

Introduction to Machine Learning and Data Analytics

TRI-DUC NGHIEM

DUC.NGHIEM@PIPEPREDICT.COM

Overview



Machine Learning

Supervised learning
Unsupervised learning
Semi-supervised learning
Reinforcement learning



Data analytics

What to do?
Big data?



Practice

Playing around with scikit-learn, matplotlib

Part I: Machine Learning

► Example:

- Given a series of integer numbers:

► 1, 3, 5, 7, ...

What is the next number?

A linear model (series of odd numbers) will give us 9, but a polynomial function (degree of 4 as below) will give us: 217341

Find the next number of the sequence

1, 3, 5, 7, ?

Correct solution
217341

because when

$$f(x) = \frac{18111}{2}x^4 - 90555x^3 + \frac{633885}{2}x^2 - 452773x + 217331$$

$$f(1)=1$$

$$f(2)=3$$

$$f(3)=5$$

$$f(4)=7$$

$$f(5)=217341$$

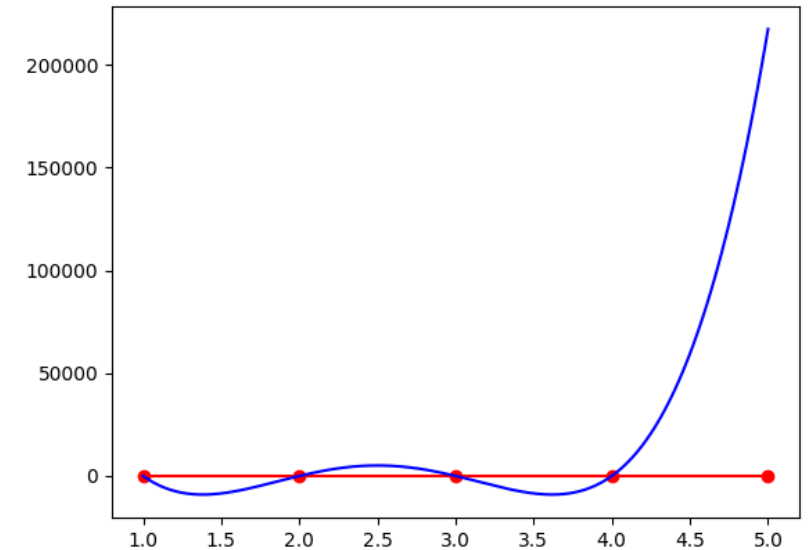
much solution

wow very logic

such function

many maths

wow



Part I: Machine Learning

► And more:

► $R = \frac{U}{I}$

► $F = G \frac{(m_1 \times m_2)}{r^2}$

► (It's true until Einstein proved it's wrong)

“All models are wrong, but some are useful”

George Box

I.I Supervised Machine Learning

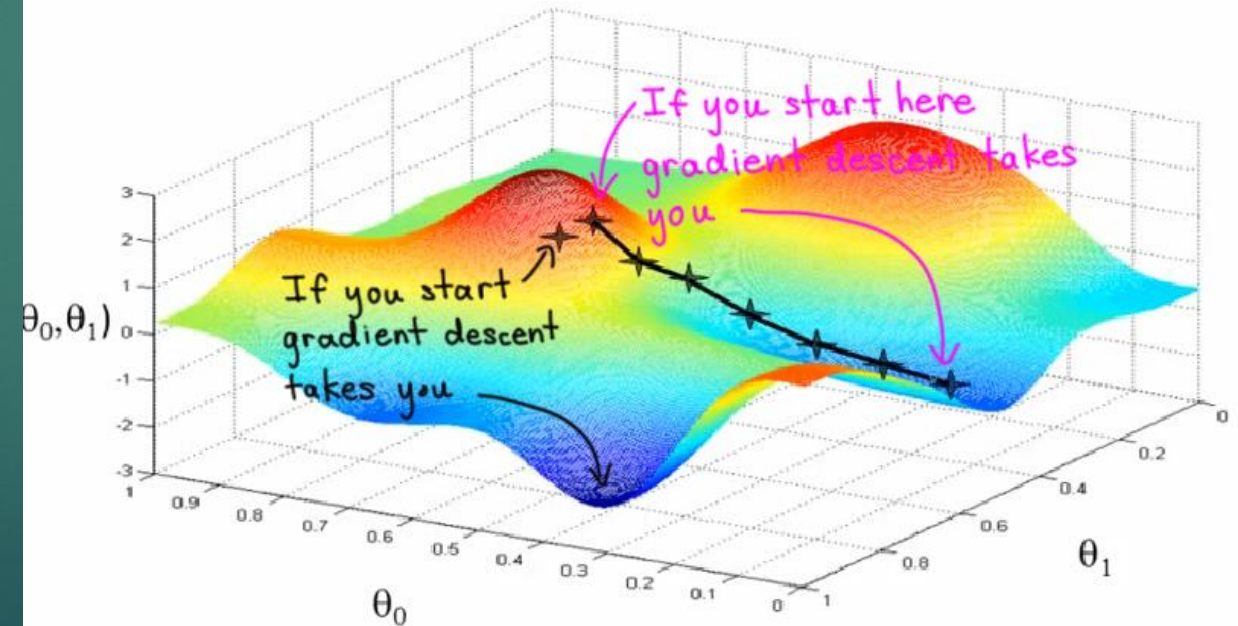
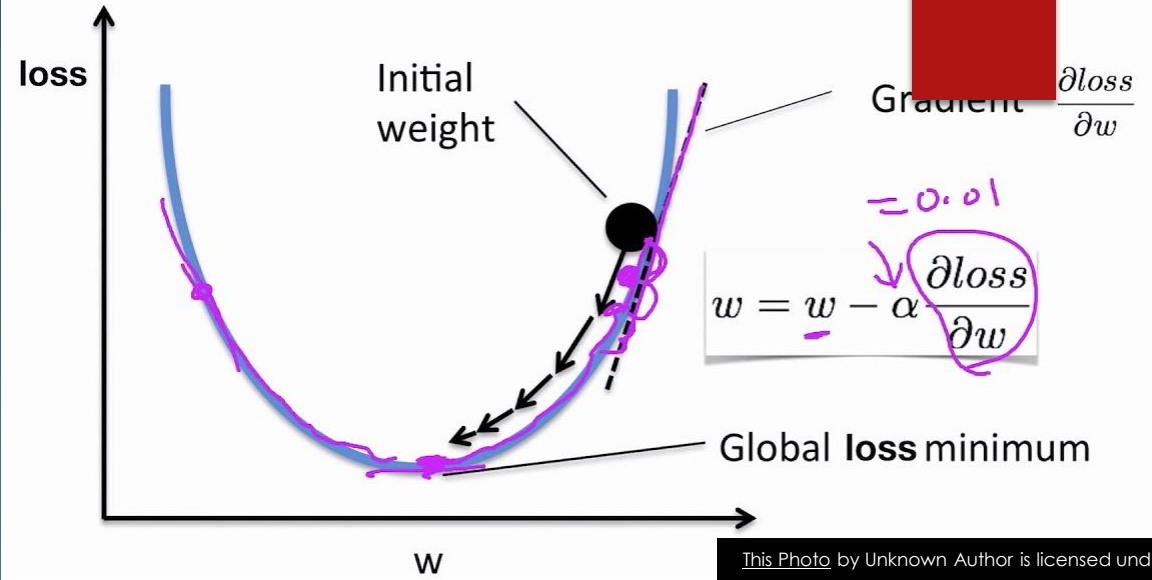
- ▶ Observation:
 - ▶ (x_i, y_i) where $i = 1..n$
- ▶ Estimate $F(x) = y$ with set of parameters and hyper-parameters
- ▶ Such that the loss over the observation set is minimum:
 - ▶ Minimize $loss = \sum_i L(F(x_i), y_i)$
- ▶ Example:
 - ▶ In the first example: $y = w_0 + w_1 \times x + w_2 \times x^2 + w_3 \times x^3 + w_4 \times x^4$
 - ▶ Hyper-parameters: degree of the polynomial: 4
 - ▶ Parameters: $w(w_0, w_1, w_2, w_3, w_4)$
 - ▶ **Deep Learning** is nothing different, it's just a chain of functions applied on the inputs



I.I Supervised Machine Learning

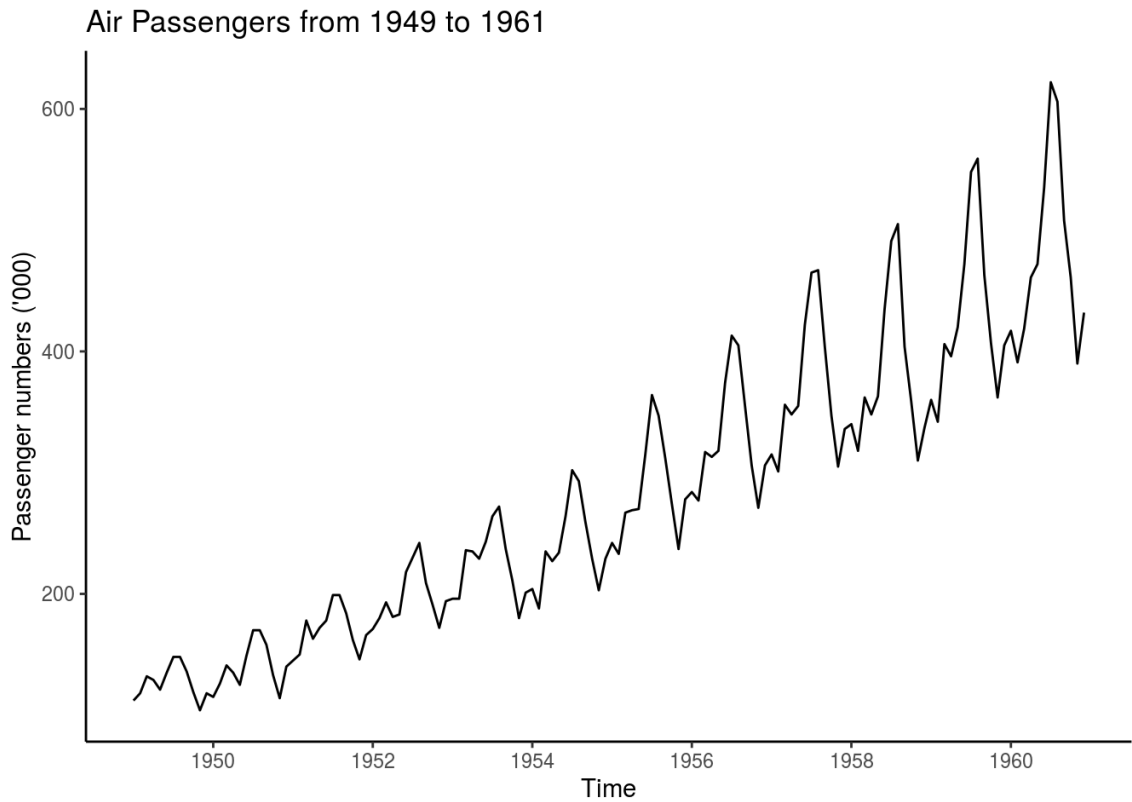
How does it work?

Gradient descent algorithm



I.I Supervised ML: Regression

- ▶ If F estimates continuous output, then the problem is a regression problem
 - ▶ Regular Regression problems:
 - ▶ Predict the price of a house / an apartment based on:
 - ▶ Location
 - ▶ Size
 - ▶ built-year
 - ▶ Timeseries: forecast future values of a series based on observed set of values
 - ▶ Price of flight ticket
 - ▶ Stock market



I.I Supervised ML: Classification

- ▶ If F's range is discrete and limited
- ▶ Examples:
 - ▶ Spam filter (binary classification)
 - ▶ Text categorization (multiple-class classification)
 - ▶ Handwriting recognition
 - ▶ Speech recognition
 - ▶ ...



Example of MNIST Dataset – handwriting digits recognition

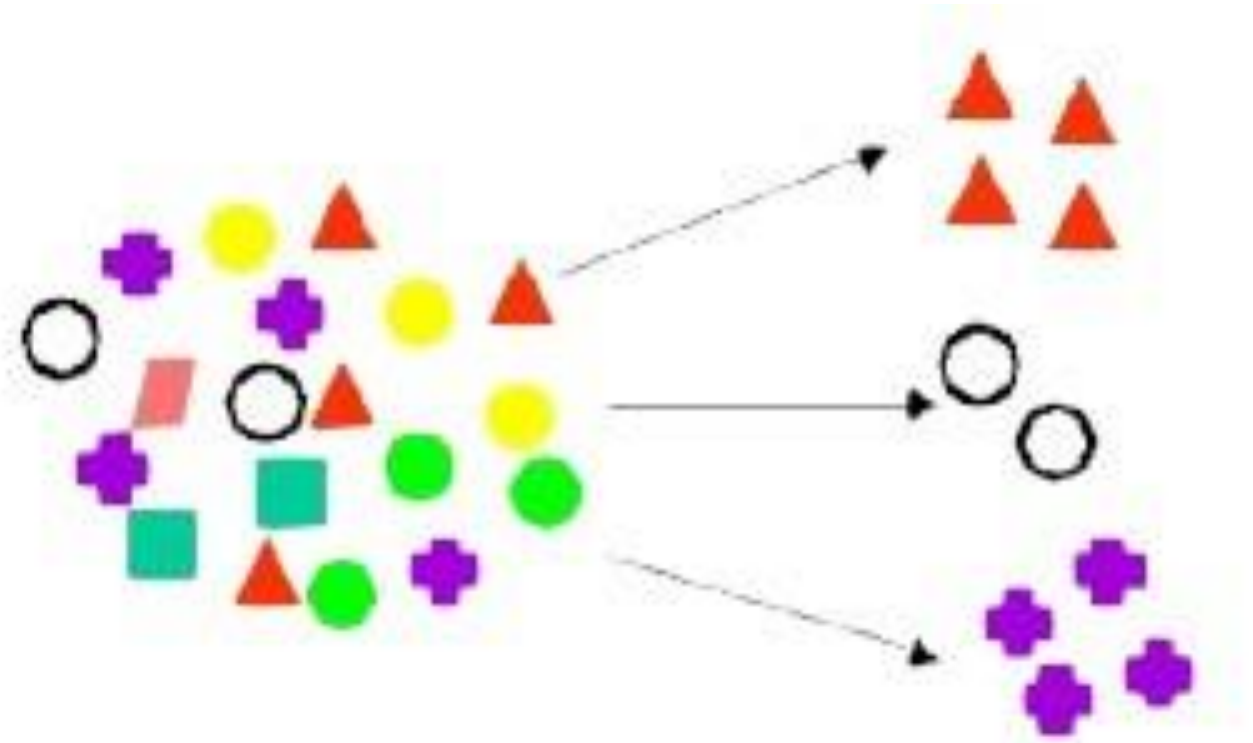
1998, 1999, 2000, 2001, 2002, 2003, 2004, 2005, 2006, 2007, 2008, 2009, 2010, 2011, 2012, 2013, 2014, 2015, 2016, 2017, 2018, 2019, 2020, 2021, 2022, 2023, 2024, 2025, 2026, 2027, 2028, 2029, 2030, 2031, 2032, 2033, 2034, 2035, 2036, 2037, 2038, 2039, 2040, 2041, 2042, 2043, 2044, 2045, 2046, 2047, 2048, 2049, 2050, 2051, 2052, 2053, 2054, 2055, 2056, 2057, 2058, 2059, 2060, 2061, 2062, 2063, 2064, 2065, 2066, 2067, 2068, 2069, 2070, 2071, 2072, 2073, 2074, 2075, 2076, 2077, 2078, 2079, 2080, 2081, 2082, 2083, 2084, 2085, 2086, 2087, 2088, 2089, 2090, 2091, 2092, 2093, 2094, 2095, 2096, 2097, 2098, 2099, 2100, 2101, 2102, 2103, 2104, 2105, 2106, 2107, 2108, 2109, 2110, 2111, 2112, 2113, 2114, 2115, 2116, 2117, 2118, 2119, 2120, 2121, 2122, 2123, 2124, 2125, 2126, 2127, 2128, 2129, 2130, 2131, 2132, 2133, 2134, 2135, 2136, 2137, 2138, 2139, 2140, 2141, 2142, 2143, 2144, 2145, 2146, 2147, 2148, 2149, 2150, 2151, 2152, 2153, 2154, 2155, 2156, 2157, 2158, 2159, 2160, 2161, 2162, 2163, 2164, 2165, 2166, 2167, 2168, 2169, 2170, 2171, 2172, 2173, 2174, 2175, 2176, 2177, 2178, 2179, 2180, 2181, 2182, 2183, 2184, 2185, 2186, 2187, 2188, 2189, 2190, 2191, 2192, 2193, 2194, 2195, 2196, 2197, 2198, 2199, 2200, 2201, 2202, 2203, 2204, 2205, 2206, 2207, 2208, 2209, 2210, 2211, 2212, 2213, 2214, 2215, 2216, 2217, 2218, 2219, 2220, 2221, 2222, 2223, 2224, 2225, 2226, 2227, 2228, 2229, 2230, 2231, 2232, 2233, 2234, 2235, 2236, 2237, 2238, 2239, 2240, 2241, 2242, 2243, 2244, 2245, 2246, 2247, 2248, 2249, 2250, 2251, 2252, 2253, 2254, 2255, 2256, 2257, 2258, 2259, 2260, 2261, 2262, 2263, 2264, 2265, 2266, 2267, 2268, 2269, 2270, 2271, 2272, 2273, 2274, 2275, 2276, 2277, 2278, 2279, 2280, 2281, 2282, 2283, 2284, 2285, 2286, 2287, 2288, 2289, 2290, 2291, 2292, 2293, 2294, 2295, 2296, 2297, 2298, 2299, 2300, 2301, 2302, 2303, 2304, 2305, 2306, 2307, 2308, 2309, 2310, 2311, 2312, 2313, 2314, 2315, 2316, 2317, 2318, 2319, 2320, 2321, 2322, 2323, 2324, 2325, 2326, 2327, 2328, 2329, 2330, 2331, 2332, 2333, 2334, 2335, 2336, 2337, 2338, 2339, 2340, 2341, 2342, 2343, 2344, 2345, 2346, 2347, 2348, 2349, 2350, 2351, 2352, 2353, 2354, 2355, 2356, 2357, 2358, 2359, 2360, 2361, 2362, 2363, 2364, 2365, 2366, 2367, 2368, 2369, 2370, 2371, 2372, 2373, 2374, 2375, 2376, 2377, 2378, 2379, 2380, 2381, 2382, 2383, 2384, 2385, 2386, 2387, 2388, 2389, 2390, 2391, 2392, 2393, 2394, 2395, 2396, 2397, 2398, 2399, 2400, 2401, 2402, 2403, 2404, 2405, 2406, 2407, 2408, 2409, 2410, 2411, 2412, 2413, 2414, 2415, 2416, 2417, 2418, 2419, 2420, 2421, 2422, 2423, 2424, 2425, 2426, 2427, 2428, 2429, 2430, 2431, 2432, 2433, 2434, 2435, 2436, 2437, 2438, 2439, 2440, 2441, 2442, 2443, 2444, 2445, 2446, 2447, 2448, 2449, 2450, 2451, 2452, 2453, 2454, 2455, 2456, 2457, 2458, 2459, 2460, 2461, 2462, 2463, 2464, 2465, 2466, 2467, 2468, 2469, 2470, 2471, 2472, 2473, 2474, 2475, 2476, 2477, 2478, 2479, 2480, 2481, 2482, 2483, 2484, 2485, 2486, 2487, 2488, 2489, 2490, 2491, 2492, 2493, 2494, 2495, 2496, 2497, 2498, 2499, 2500, 2501, 2502, 2503, 2504, 2505, 2506, 2507, 2508, 2509, 2510, 2511, 2512, 2513, 2514, 2515, 2516, 2517, 2518, 2519, 2520, 2521, 2522, 2523, 2524, 2525, 2526, 2527, 2528, 2529, 2530, 2531, 2532, 2533, 2534, 2535, 2536, 2537, 2538, 2539, 2540, 2541, 2542, 2543, 2544, 2545, 2546, 2547, 2548, 2549, 2550, 2551, 2552, 2553, 2554, 2555, 2556, 2557, 2558, 2559, 2560, 2561, 2562, 2563, 2564, 2565, 2566, 2567, 2568, 2569, 2570, 2571, 2572, 2573, 2574, 2575, 2576, 2577, 2578, 2579, 2580, 2581, 2582, 2583, 2584, 2585, 2586, 2587, 2588, 2589, 2590, 2591, 2592, 2593, 2594, 2595, 2596, 2597, 2598, 2599, 2600, 2601, 2602, 2603, 2604, 2605, 2606, 2607, 2608, 2609, 2610, 2611, 2612, 2613, 2614, 2615, 2616, 2617, 2618, 2619, 2620, 2621, 2622, 2623, 2624, 2625, 2626, 2627, 2628, 2629, 2630, 2631, 2632, 2633, 2634, 2635, 2636, 2637, 2638, 2639, 2640, 2641, 2642, 2643, 2644, 2645, 2646, 2647, 2648, 2649, 2650, 2651, 2652, 2653, 2654, 2655, 2656, 2657, 2658, 2659, 2660, 2661, 2662, 2663, 2664, 2665, 2666, 2667, 2668, 2669, 2670, 2671, 2672, 2673, 2674, 2675, 2676, 2677, 2678, 2679, 26

- [illegible]

Word vectors visualization showed that it can capture semantic level of words

I.II Unsupervised Machine Learning

- ▶ No labels are given to the learning algorithm, leaving it on its own to find structure in its input
- ▶ Examples:
 - ▶ Group objects according to their shape

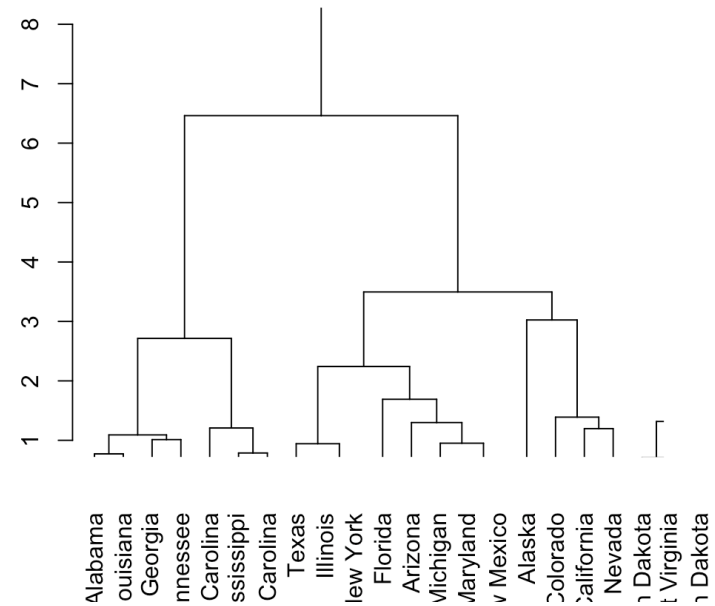


This Photo by Unknown Author is licensed under [CC BY-SA-NC](#)

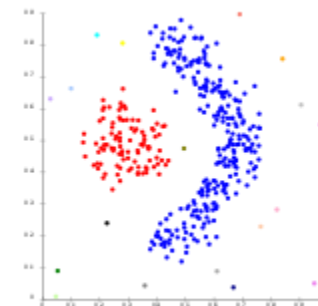
I.II Unsupervised ML

Clustering

- ▶ Discovering clusters
 - ▶ Clustering data into groups
- ▶ Hierarchical clustering
- ▶ Partitional clustering



Hierarchical clustering

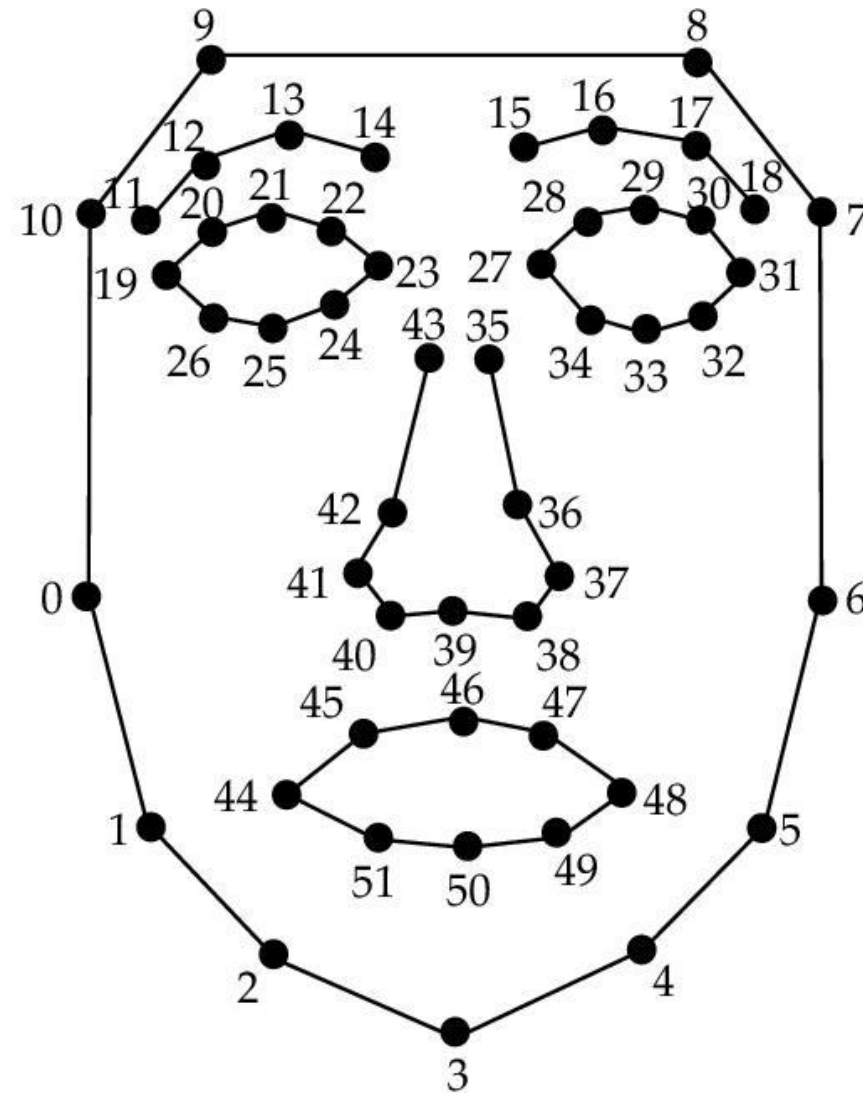


Partitional Clustering

I.II Unsupervised ML

Discovering Latent Factors

- ▶ Discovering latent factors
 - ▶ Dimensionality reduction: reduce high dimensional data to a lower dimensional subspace.
 - ▶ E.g: there are many features representing an images, but just few are important: lighting, pose, identity...



This Photo by Unknown Author is licensed under CC BY

I.II Unsupervised ML

Matrix Completion

- ▶ If we have missing data, then the goal of this task is filling the gaps with numbers based on the surrounding values
- ▶ Tasks:
 - ▶ Image inpainting
 - ▶ Collaborative filtering
 - ▶ E.g: Given rating matrix of users with watched movies, predict how likely they would rate an unwatched one.



Figure 11: Examples of object removal and image editing using our EdgeConnect model. (Left) Original image. (Center) Unwanted object removed with optional edge information. (Right) Inpainted image.

	target	movie_0	movie_1	movie_2	movie_3	movie_4
0	?	1	2	4	4	6
1	?	2	6	3	4	2
2	?	2	3	1	5	3
3	?	1	3	1	2	4
4	?	1	3	1	2	4

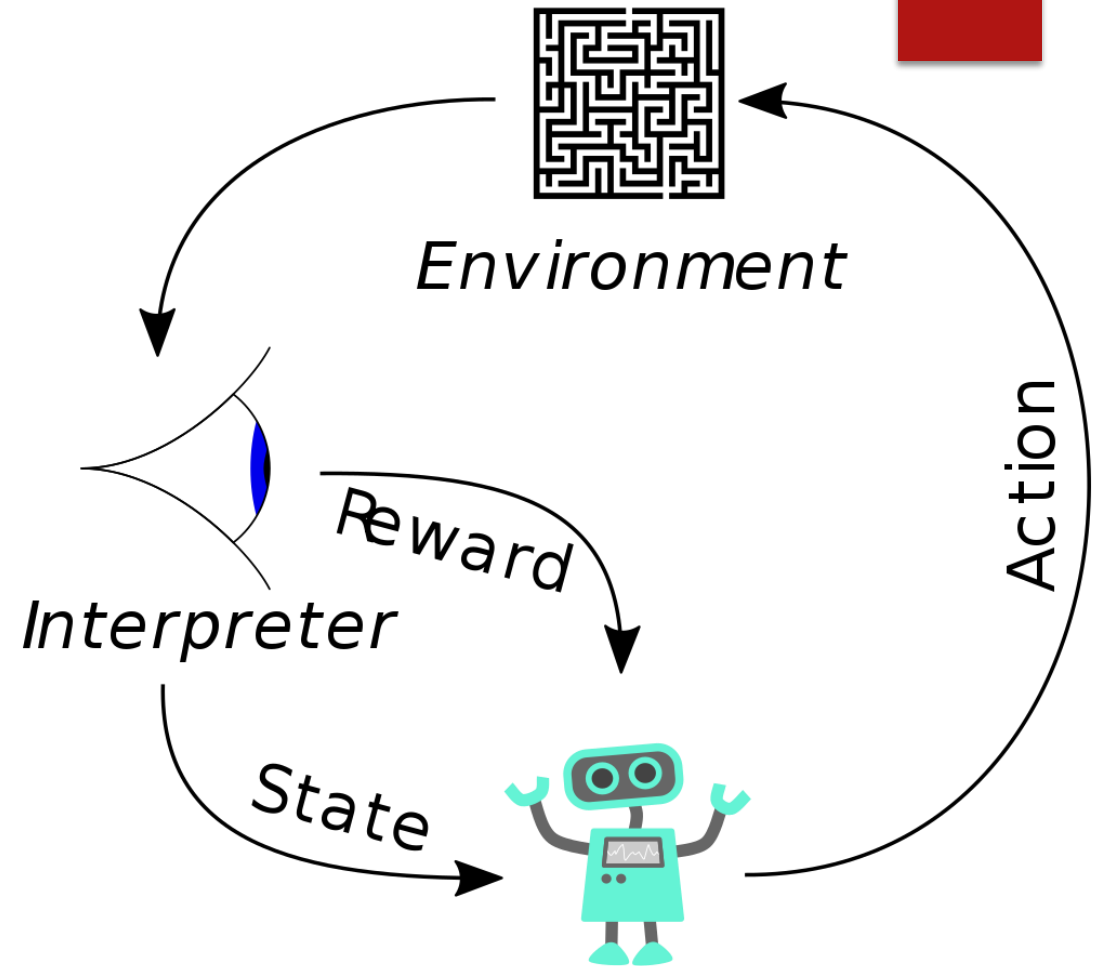
This Photo by Unknown Author is licensed under CC BY-SA

I.III Semi-supervised Learning

- ▶ If you have a small set of labelled data, and one to use them as a seed to grow the data bigger:
 - ▶ Train your supervised model on that small dataset
 - ▶ Run prediction on the bigger data set and get labels
 - ▶ Revise the quality
 - ▶ Use that data as training data...

I.IV Reinforcement Learning

► Reinforcement learning (RL) is an area of machine learning concerned with how software agents ought to take actions in an environment in order to maximize the notion of cumulative reward.



Source: wikipedia -
https://en.wikipedia.org/wiki/Reinforcement_learning

Part II: Data Analytics – What to do?

- ▶ Here are questions you have to answer:
 - ▶ What problem you are going to solve?
 - ▶ Is the data in good quality and in the format that you can load into memory and work with?
 - ▶ Is it big enough?
 - ▶ What model am I going to use?
 - ▶ How do I improve the performance of the model I am using?
 - ▶ How do I evaluate the result?
 - ▶ Visualize results?

II.I Data Analytics – What to do?

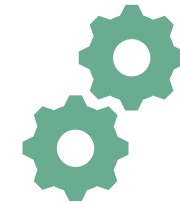
Preprocessing



Filtering, cleaning up data



Data conversion



Feature extraction

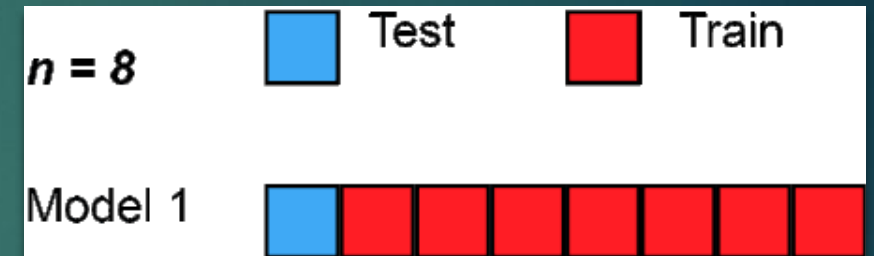
Feature engineering:

- Extract more features from given data
- Transform the features (feature scaling)

II.I Data Analytics – What to do?

Model Selection

- ▶ Split data into training set, test set, validation set (optional) if your problem is supervised
- ▶ Choose some models with different hyper-parameters and train them on the training set.
 - ▶ To tune the hyper-parameters, run it on validation set or n-fold validation on training set
 - ▶ Never do this:
 - ▶ Train and test on the same data set -> over fitting your model
 - ▶ Train on training set and test on test set but repeatedly until archive the best set of hyper-parameters -> overfitting



II.I Data Analytics – What to do?

Evaluation

- ▶ Regression:

- ▶ mean square error, root mean square error

- ▶ $mse = \frac{(\sum_i (y_i - y_i^*)^2)}{n}$

- ▶ Mean absolute error

- ▶ $mae = \sum_i \frac{|y_i - y_i^*|}{n}$

- ▶ R square:

- ▶ $R^2 = 1 - \frac{\sum_i (y_i - y_i^*)^2}{\sum_i (y_i - \bar{y})^2}$

II.I Data Analytics – What to do?

Evaluation

- ▶ Classification: Precision, Recall, F-1
 - ▶ True Positive: accurately classified as positive label (TP)
 - ▶ E.g: AIDS positive true, corona positive true
 - ▶ True negative: accurately classified as negative label (TN)
 - ▶ E.g: AIDS negative true, corona negative true
 - ▶ False Positive: wrongly classified as positive label (FP)
 - ▶ E.g: don't have AIDS, corona but got positive test
 - ▶ False Negative: wrongly classified as negative label (FN)
 - ▶ E.g: have AIDS, corona but got negative result

$$✓ \text{ Precision} = \frac{TP}{TP+FP},$$

$$✓ \text{ Recall} = \frac{TP}{TP+FN},$$

$$✓ F1 = \frac{2(P \times R)}{P+R},$$

$$✓ \text{ accuracy} = \frac{TP+TN}{TP+TN+FP+FN}$$

II.I Data Analytics – What to do? Evaluation

► Clustering:

► $Purity(\Omega, C) = \frac{1}{N} \sum_k \max_j |\omega_k \cap c_j|$

- N is number of members
- Ω is a set of result-clusters
- C is a set of gold standard clusters

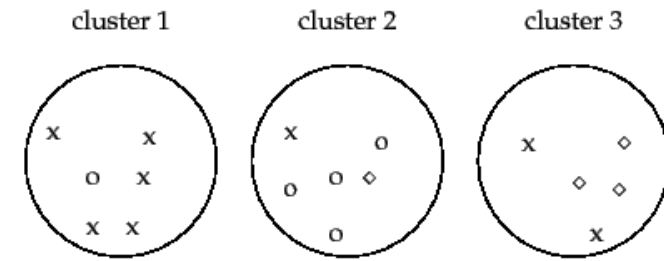
► B-cubed (B^3)

► $Precision = \frac{1}{N} \sum_{p \in Dataset} Precision(p)$

► $Recall = \frac{1}{N} \sum_{p \in Dataset} Recall(p)$

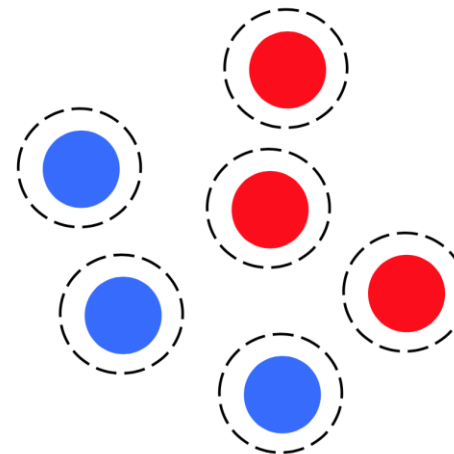
► $F - Score = \frac{1}{N} \sum_{p \in Dataset} F(p)$

- Where $n = \#clusters$

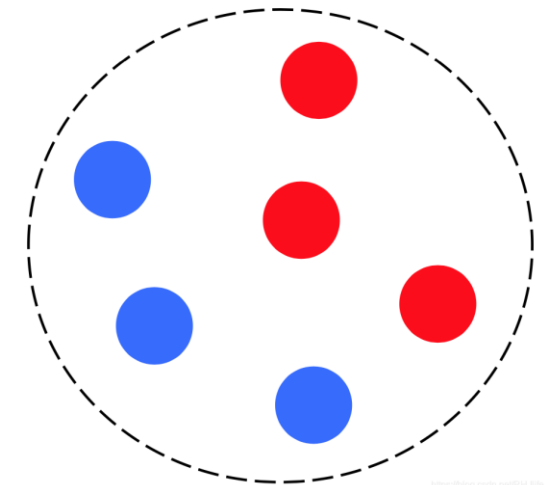


► **Figure 16.1** Purity as an external evaluation criterion for cluster quality. Majority class and number of members of the majority class for the three clusters are: x, 5 (cluster 1); o, 4 (cluster 2); and o, 3 (cluster 3). Purity is $(1/17) \times (5 + 4 + 3) \approx 0.71$.

100% Precision, 33% Recall



50% Precision, 100% Recall,



II.I Data Analytics – What to do?

Visualization

- ▶ Features visualization
 - ▶ Sub-space selection – our intuition is limited to 3-d space
- ▶ Result visualization
 - ▶ Selection of positive results
 - ▶ Selection of negative results (no system is perfect, even human)
 - ▶ Be creative! People love to see charts than table of numbers

Part II: Big Data

- ▶ If you have a big data (how big should be considered as big, well say 5TB data with the RAM and Hard Drive capacity at the moment – 2020), then you have to concern the technical issue to work with it as you cannot load them all to memory to analyze
- ▶ Solution: “divide and conquer”, take advantage of parallel computing
- ▶ There are many techniques out there, but the most famous in the recent years is: Hadoop (Map – Reduce)

Materials and Classes

- ▶ Books:

- ▶ Machine Learning – Tom Mitchell
- ▶ Machine Learning A probabilistic Perspective – Kevin P. Murphy (Advanced)
- ▶ Timeseries analysis – Hamilton
- ▶ Applied Multivariate Statistical Analysis – Richard A. Johnson and Dean W. Wichern

- ▶ Online course:

- ▶ Machine Learning – Andrew Ng (Recommended)
 - ▶ <https://www.coursera.org/learn/machine-learning>
- ▶ DeepLearning.ai – Andrew Ng (Recommended)
 - ▶ <https://www.coursera.org/specializations/deep-learning>
- ▶ Tip:
 - ▶ Select audit mode to attend those courses without certificate (and so they're free)

Part III Practice



PLAYING WITH
REGRESSION



PLAYING WITH
CLASSIFICATION

Part III

Practice

General TIPS



Import libraries

from library **import** classes, functions, definitions...

Import library



Numpy is library for matrix computation



Pandas is for loading structured data into DataFrame (similar concept as excel sheet with rows and columns), Series



Matplotlib is data visualization library



Sklearn is library for basic machine learning algorithms

Part III Practice

General TIPS

- ▶ Dataset:
 - ▶ <https://scikit-learn.org/stable/datasets/index.html>
- ▶ Loading dataset:
 - ▶ Regression:
 - ▶ `from sklearn.datasets import load_boston`
 - ▶ `X, y = load_boston(return_X_y=True)`
 - ▶ Classification:
 - ▶ `sklearn.datasets.load_breast_cancer`

Part III Practice

General TIPS

- ▶ Feature Engineering:
 - ▶ Scaling and transformation:
 - ▶ `from sklearn import preprocessing`
 - ▶ More here:
 - ▶ <https://scikit-learn.org/stable/modules/preprocessing.html>

Part III Practice

General TIPS

Linear Regression:

- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html

Logistic Regression (it's actually a classifier)

- https://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

Support Vector Machine

- Regressor: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVR.html>
- Classifier: <https://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html>

Part III Practice

General TIPS

- ▶ Train and predict
 - ▶ Choose your model: `model = YourModel(hyper_parameters)`
 - ▶ Fitting model with training set: `model.fit(x,y)`
 - ▶ Predict:
 - ▶ `y_hat = model.predict(x_test)`
 - ▶ Remember to scale the `x_test` before predict, if you scaled your training features
- ▶ Evaluation:
 - ▶ https://scikit-learn.org/stable/modules/model_evaluation.html