Before we get started:

Open with Google Colaborat... ▾

→

In-call messages   ✕

Let everyone send mess...   ✓

Messages can only be ... n by people in the call and
are deleted ... n the call ends

You 12:30 PM
https://drive.google.com/fle.../1Xqac_Y8vDpAFzA_Kl
SOLDDBZXs9F5YHh/view?us..._link

File Edit View Insert Runtime Tools Help   Last edited on September 19

Locate in Drive

Open in playground mode

New notebook

Open notebook   Ctrl+O

Upload notebook

Rename

Move

Move to trash

Save a copy in Drive

Save a copy as a GitHub Gist

Save a copy in GitHub

om sc

From A to Z. From zero t

NLP from scratch

{"cells":[{"cell_type":"markdown","metadata":{"id":"fwukZZnNTYWE","tags":[]},"source":["<center><img src=\"https://drive.google.com/uc?export=view&id=1-1t6Uft8lgBG9jPD0dO6w3dAcv_EUQRP\"></center>\n","\n","## Learn natural language processing. From A to Z. From zero to hero. Fast.\n","\n","Copyright, NLP from scratch, 2023.\n","\n","[nlpfromscratch.com](https://www.nlpfromscratch.com)\n","\n","------------"]},
{"cell_type":"markdown","metadata":{"id":"gmAd3hBe94HF"},"source":["# Webinar #2 - What the Heck is an LLM?"]},

{"cell_type":"markdown","metadata":{"id":"uuznFjxWVVOn"},"source":["## Introduction 📖\n","In this notebook, we will see several different applications of Large Language Models (LLMs), and show how they can be leveraged the open source libraries from [Hugging Face](https://huggingface.co/).\n","\n","This notebook is best run in [Google Colab](https://colab.research.google.com/), where the majority of dependencies are already installed. However, if you wish to run the notebook locally, please follow the [directions for setting up a local environment](https://drive.google.com/file/d/1EV1seK-dUHRCzj2EDuu3ETAhUyjzOGRd/view?usp=drive_link) and you may then download the notebook as a `.ipynb` and run in either Jupyter or Jupyterlab.\n","\n","Though Google Colab comes with many useful data science libraries included by default (including Pytorch), the Hugging Face libraries are not, so we will first install those here using `pip`, as they will be used in the remainder of the notebook.\n","\n","- The `transformers` library, for general usage of transformer models\n","- The `datasets` library, for working with datasets hosted on Hugging Face\n","- The `diffusers` library, for working with diffusion models for image generation\n","- The `accelerate` library, for using GPU for inference"]},{"cell_type":"code","execution_count":1,"metadata":{"colab":{"base_uri":"https://localhost:8080/"},"executionInfo":{"elapsed":9449,"status":"ok","timestamp":1702487727798,"user":{"displayName":"NLP from scratch","userId":"13636460506782883737"},"user_tz":300},"id":"fjcuFD9_-ajD","outputId":"1480bf7f-0a0e-4d55-bf16-70b955a1aeed"},"outputs":[{"output_type":"stream","name":"stdout","text":["Requirement already satisfied: transformers in /usr/local/lib/python3.10/dist-packages (4.35.2)\n","Collecting datasets\n"," Downloading datasets-2.15.0-py3-none-any.whl (521 kB)\n","\u001b[2K   \u001b[90m━━\u001b[0m \u001b[32m521.2/521.2 kB\u001b[0m \u001b[31m6.9 MB/s\u001b[0m eta \u001b[36m0:00:00\u001b[0m\n","\u001b[?25hCollecting diffusers\n"," Downloading diffusers-0.24.0-py3-none-any.whl (1.8 MB)\n","\u001b[2K   \u001b[90m━━\u001b[0m \u001b[32m1.8/1.8 MB\u001b[0m \u001b[31m24.6 MB/s\u001b[0m eta \u001b[36m0:00:00\u001b[0m\n","\u001b[?25hCollecting accelerate\n"," Downloading accelerate-0.25.0-py3-none-any.whl (265 MB/s eta \u001b[36m0:00:00\u001b[0m\n","\u001b[?25h"," Requirement already satisfied: filelock in /usr/local/lib/python3.10/dist-packages (from transformers) (3.13.1)\n"," Requirement already satisfied: huggingface-hub<1.0,>=0.16.4 in /usr/local/lib/python3.10/dist-packages (from transformers) (0.19.4)\n"," Requirement already satisfied: numpy>=1.17 in /usr/local/lib/python3.10/dist-packages (from transformers) (1.23.5)\n"," Requirement already satisfied: packaging>=20.0 in /usr/local/lib/python3.10/dist-packages (from transformers) (23.2)\n"," Requirement already satisfied: pyyaml>=5.1 in /usr/local/lib/python3.10/dist-packages (from transformers) (6.0.1)\n"," Requirement already satisfied: regex!=2019.12.17 in /usr/local/lib/python3.10/dist-packages (from transformers)"}]}
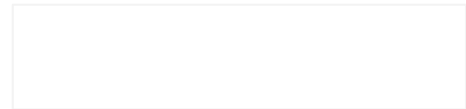
NLP from scratch

Machine Learning: From Zero

www.nlpfromscratch.com

Housekeeping

👤 **Camera on if comfortable doing so**

🚫📹 **This meeting will not be recorded**

🎤 **Stay muted unless speaking**

🤝 **Be professional**

NLP from scratch

Who am I?

- Data Scientist

- Career consultant (SapientNitro, PwC, Accenture)
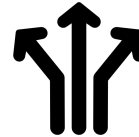
- Community builder

- Accidental entrepreneur

Manifesto

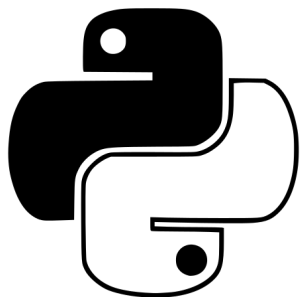Knowledge is only valuable if it is useful.

The best way to learn is by doing.

Learning is a non-linear process.

Learning is exploration, not a journey.

Teaching and learning are complementary.

NLP from scratch

Tools of the Trade

Python 3

Google Colab / Jupyter

Machine Learning

NLP from scratch

What is Machine Learning?

Data ⟶

Rules ⟶ **Classical Programming** ⟶ Answers

Data ⟶

Answers ⟶ **Machine Learning** ⟶ Rules

NLP from scratch
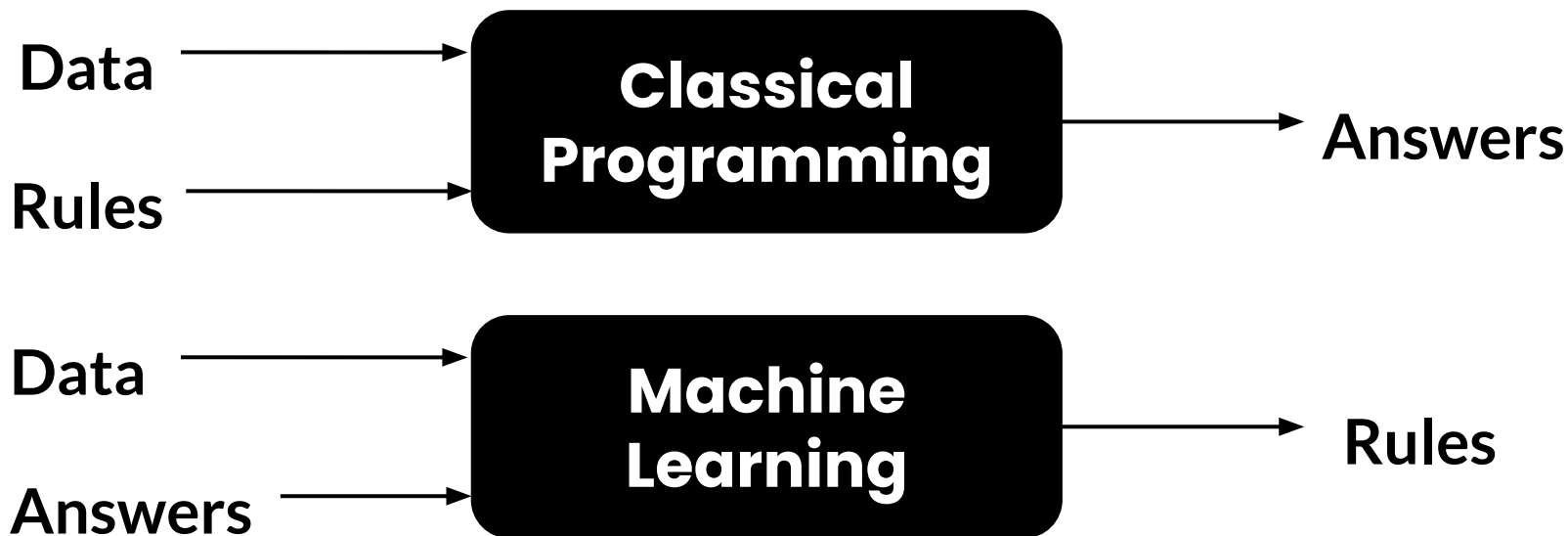
Types of Machine Learning
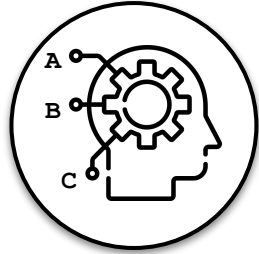
Supervised Learning
Making predictions

Make predictions from a dataset and *data labels* - categorical or numeric values associated with each observation in the training data.
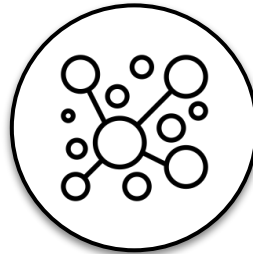
Unsupervised Learning
Finding patterns in data

Uses statistical techniques to uncover patterns in a dataset based upon its features. Unlike supervised learning, it does not require data labels.

Reinforcement Learning
Feedback & Rewards

Reinforcement Learning (RL) teaches an *agent* a behavior by optimizing against a target objective with a reward function.

NLP from scratch

Supervised Learning

TRAINING

$$X \quad y \quad \longrightarrow \quad \text{[ML Model]} \quad \longrightarrow \quad \hat{y}$$

Features   Labels   ML Model   Predictions

PREDICTION / EVALUATION

$$X \quad \longrightarrow \quad \text{[ML Model]} \quad \longrightarrow \quad \hat{y} \qquad \text{Error} = y - \hat{y}$$

Unseen data   ML Model   Predictions

NLP from scratch
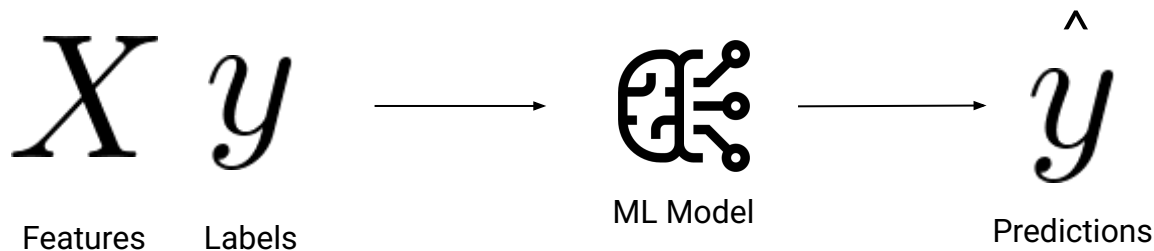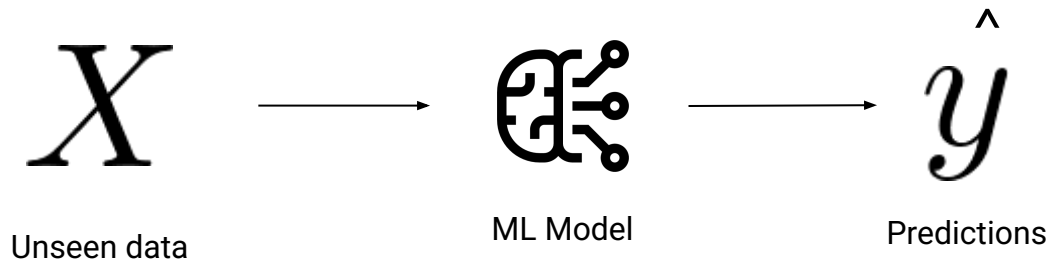
Linear Models

Regression

Linear Regression
(Ordinary Least Squares / OLS)

$$\hat{y} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \dots$$
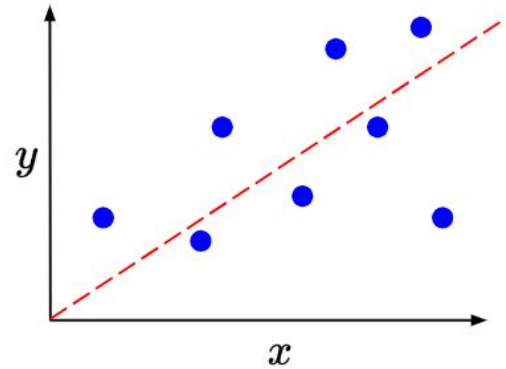
Classification Logistic Regression

$$P = \frac{e^{a+bX}}{1+e^{a+bX}}$$

$$Y = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$

$$\ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n$$
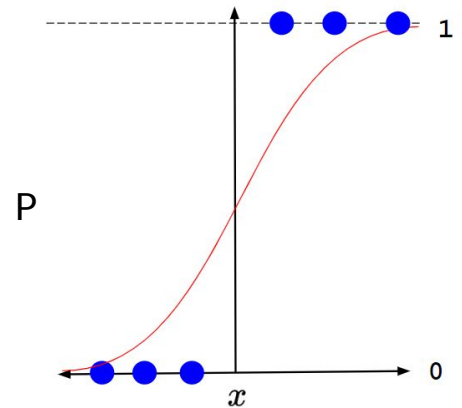
A Machine Learning Problem

Overfitting

The goal of a given ML model is to be as generalizable as possible and describe the underlying behavior or phenomenon of interest.

Overfitting refers to the model learning specifics of the particular dataset that was used for training, such that it does not perform well on new data.
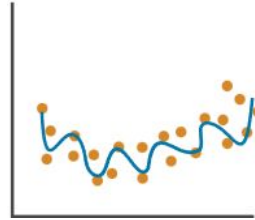
Overfitting:

The model has learned "too much" about the dataset used for training, including outlying observations and is fit too tightly to the data.
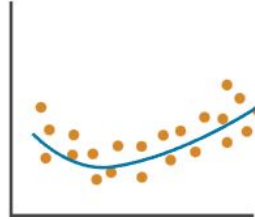
Good Fit:

The model is fit well, describing the underlying pattern in the data well in a general way. Its predictions are not perfect, but it does not exhibit excessive variance in order to capture all data points.
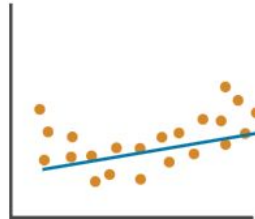
Underfitting:

The model has not learned enough about the patterns in the data and fails to adequately capture the relationship it is supposed to describe.

Image Attribution: https://www.mathworks.com/discovery/overfitting.html   *NLP from scratch*

Training and Testing an ML Model from Data

DATA

TRAINING DATA

TEST DATA

TRAIN MODEL

TRAINED MODEL

EVALUATE MODEL

SCORES (METRICS)
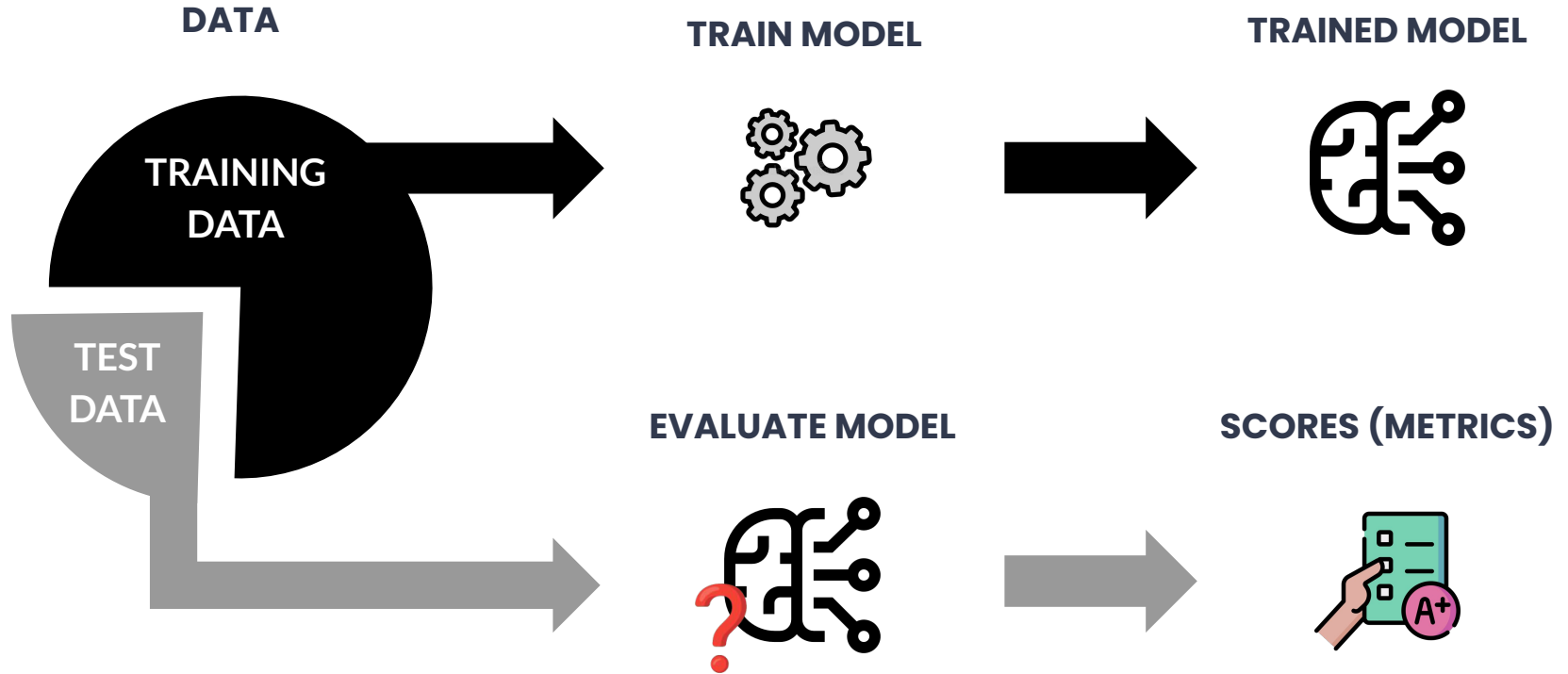
NLP from scratch

QUESTIONS

NLPfromscratch.com

NLP from scratch is a collection of free and Pay-What-You-Can (PWYC) courses and workshops created by Myles Harrison.

This official homepage contains all the files and relevant resources, hosted on Github.

- Webinars: Webinars on NLP, LLMs, and OpenAI delivered from October 2023 - Nov 2024.
- Workshops: Various workshops on Generative AI and related topics.
- Datasets: NLP / language datasets used in the above.
- Talks: Recordings of talks at various meetup events and conferences.

If you find the materials I've created useful, please consider supporting me in developing more by contributing a donation:

Donate

NLP from scratch