Machine Learning with Python

Session 1

By Dr. Maryam Rahbaralam

31 July 2020

What is machine learning?

زیر مجموعه‌ای از هوش مصنوعی، الگوریتم‌های یادگیری ماشین یک مدل ریاضی بر اساس داده‌های نمونه یا **"داده‌های آموزش"** به منظور پیش‌بینی یا تصمیم‌گیری، ایجاد می‌کنند

اهداف و انگیزه‌ها:

- **هدف یادگیری ماشین این است که رایانه‌ها و سامانه‌ها بتوانند به تدریج و با افزایش داده‌ها کارایی بهتری در انجام وظیفهٔ مورد نظر پیدا کند.**
- **یادگیری ماشین کمک فراوانی به صرفه جویی در هزینه‌های عملیاتی و بهبود سرعت عمل تجزیه و تحلیل داده‌ها می‌کند.**

تقسیم‌بندی‌های متداول در یادگیری ماشینی

تقسیم‌بندی بر اساس نوع داده‌های در اختیار کارگزار هوشمند است.

یادگیری با نظارت

یادگیری تقویتی

یادگیری بی نظارت

یادگیری با نظارت

در این حالت شما به کامپیوتر گفته‌اید که چه ورودی را به چه خروجی مربوط کند. دقت کنید که هم ورودی و هم خروجی مشخص است و در اصطلاح خروجی برچسب‌دار[1] است. به این شیوهٔ یادگیری، **یادگیری با نظارت** می‌گویند.

یادگیری تقویتی

اینک حالت دیگری را فرض کنید. برخلاف دفعهٔ پیشین که به رباتتان می‌گفتید چه محرکه‌ای را به چه خروجی ربط دهد، این‌بار می‌خواهید ربات خودش چنین چیزی را یاد بگیرد. به این صورت که اگر درست تشخیص داد به نحوی به او پاداش دهید و اگر به اشتباه، او را تنبیه کنید. در این حالت به ربات نمی‌گویید به ازای هر شرایطی چه کاری مناسب است، بلکه اجازه می‌دهید ربات خود کاوش کند و تنها شما نتیجهٔ نهایی را **تشویق** یا **تنبیه** می‌کنید. به این شیوهٔ یادگیری، **یادگیری تقویتی** می‌گویند.

یادگیری بی نظارت

در دو حالت پیش قرار بود ربات ورودی را به خروجی مرتبط کند. اما گاهی وقت‌ها تنها می‌خواهیم ربات بتواند تشخیص دهد که آنچه می‌بیند را به نوعی به آنچه پیش‌تر دیده‌است ربط دهد بدون این‌که به‌طور مشخص بداند آن‌چیزی که دیده شده‌است چه چیزی است یا این‌که چه کاری در موقع دیدنش باید انجام دهد. ربات هوشمند شما باید بتواند بین صندلی و انسان تفاوت قایل شود بی‌آنکه به او بگوییم این نمونه‌ها صندلی‌اند و آن نمونه‌های دیگر انسان. در این‌جا برخلاف یادگیری با نظارت هدف ارتباط ورودی و خروجی نیست، بلکه تنها **دسته‌بندی** آن‌ها است. این نوع یادگیری که به آن **یادگیری بی نظارت**[۲] می‌گویند بسیار مهم است چون دنیای ربات پر از ورودی‌هایی است که کسی برچسبی به آن‌ها اختصاص نداده اما به وضوح جزئی از یک دسته هستند.

یادگیری تحت نظارت

Supervised Learning

یادگیری تحت نظارت، یک روش عمومی در یادگیری ماشین است که در آن به یک سیستم، مجموعه‌ای از جفت‌های ورودی – خروجی ارائه شده و سیستم تلاش می‌کند تا تابعی از ورودی به خروجی را فرا گیرد. یادگیری تحت نظارت نیازمند تعدادی داده ورودی به منظور آموزش سیستم است.

یادگیری تحت نظارت خود به دو دسته تقسیم می‌شود: رگرسیون و طبقه‌بندی.

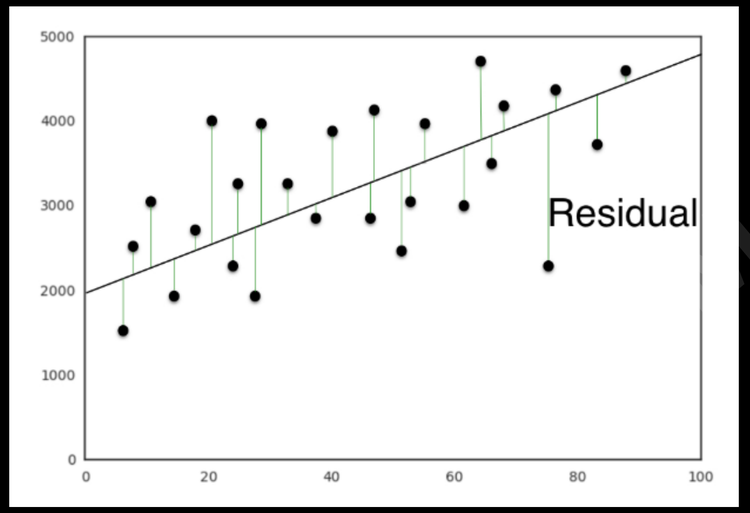
Regression
Classification

رگرسیون آن دسته از مسائل هستند که خروجی یک عدد **پیوسته** یا یک سری اعداد پیوسته هستند مانند پیش‌بینی قیمت خانه بر اساس اطلاعاتی مانند مساحت، تعداد اتاق خوابها، و غیره

طبقه‌بندی به آن دسته از مسائل گفته می‌شود که خروجی یک عضو از یک مجموعه باشد مانند پیش‌بینی اینکه یک ایمیل هرزنامه **هست یا خیر** یا پیش‌بینی نوع بیماری یک فرد از میان ۱۰ بیماری.

رگرسیون خطی

فرق رگرسیون خطی با سایر مدلهای رگرسیون در این است که در این مدل رابطهٔ بین متغیرهای مستقل و متغیر وابسته یک **رابطهٔ خطی** فرض می‌شود.[۱۷] رگرسیون خطی، که خود نوعی تابع پیش‌بینی‌کنندهٔ خطی است، پیش‌بینی متغیر وابسته را از حاصلجمع ضرب متغیرهای مستقل در یک سری ضرایب به‌دست می‌آورد. در **رگرسیون خطی ساده** که تنها یک متغیر مستقل وجود دارد، پیش‌بینی متغیر وابسته شکل یک خط مستقیم به خود می‌گیرد.



مثلاً تحلیل رگرسیونی سادهٔ با $y = ax + b$

روش رایج برای به‌دست‌آوردن پارامترها، روش **کمترین مربعات** است.

در این روش پارامترها را با کمینه‌کردن مجموع مربعات خطا به دست می‌آورند

x: متغیر مستقل

Y: متغیر وابسته

a:b پارامترهای مدل

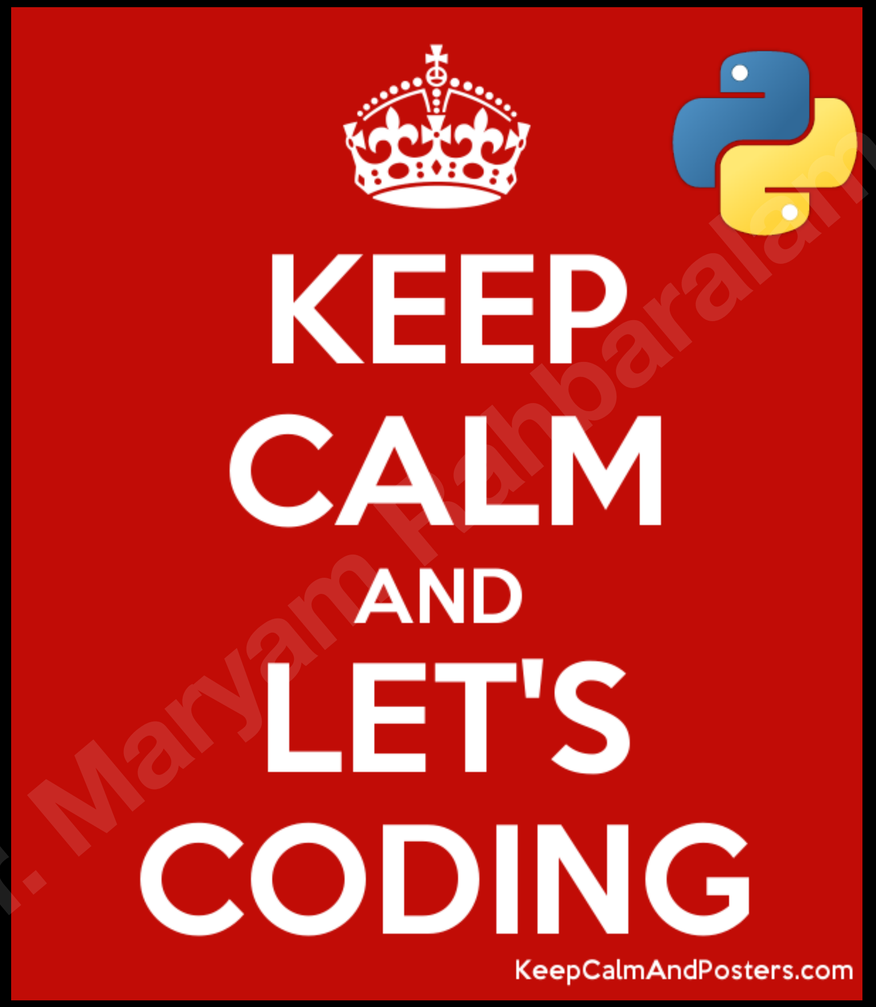
ضریب تشخیص

(به انگلیسی: Coefficient of Determination) که با علامت R2

- بیانگر میزان احتمال هم‌بستگی میان دو دسته داده در آینده می‌باشد. این ضریب در واقع نتایج تقریبی پارامتر موردنظر در آینده را بر اساس مدل ریاضی تعریف شده که منطبق بر داده‌های موجود است، بیان می‌دارد.

- ضریب تعیین، معیاری است از این که خط رگرسیون، چقدر خوب خواندها را معرفی می‌کند. اگر خط رگرسیون از تمام نقاط بگذرد توانائی معرفی همه متغیرها را دارد و هرچه از نقاط دورتر باشد نشان دهنده توانائی کمتر است

Compute and print the R2 score using the .score() method.



Exploring the Gapminder data

- In this chapter, you will work with Gapminder data, CSV file available in the workspace as ‘gapminder_all.csv'. Specifically, your goal will be to use this data **to predict the life expectancy** in a given country **based on features** such as the country's GDP, fertility rate, and population.
- Explore the DataFrame using pandas methods such as .info(), .describe(), .head()
- Heatmap showing the correlation between the different features of the Gapminder dataset
    - show positive correlation and negative correlation.

Let’s start coding in Python:

رگرسیون خطی ساده

- Since the target variable here is quantitative, this is a regression problem.
- To begin, you will fit a linear regression with just one feature: 'fertility', which is the average number of children a woman in a given country gives birth to.
- In later exercises, you will use all the features to build regression models.

- Before that, however, you need to

رگرسیون خطی ساده

1. Importing libraries and data for supervised learning
2. import the data and get it into the form needed by scikit-learn.
3. This involves creating feature and target variable arrays.
4. Furthermore, since you are going to use only one feature to begin with, you need to do some reshaping using NumPy's .reshape() method. Don't worry too much about this reshaping right now, but it is something you will have to do occasionally when working with scikit-learn so it is useful to practice.

Regression in Python:

- Use the function LinearRegression() to create the regressor.
- Use the .fit() method on reg with X(fertility) and y as arguments to fit the model.
- Use the .predict() method on reg with prediction_space as the argument to compute the predictions.
- Use the .score() method with X(fertility) and y as arguments to compute the R2 score.

References:

- Pattern Recognition and Machine Learning Book by Christopher Bishop
- https://en.wikipedia.org/wiki/Pearson_correlation_coefficient
- https://en.wikipedia.org/wiki/Coefficient_of_determination

Thank you!