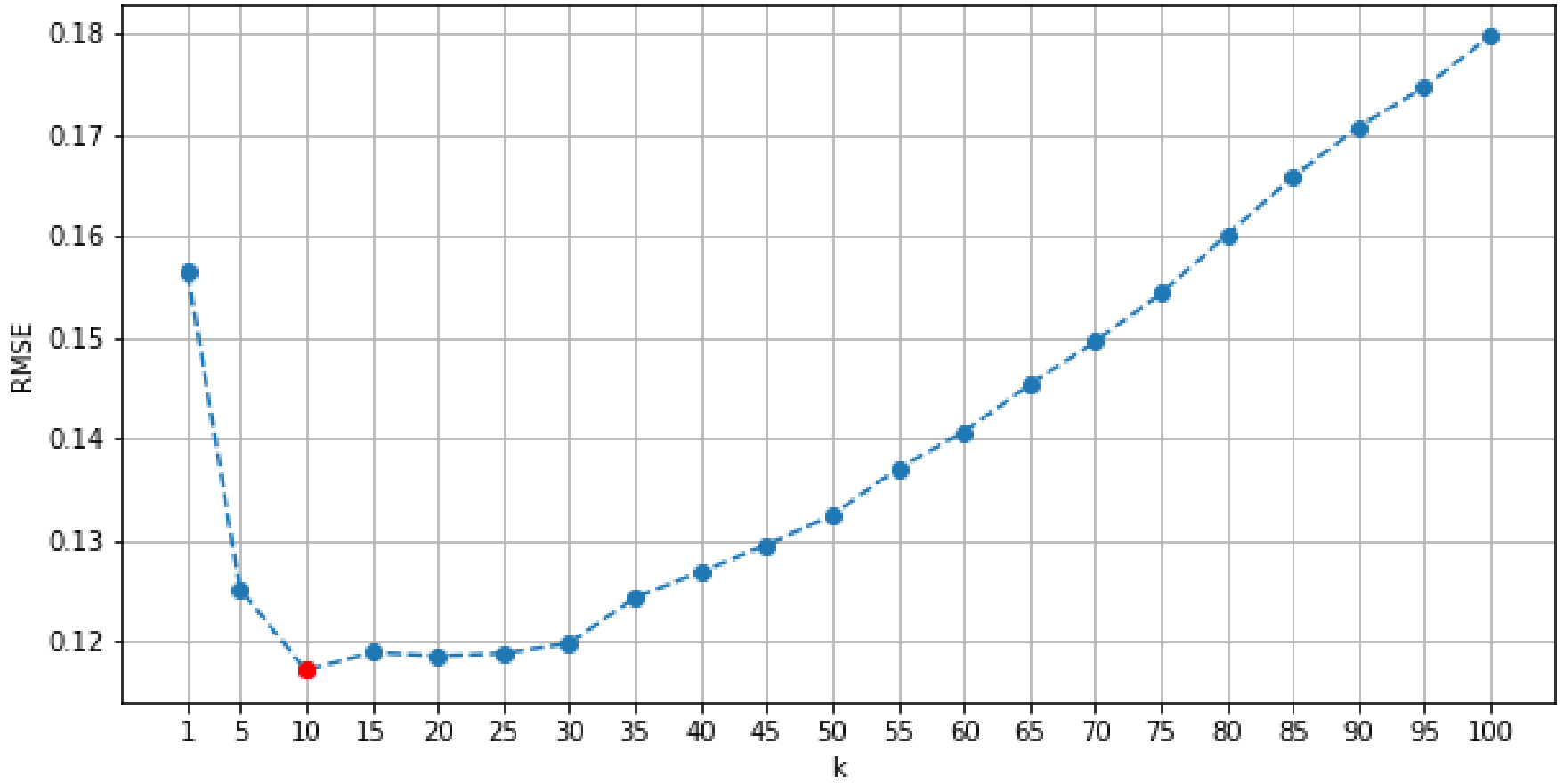


الگوریتم k-Fold Cross-Validation

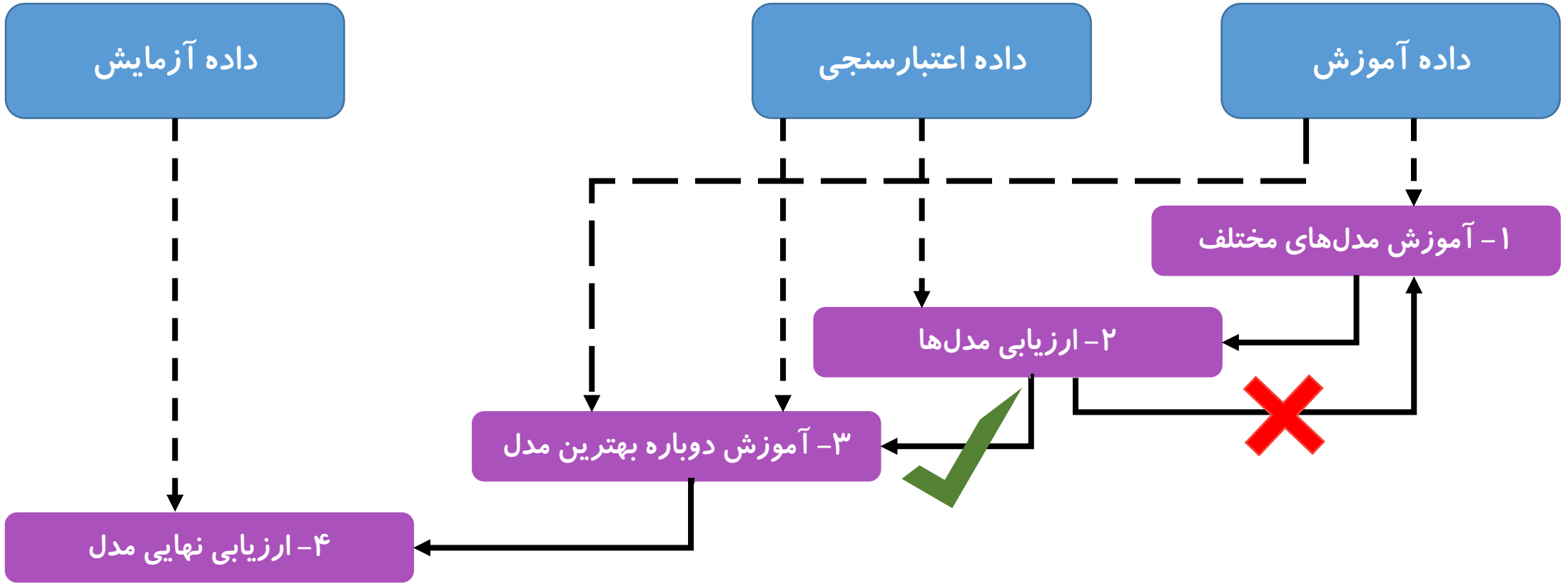
Hyperparameter (k)
1
5
. . .
100



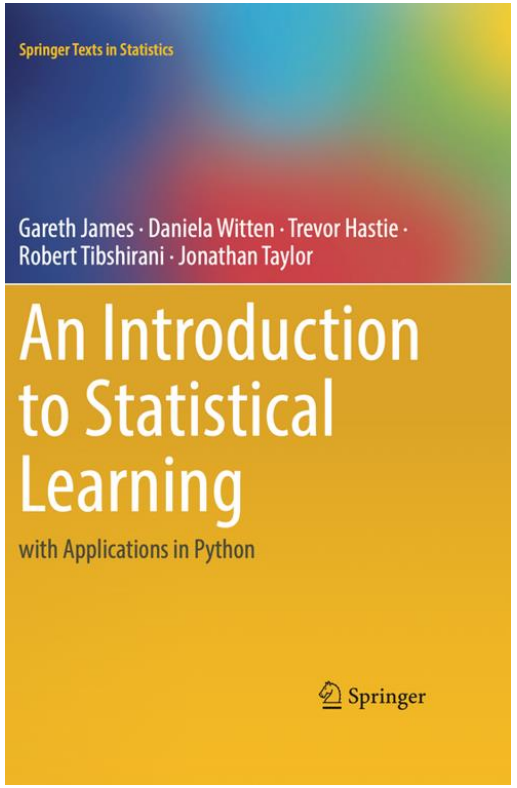
تعادل بین سوگیری و واریانس در الگوریتم kNN



فرآیند مدل‌سازی در یادگیری ماشین



پیشنهاد مطالعه



https://www.statlearning.com/

برنامه‌نویسی در پایتون

اجرای الگوریتم kNN و آشنایی با
مفهوم تعادل بین سوگیری و
واریانس در یادگیری ماشین