CMU Portugal Academy

CMU Portugal
Advanced Training Program

Foundations of Data Science

DAVID SEMEDO

RAFAEL FERREIRA

NOVA SCHOOL OF SCIENCE AND TECHNOLOGY

Email: df.semedo@fct.unl.pt

CMU Portugal Academy

Today's Topics

1. INTRO TO MACHINE LEARNING

2. FEATURE ENGINEERING

3. MODEL EVALUATION

4. SUPERVISED LEARNING
    - Linear Models
    - Classification
    - Regression

Use-Case Details

Requirements - One operation of each of the following:

- Dataset Descriptive Statistics
- Data Cleaning (e.g. checking for NaNs, column removal, etc.)
- Model Selection, Feature Engineering, and Normalization
- Plotting (frequency, correlation between feature pairs)
- Supervised Learning:
  - Training a linear classifier
  - Evaluate its performance over multiple metrics

Use-Case Discussion Session

Let's simulate a Data Science team discussion:

- Bring your expertise and point of view!

We Invite all groups to present and discuss their use-case with the class:

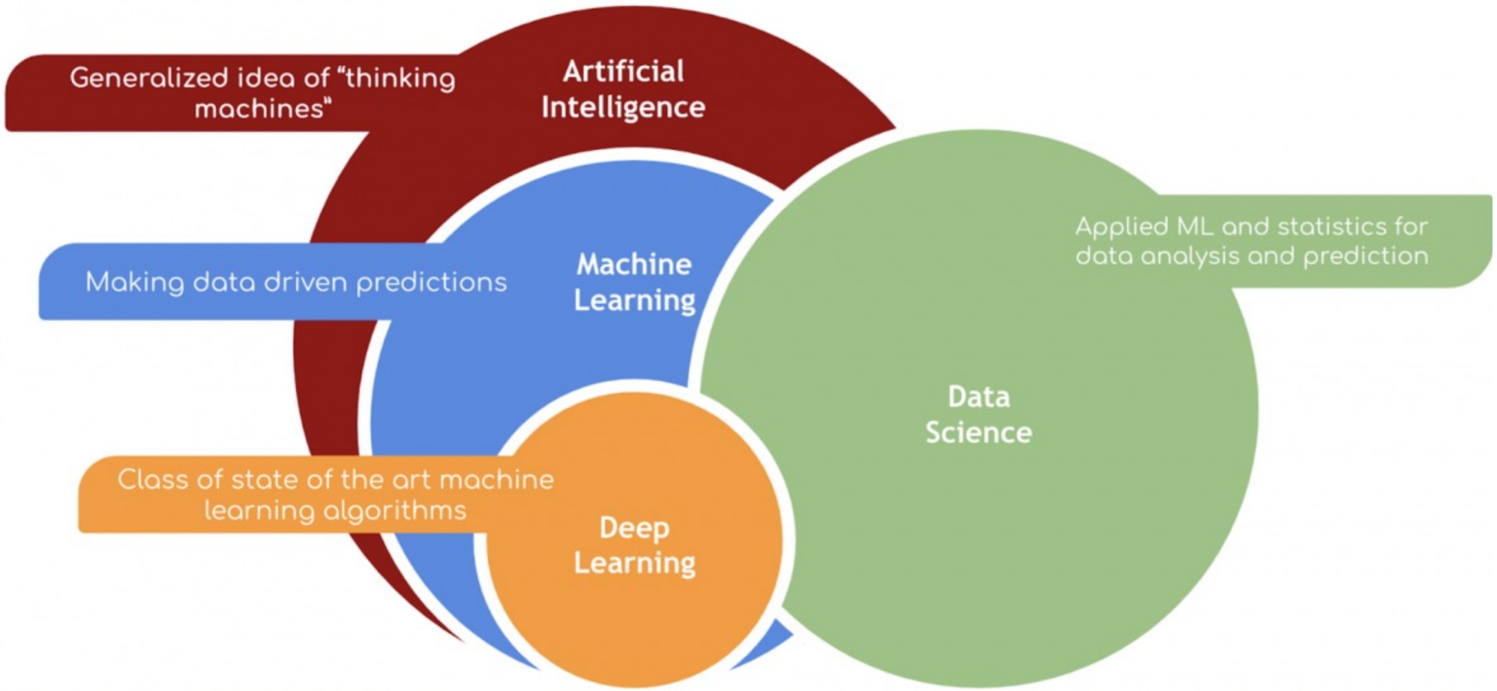
- Show and discuss your notebook to the class
- 5 to 7 minutes per presentation

Academy

01

Intro to Machine Learning

What is machine learning?



What is Machine Learning?

- [Arthur Samuel, 1959]
  - Field of study that gives computers the ability to learn without being explicitly programmed

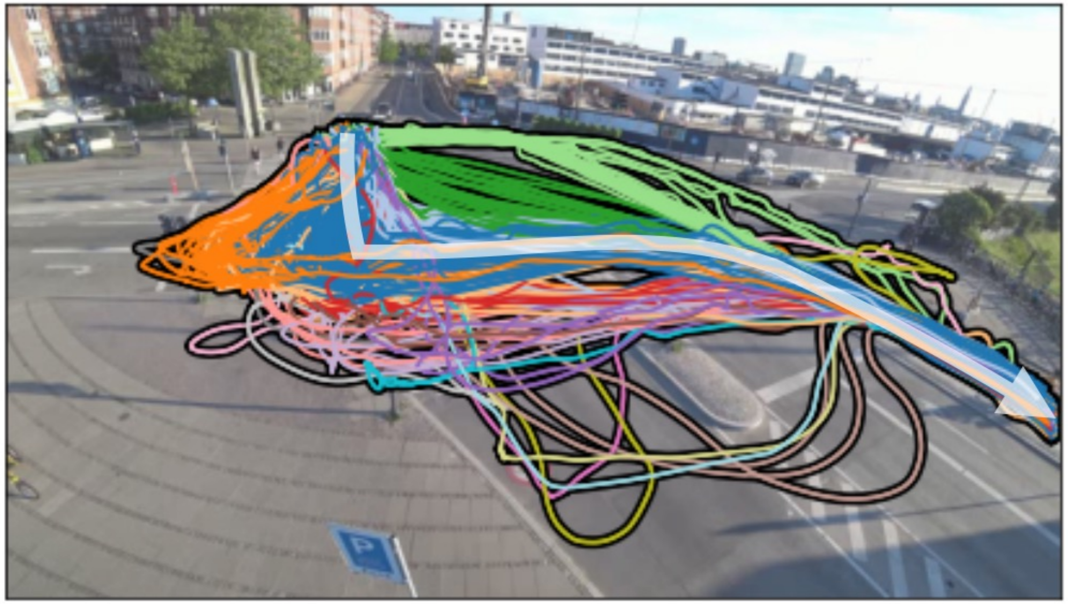
- [Kevin Murphy] algorithms that
  - automatically detect patterns in data
  - use the uncovered patterns to predict future data or other outcomes of interest

- [Tom Mitchell] algorithms that
  - improve their performance (P)
  - at some task (T)
  - with experience (E)

Why do we need learning?

Design

Reality



Breum, Kostic & Szell. Computational Desire Line Analysis of Cyclists on the Dybbølsbro Intersection in Copenhagen, Transport Findings 56683 (2022)

ML in a (tiny)Nutshell

- Tens of thousands of machine learning algorithms
  - Hundreds new every year

- Decades of ML research oversimplified:
  - All of Machine Learning:
  - Learn a mapping from input to output f: X → Y
  - X: emails, Y: {spam, notspam}

Slide Credit: Pedro Domingos

ML in a Nutshell

- Input: x (images, text, emails…)
- Output: y (spam or non-spam…)
- (Unknown) Target Function
  - $f: X \rightarrow Y$ (the "true" mapping / reality)
- Data
  - $(x_1,y_1), (x_2,y_2), \ldots, (x_N,y_N)$
- Model / Hypothesis Class
  - $g: X \rightarrow Y$
  - $y = g(x) = \text{sign}(w^T x)$

Machine Learning Components

- Representation / Model Class
- Optimization
- Objective Function / Evaluation

Machine Learning Models

- Decision trees
- Sets of rules / Logic programs
- Instance-based Models
- Graphical models (Bayes/Markov nets)
- Neural networks
- Support vector machines
- Model ensembles
- And many others

You will cover these in detail in the **Machine Learning** Module

Optimization

- Discrete/Combinatorial optimization
  - Greedy search
  - Graph algorithms (cuts, flows, etc)
- Continuous optimization
  - Convex/Non-convex optimization (gradient descent)
  - Linear programming

Evaluation / Objective Function

- Accuracy
- Precision and recall
- Squared error
- Likelihood
- Posterior probability
- Cost / Utility
- Margin
- Entropy
- K-L divergence
- Etc.

Source: Pedro Domingos

Learning Settings

- Supervised learning
  - Training data includes desired outputs
- Unsupervised learning
  - Training data does not include desired outputs
- Weakly or Semi-supervised learning
  - Training data includes a few desired outputs
- Self-Supervised Learning
  - Model supervises itself
- Reinforcement learning
  - Rewards from sequence of actions

CMU Portugal Academy

Supervised Learning Example

Credit Card Fraud Detection

Binary classification (Fraud, Not Fraud)

trans_date_trans...	merchant	category	amt	gender	street	zip	unix_time	merch_lat	merch_long
2020-06-21 2020-12-31	693 unique values	gas_transport 10% grocery_pos 9% Other (446796) 80%	1 22.8k	F 55% M 45%	924 unique values	1257 99.9k	1375829067.16 - 1376163417.34 Count: 11,371 1.37b 1.39b	19 66.7	-167 -67
2020-06-21 12:14:25	fraud_Kirlin and Sons	personal_care	2.86	M	351 Darlene Green	29209	1371816865	33.986391	-81.200714
2020-06-21 12:14:33	fraud_Sporer-Keebler	personal_care	29.84	F	3638 Marsh Union	84002	1371816873	39.450497999999996	-109.960431
2020-06-21 12:14:53	fraud_Swaniawski, Nitzsche and Welch	health_fitness	41.28	F	9333 Valentine Point	11710	1371816893	40.49581	-74.196111
2020-06-21 12:15:15	fraud_Haley Group	misc_pos	60.05	M	32941 Krystal Mill Apt. 552	32780	1371816915	28.812397999999998	-80.883061
2020-06-21 12:15:17	fraud_Johnston-Casper	travel	3.19	M	5783 Evan Roads Apt. 465	49632	1371816917	44.959148	-85.884734

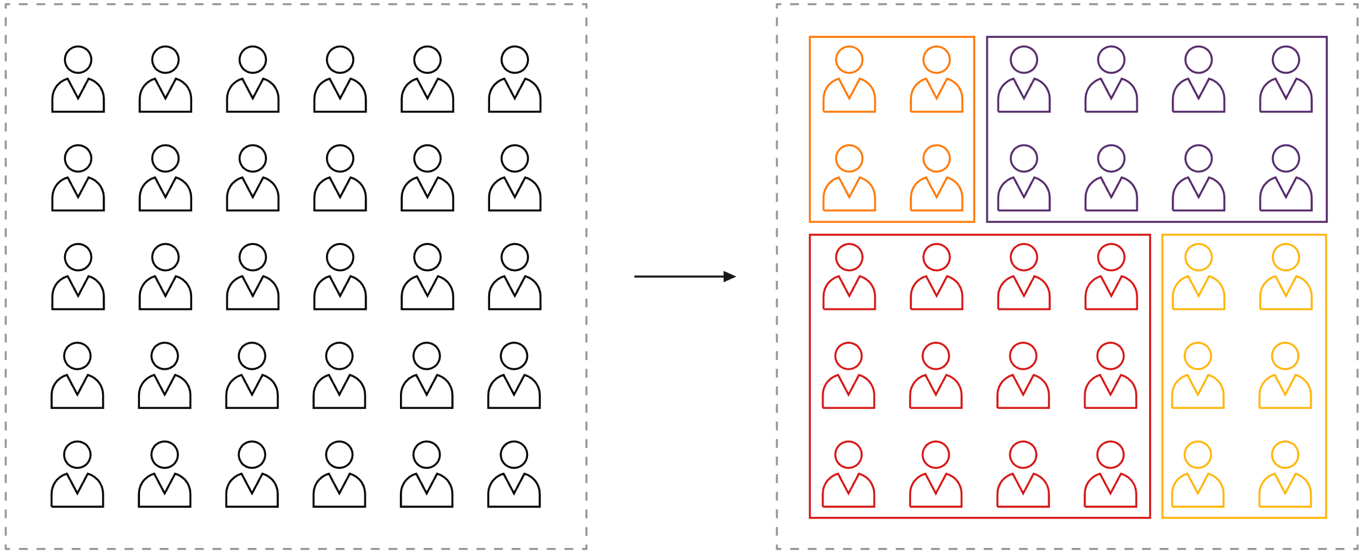
Dataset source: https://www.kaggle.com/datasets/kartik2112/fraud-detection

CMU Portugal Academy

Unsupervised Learning Example

Market Segmentation

Learning from data without guidance



Source: https://www.univio.com/blog/machine-learning-and-customer-segmentation-meet-the-perfect-couple/

CMU Portugal Academy

Tasks

Supervised Learning

x → Classification → y Discrete

x → Regression → y Continuous

Unsupervised Learning

x → Clustering → y Discrete ID

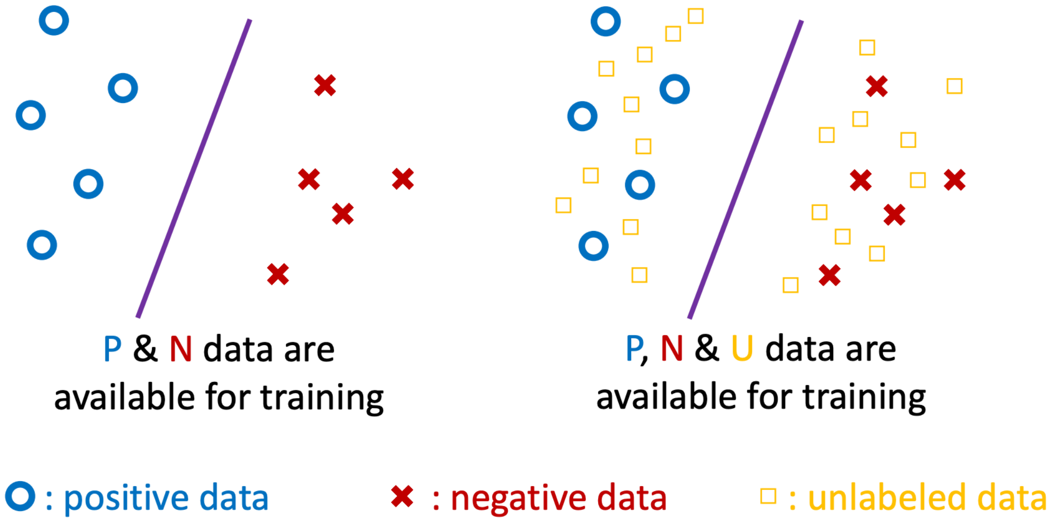
x → Dimensionality Reduction → y Continuous

Weakly-Supervised Learning Example

Breast Cancer Tumor Prediction - Scarce labeled data!

Scenarios:

- Low-quality labels
- Proxy process to label data

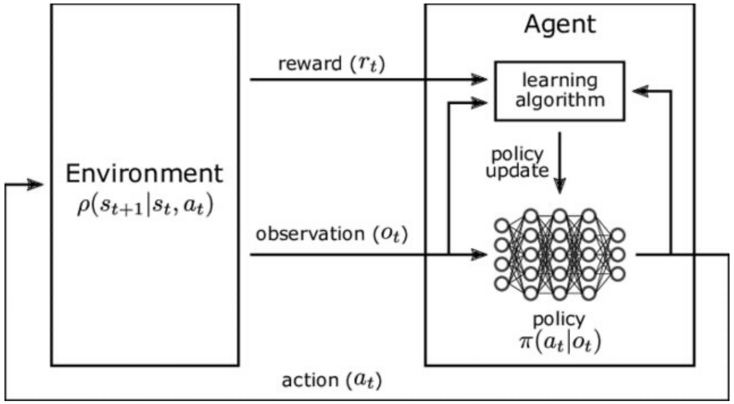


Source: https://niug1984.github.io/paper/niu_tdlw2018.pdf

Reinforcement Learning

An agent:

- Interacts with an Environment
- Learns by Trial-and-Error
- Receives rewards from actions



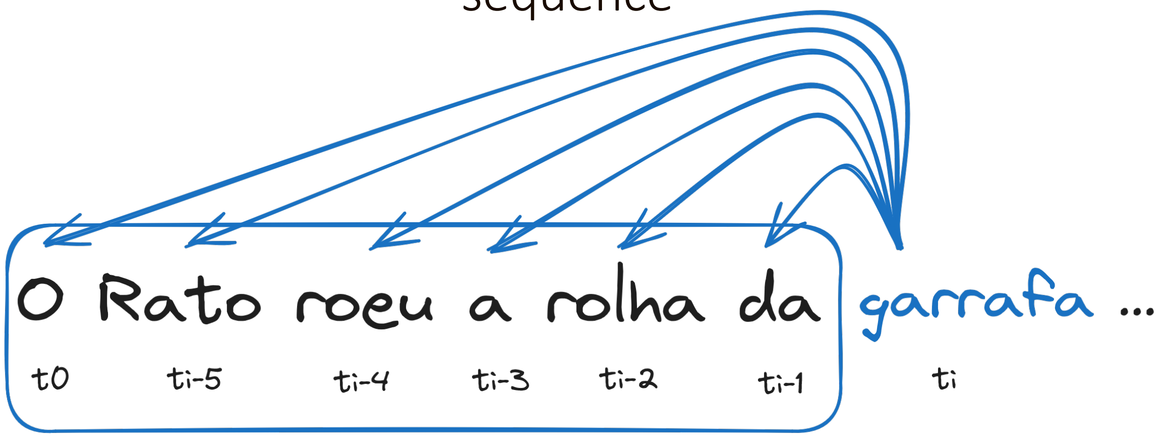
Source: https://bernardmarr.com/how-tesla-is-using-artificial-intelligence-to-create-the-autonomous-cars-of-the-future/

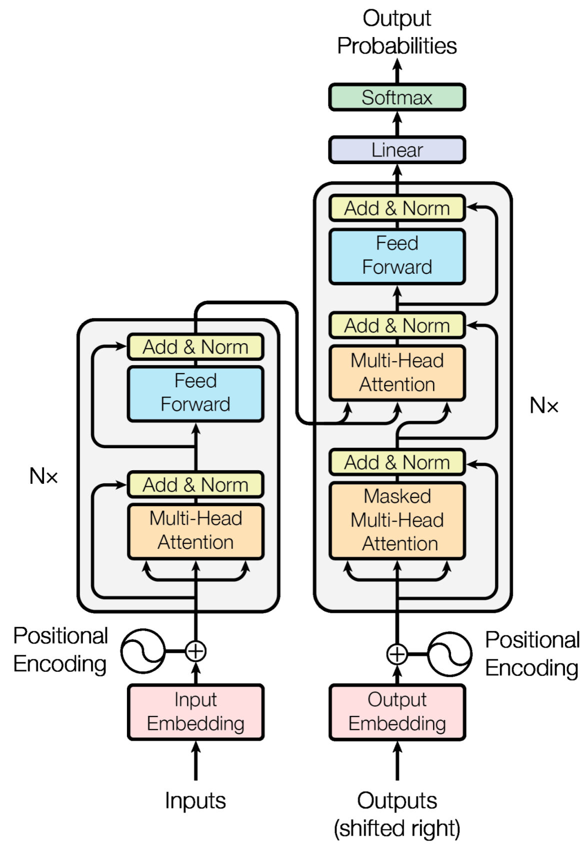
Self-supervised Learning

Causal Language Modeling Objective

$$\boldsymbol{P(t_1,t_2,\ldots,t_W)} = P(t_1) \cdot P(t_2|t_1) \cdot P(t_3|t_2,t_1) \cdot \ldots \cdot P(t_W \mid t_{W-1},\ldots,t_1),$$

$\boldsymbol{t_i}$ is the *i*-th word, and **W** is the total amount of words in a sequence

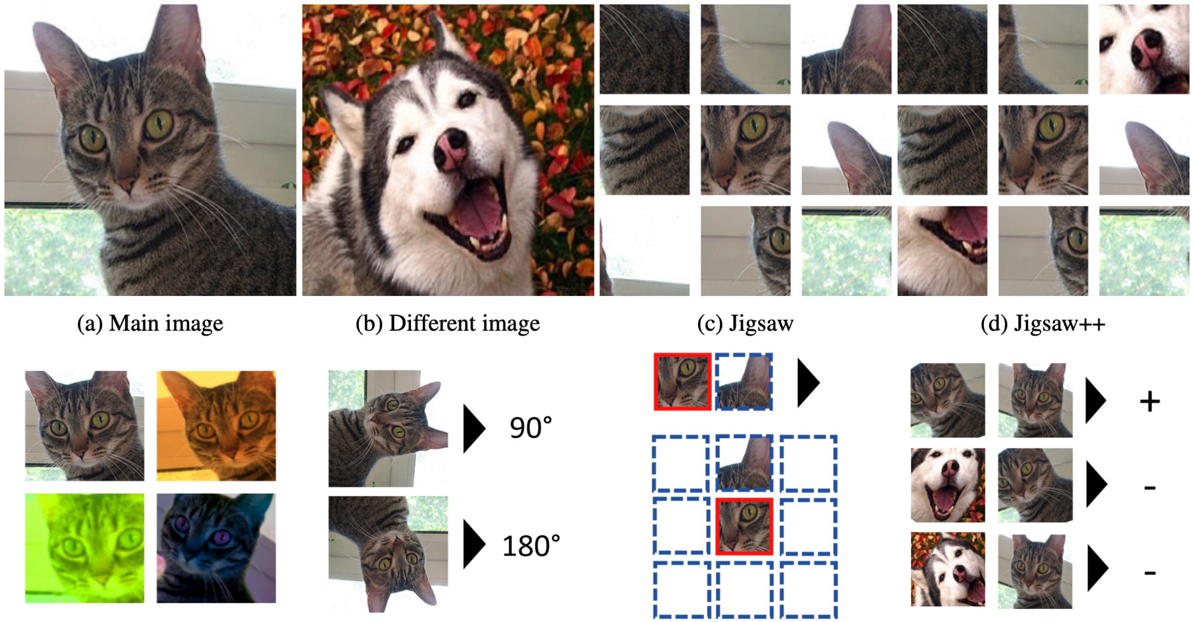




Self-supervised Learning

Pretext Tasks

Learning data representations



(a) Main image (b) Different image (c) Jigsaw (d) Jigsaw++

(e) Exemplar (f) Rotation (g) Context (h) Contrastive Learning

Schmarje, L., Santarossa, M., Schroder, S., & Koch, R. A Survey on Semi-, Self- and Unsupervised Learning for Image Classification. *IEEE Access, 2020.*

Supervised Learning

- Input: x (images, text, emails…)
- Output: y (spam or non-spam…)

- (Unknown) Target Function
  - $f: X \rightarrow Y$ (the "true" mapping / reality)

- Data
  - $\{ (x_1, y_1), (x_2, y_2), \ldots, (x_N, y_N) \}$

- Model / Hypothesis Class
  - $H = \{h: X \rightarrow Y\}$
  - e.g. $y = h(x) = \text{sign}(w^T x)$

- Loss Function
  - How good is a model w.r.t. my data D?

- Learning = Search in hypothesis space
  - Find best h in model class.

Source: Pedro Domingos

Appropriate Applications for Supervised Learning

- **Situations where there is no human expert**

  $\mathbf{x}$: Bond graph for a new molecule.

  $f(\mathbf{x})$: Predicted binding strength to AIDS protease molecule.

- **Situations where humans can perform the task but can't describe how they do it.**

  $\mathbf{x}$: Bitmap picture of hand-written character

  $f(\mathbf{x})$: Ascii code of the character

- **Situations where the desired function is changing frequently**

  $\mathbf{x}$: Description of stock prices and trades for last 10 days.

  $f(\mathbf{x})$: Recommended stock transactions

- **Situations where each user needs a customized function f**

  $\mathbf{x}$: Incoming email message.

  $f(\mathbf{x})$: Importance score for presenting to user (or deleting without presenting).

Feature Extraction and Engineering

Obtaining feature vectors:

- Numeric representation of a sample
- Implies the application of translation function

trans_date_trans...	merchant	category	amt	gender	street	zip	unix_time	merch_lat	merch_long
2020-06-21 2020-12-31	693 unique values	gas_transport 10% grocery_pos 9% Other (446796) 80%	1 22.8k	F 55% M 45%	924 unique values	1257 99.9k	1375829067.16 - 1376163417.34 Count: 11,371 1.37b 1.39b	19 66.7	-167 -67
2020-06-21 12:14:25	fraud_Kirlin and Sons	personal_care	2.86	M	351 Darlene Green	29209	1371816865	33.986391	-81.200714
2020-06-21 12:14:33	fraud_Sporer-Keebler	personal_care	29.84	F	3638 Marsh Union	84002	1371816873	39.450497999999996	-109.960431
2020-06-21 12:14:53	fraud_Swaniawski, Nitzsche and Welch	health_fitness	41.28	F	9333 Valentine Point	11710	1371816893	40.49581	-74.196111
2020-06-21 12:15:15	fraud_Haley Group	misc_pos	60.05	M	32941 Krystal Mill Apt. 552	32780	1371816915	28.812397999999998	-80.883061
2020-06-21 12:15:17	fraud_Johnston-Casper	travel	3.19	M	5783 Evan Roads Apt. 465	49632	1371816917	44.959148	-85.884734

Feature Extraction and Engineering

Obtaining feature vectors:

- Numeric representation of a sample
- Implies the application of translation function

$$t(x_i) \rightarrow x_i'$$

Might introduce noise or an inaccurate approximation

trans_date_trans...	merchant	category	amt	gender	street	zip	unix_time	merch_lat	merch_long
2020-06-21 2020-12-31	693 unique values	gas_transport 10% grocery_pos 9% Other (446796) 80%	1 22.8k	F 55% M 45%	924 unique values	1257 99.9k	1375829067.16 - 1376163417.34 Count: 11,371 1.37b 1.39b	19 66.7	-167 -67
2020-06-21 12:14:25	fraud_Kirlin and Sons	personal_care	2.86	M	351 Darlene Green	29209	1371816865	33.986391	-81.200714
2020-06-21 12:14:33	fraud_Sporer-Keebler	personal_care	29.84	F	3638 Marsh Union	84002	1371816873	39.450497999999996	-109.960431
2020-06-21 12:14:53	fraud_Swaniawski, Nitzsche and Welch	health_fitness	41.28	F	9333 Valentine Point	11710	1371816893	40.49581	-74.196111
2020-06-21 12:15:15	fraud_Haley Group	misc_pos	60.05	M	32941 Krystal Mill Apt. 552	32780	1371816915	28.812397999999998	-80.883061
2020-06-21 12:15:17	fraud_Johnston-Casper	travel	3.19	M	5783 Evan Roads Apt. 465	49632	1371816917	44.959148	-85.884734

CMU Portugal Academy

Feature Extraction and Engineering

Feature Transformations $t(x_i) \rightarrow x_i'$

Gender: {F, M} -> {0, 1}

trans_date_trans...	merchant	category	amt	gender	street	zip	unix_time	merch_lat	merch_long
2020-06-21 2020-12-31	693 unique values	gas_transport 10% grocery_pos 9% Other (446796) 80%	1 22.8k	F 55% M 45%	924 unique values	1257 99.9k	1375829067.16 - 1376163417.34 Count: 11,371 1.37b 1.39b	19 66.7	-167 -67
2020-06-21 12:14:25	fraud_Kirlin and Sons	personal_care	2.86	M	351 Darlene Green	29209	1371816865	33.986391	-81.200714
2020-06-21 12:14:33	fraud_Sporer-Keebler	personal_care	29.84	F	3638 Marsh Union	84002	1371816873	39.450497999999996	-109.960431
2020-06-21 12:14:53	fraud_Swaniawski, Nitzsche and Welch	health_fitness	41.28	F	9333 Valentine Point	11710	1371816893	40.49581	-74.196111
2020-06-21 12:15:15	fraud_Haley Group	misc_pos	60.05	M	32941 Krystal Mill Apt. 552	32780	1371816915	28.812397999999998	-80.883061
2020-06-21 12:15:17	fraud_Johnston-Casper	travel	3.19	M	5783 Evan Roads Apt. 465	49632	1371816917	44.959148	-85.884734

CMU Portugal Academy

Feature Extraction and Engineering

Feature Transformations $t(x_i) \rightarrow x_i'$

Gender: {F, M} -> {0, 1}

Category: {travel, personal_care, ...} -> {0, 1, 2, ... }

trans_date_trans...	merchant	category	amt	gender	street	zip	unix_time	merch_lat	merch_long
2020-06-21 2020-12-31	693 unique values	gas_transport 10% grocery_pos 9% Other (446796) 80%	1 22.8k	F 55% M 45%	924 unique values	1257 99.9k	1375829067.16 - 1376163417.34 Count: 11,371 1.37b 1.39b	19 66.7	-167 -67
2020-06-21 12:14:25	fraud_Kirlin and Sons	personal_care	2.86	M	351 Darlene Green	29209	1371816865	33.986391	-81.200714
2020-06-21 12:14:33	fraud_Sporer-Keebler	personal_care	29.84	F	3638 Marsh Union	84002	1371816873	39.450497999999996	-109.960431
2020-06-21 12:14:53	fraud_Swaniawski, Nitzsche and Welch	health_fitness	41.28	F	9333 Valentine Point	11710	1371816893	40.49581	-74.196111
2020-06-21 12:15:15	fraud_Haley Group	misc_pos	60.05	M	32941 Krystal Mill Apt. 552	32780	1371816915	28.812397999999998	-80.883061
2020-06-21 12:15:17	fraud_Johnston-Casper	travel	3.19	M	5783 Evan Roads Apt. 465	49632	1371816917	44.959148	-85.884734

CMU Portugal Academy

Feature Extraction and Engineering

Feature Transformations $t(x_i) \rightarrow x_i'$

Unix Timestamp -> Seconds since January 01, 1970.

Unix_time: [0, 2147483647] -> ???

Assuming 32bit integers.

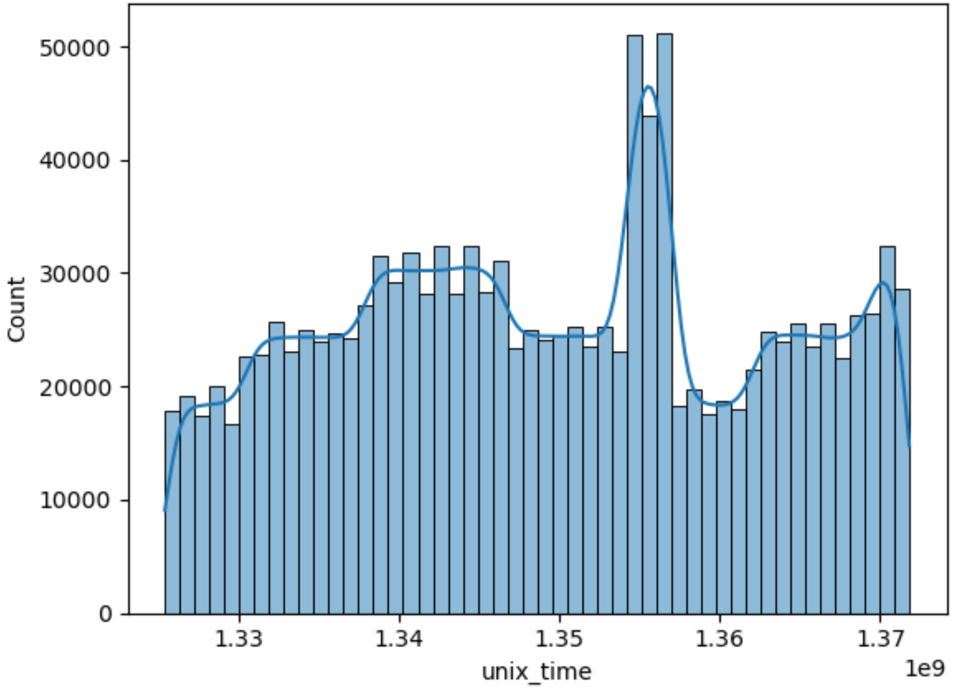
trans_date_trans...	merchant	category	amt	gender	street	zip	unix_time	merch_lat	merch_long
2020-06-21 2020-12-31	693 unique values	gas_transport 10% grocery_pos 9% Other (446796) 80%	1 22.8k	F 55% M 45%	924 unique values	1257 99.9k	Count: 11,371 1.37b 1.39b	19 66.7	-167 -67
2020-06-21 12:14:25	fraud_Kirlin and Sons	personal_care	2.86	M	351 Darlene Green	29209	1371816865	33.986391	-81.200714
2020-06-21 12:14:33	fraud_Sporer-Keebler	personal_care	29.84	F	3638 Marsh Union	84002	1371816873	39.450497999999996	-109.960431
2020-06-21 12:14:53	fraud_Swaniawski, Nitzsche and Welch	health_fitness	41.28	F	9333 Valentine Point	11710	1371816893	40.49581	-74.196111
2020-06-21 12:15:15	fraud_Haley Group	misc_pos	60.05	M	32941 Krystal Mill Apt. 552	32780	1371816915	28.812397999999998	-80.883061
2020-06-21 12:15:17	fraud_Johnston-Casper	travel	3.19	M	5783 Evan Roads Apt. 465	49632	1371816917	44.959148	-85.884734

Feature Extraction and Engineering

Unix Timestamp -> Seconds since January 01, 1970.

Unix_time: [0, 2147483647] -> ???

Assuming 32bit integers.



What should we do?

CMU Portugal Academy

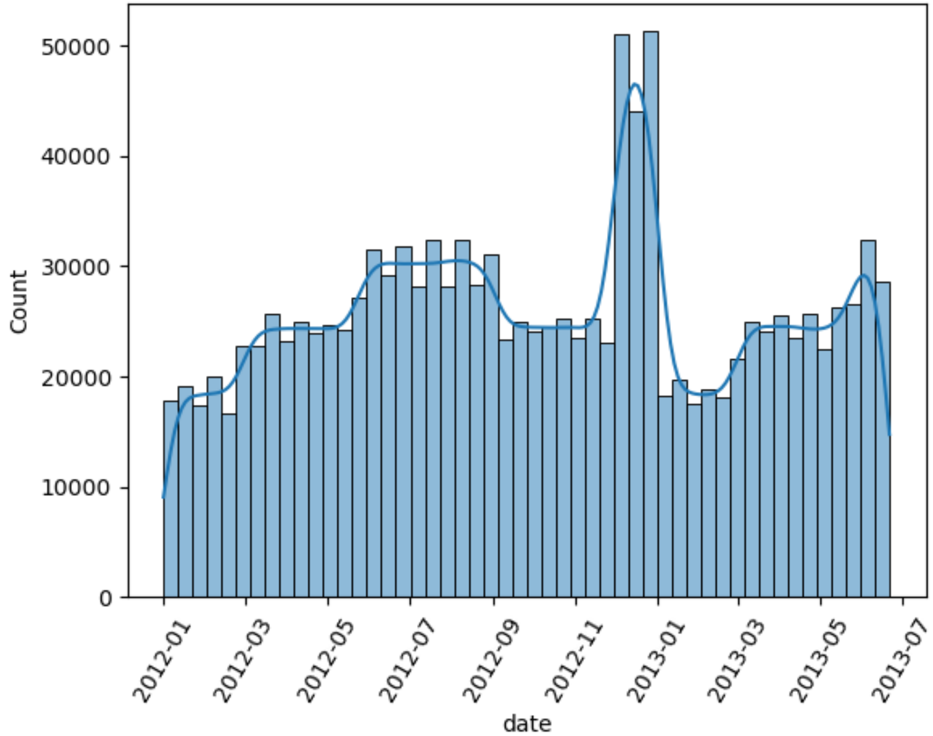
Feature Extraction and Engineering

Unix Timestamp -> Seconds since January 01, 1970.

Unix_time: [0, 2147483647] -> ???

Assuming 32bit integers.

The domain is actually quite narrow!



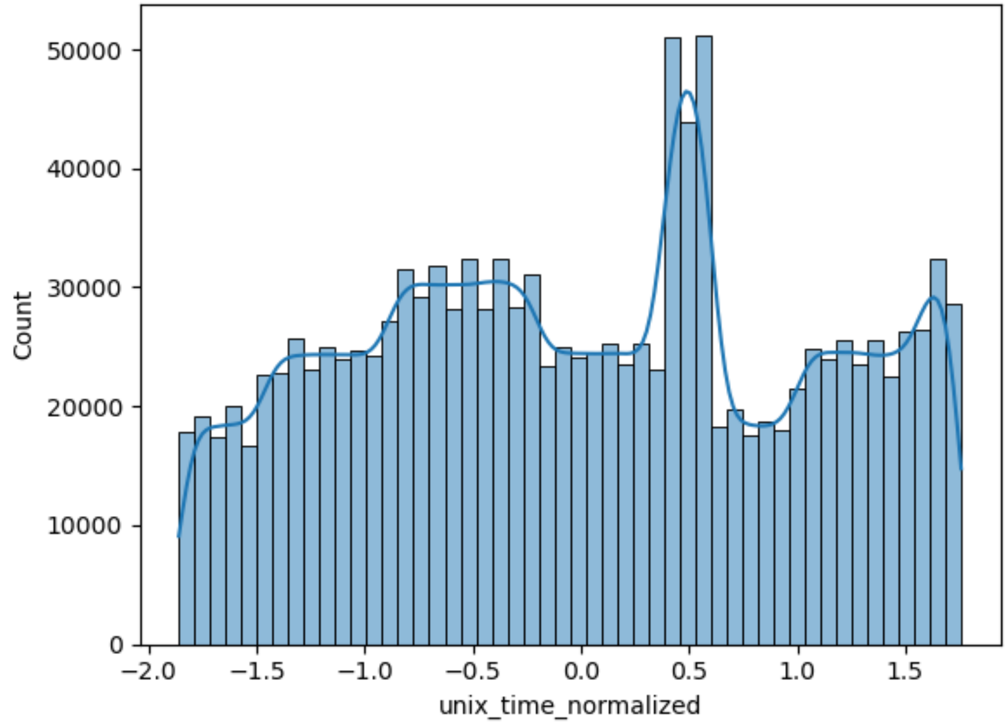
Carnegie Mellon Portugal
TÉCNICO LISBOA
Ciências ULisboa
NOVA SCHOOL OF SCIENCE & TECHNOLOGY

Feature Extraction and Engineering

Normalization:

- Standard Scaling (z-scoring):

$$x_i' = \frac{x_i - \mu}{\sigma}$$ Maps to the range $]-\infty, +\infty[$



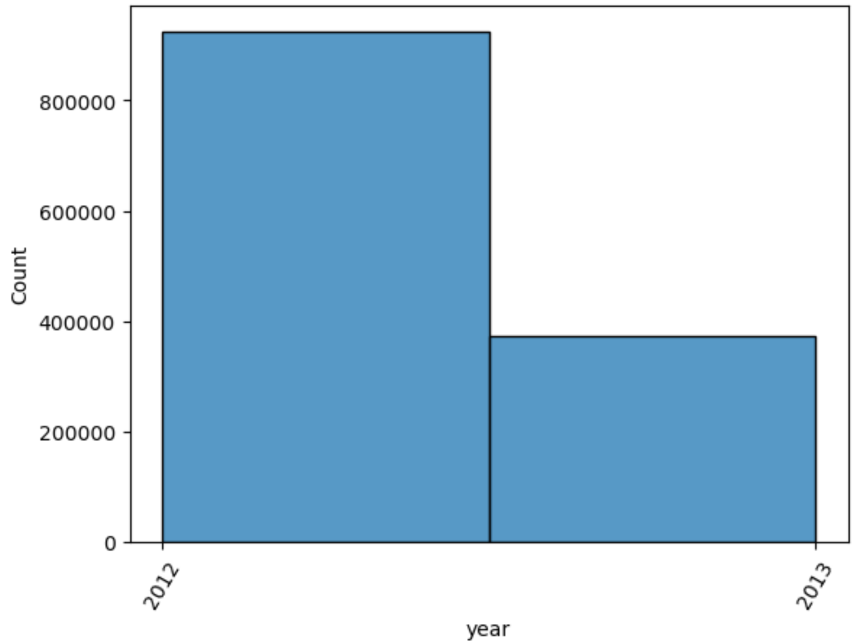
- Min-max Normalization:

$$x_i' = \frac{x_i - x_{min}}{x_{max} - x_{min}}$$ Maps to the range $[0, 1]$

Feature Extraction and Engineering

Other Approaches:

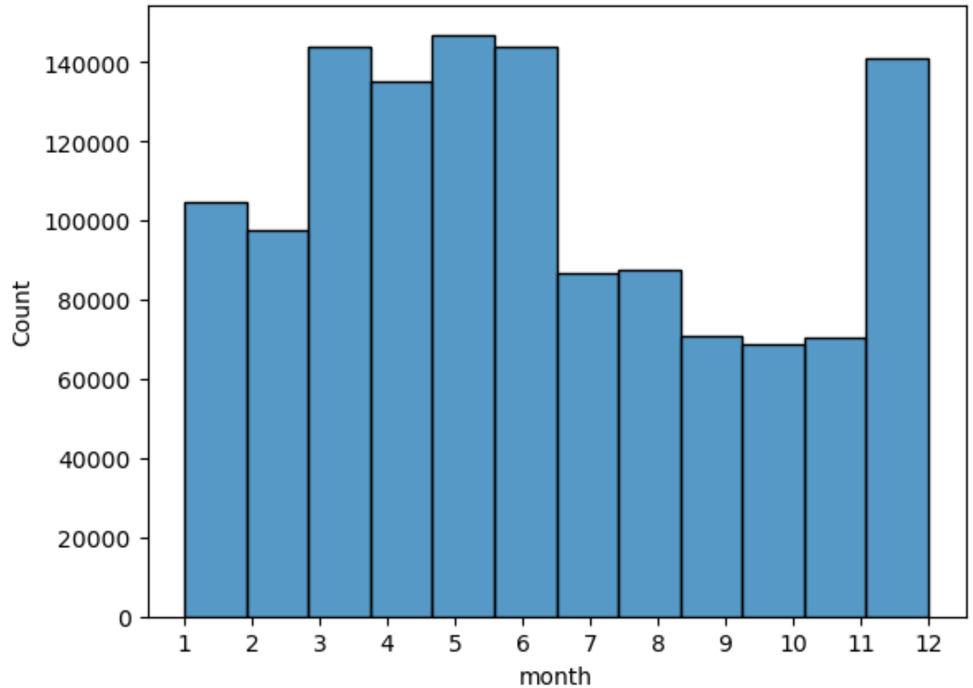
- Discretization through binning
- Discretization through clustering
- Granularity / Precision changes (e.g. use only the date year)



Feature Extraction and Engineering

Other Approaches:

- Discretization through binning
- Discretization through clustering
- Granularity / Precision changes (e.g. use only the date year)
- New features derivation (e.g. difference between locations in kilometers)



Linear Models – Linear Regression

We assume that the relationship between features x and target y is approximately linear

- The conditional mean $E[Y|X = x]$ can be expressed as a weighted sum of the features x.
- We assume there will some well behaved noise (e.g. following a Gaussian distribution)

Linear Models – Linear Regression

For example, the price can be defined as a weighted sum of features:

$$price = \; w_{area} * area + w_{age} * age + \boxed{b}$$ Bias/intercept/offset

Generalizing: $$\hat{y} = w_1 x_1 + \cdots + w_d x_d + b$$

If we represent all features into a vector $\boldsymbol{x} \in \mathbb{R}^d$ and all weights into a vector $\boldsymbol{w} \in \mathbb{R}^d$:

$$\hat{y} = \mathbf{w}^\top \mathbf{x} + b$$

Usually, we represent the whole dataset of n example as a design matrix $\boldsymbol{X} \in \mathbb{R}^{n \times d}$.
Then, it becomes:

$$\hat{\mathbf{y}} = \mathbf{X}\mathbf{w} + b$$

CMU Portugal Academy

Linear Models – Linear Regression

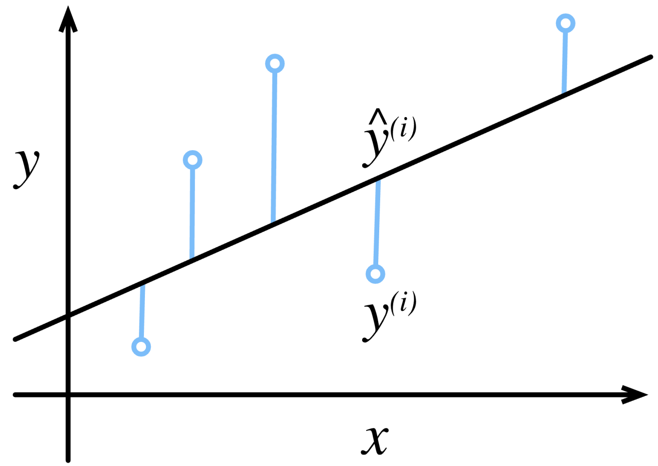
Regression Loss function: $$l^{(i)}(\mathbf{w}, b) = \frac{1}{2}\left(\hat{y}^{(i)} - y^{(i)}\right)^2$$

Generalizing to the whole dataset:

$$L(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} l^{(i)}(\mathbf{w}, b) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{2}\left(\mathbf{w}^\top \mathbf{x}^{(i)} + b - y^{(i)}\right)^2$$

Find the optimal weights:

$$\mathbf{w}^*, b^* = \underset{\mathbf{w}, b}{\operatorname{argmin}}\ L(\mathbf{w}, b)$$



CMU Portugal Academy

Linear Regression – Closed Form Solution

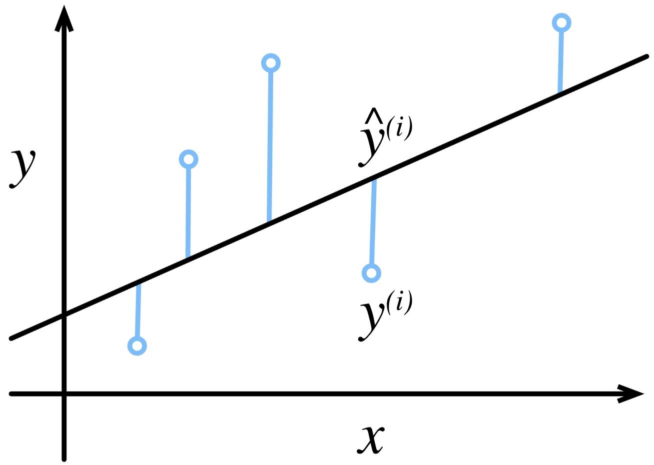
Represent minimization problem as: $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$ → Bias is added to **w** and a 1s column is added to X

We take the derivative, w.r.t. w, to find the minimum:

$$\partial_{\mathbf{w}} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = 2\mathbf{X}^\top(\mathbf{X}\mathbf{w} - \mathbf{y}) = 0 \text{ and hence } \mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\mathbf{w}$$

Solving for w:

$$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$$



The solution is unique when the matrix $X^T X$ is invertible, i.e., the columns of the design matrix are linearly independent (Golub and Van Loan, 1996).

Golub, G. H., & Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins University Press.

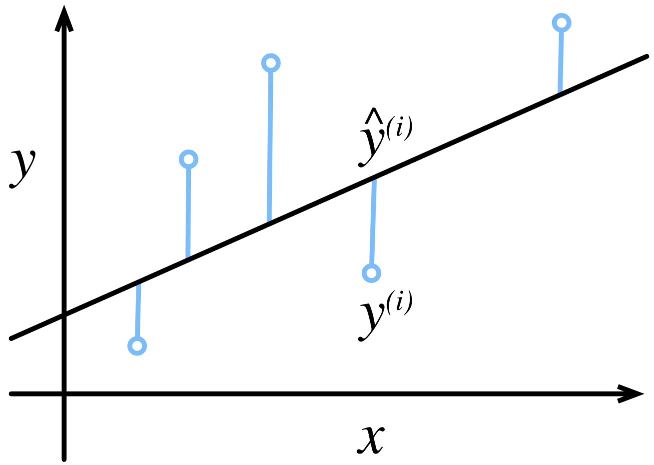
Linear Regression – Closed-Form Solution

Represent minimization problem as: $\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2$

Bias is added to **w** and a 1s column is added to X

Solving for w: $$\mathbf{w}^* = (\mathbf{X}^\top \mathbf{X})^{-1}\mathbf{X}^\top \mathbf{y}$$

- Linear regression has a closed-form solution.
- More complex models (e.g. Neural Networks) don't.
- In the Machine Learning module you will learn about other optimization approaches, like Stochastic Gradient Descent.



Logistic Regression

Vanilla linear regression is not so suitable to classification problems:

- Targets are continuous values, along a straight line
- Decision boundary too sensitive to outliers

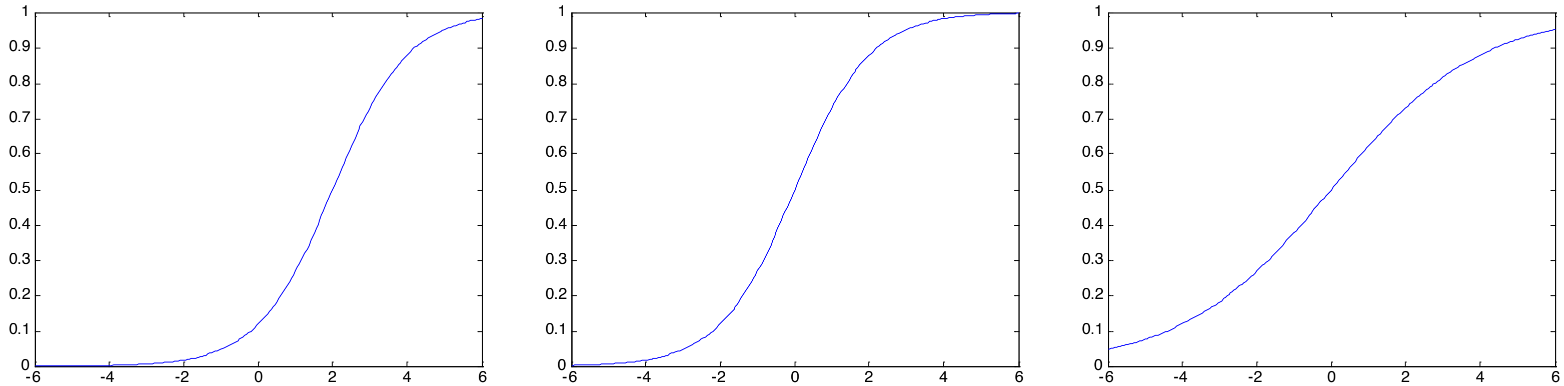
For a classification problem, we want to obtain a probability $\boldsymbol{P(Y = label|X = x)}$.

We can squash the linear regression outputs to the [0,1]
range with the sigmoid function!

Logistic Regression – Sigmoid

$$\sigma(w_0 + \sum_i w_i x_i) = \frac{1}{1 + e^{-w_0 - \sum_i w_i x_i}} = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{X}}}$$

$w_0=2$, $w_1=1$ $\qquad$ $w_0=0$, $w_1=1$ $\qquad$ $w_0=0$, $w_1=0.5$



Source: Carlos Guestrin

Logistic Regression – Binary Classification

We can assume the following:

- Class 1 corresponds to y = 1
- Class 2 corresponds to y = 0

$$\boldsymbol{P}(\boldsymbol{Y} = \mathbf{1}|\boldsymbol{X}, \boldsymbol{w}) = \sigma(\boldsymbol{w}^T * \boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{w}^T * \boldsymbol{X}}}$$

$$\boldsymbol{P}(\boldsymbol{Y} = \mathbf{0}|\boldsymbol{X}, \boldsymbol{w}) = 1 - P(Y = 1|X, w)$$

For a classification problem, we want to obtain a probability $\boldsymbol{P}(\boldsymbol{Y} = \boldsymbol{label}|\boldsymbol{X} = \boldsymbol{x})$.

We can squash the linear regression outputs to the [0,1] range with the sigmoid function!

Logistic Regression – Binary Classification

We can assume the following:

- Class 1 corresponds to y = 1
- Class 2 corresponds to y = 0

$$\boldsymbol{P}(\boldsymbol{Y} = \mathbf{1}|\boldsymbol{X}, \boldsymbol{w}) = \sigma(\boldsymbol{w}^T * \boldsymbol{X}) = \frac{1}{1 + e^{-\boldsymbol{w}^T * \boldsymbol{X}}}$$

$$\boldsymbol{P}(\boldsymbol{Y} = \mathbf{0}|\boldsymbol{X}, \boldsymbol{w}) = 1 - P(Y = 1|X, w)$$

To find w, logistic regressions uses negative log-likelihood (or cross-entropy):

$$L(\mathbf{w}) = \sum_j -ln\ P(y^j|x^j, \mathbf{w})$$

$$L(\mathbf{w}) = \sum_j -y^j \cdot ln\ P(y^j = 1|x^j, \mathbf{w}) - (1 - y^j) \cdot ln\ P(y^j = 0|x^j, \mathbf{w})$$

Logistic Regression – Binary Classification

To find w, logistic regression uses negative log-likelihood (or cross-entropy):

$$L(\mathbf{w}) = \sum_j -y^j \cdot ln\ P(y^j = 1|x^j, \mathbf{w}) - (1 - y^j) \cdot ln\ P(y^j = 0|x^j, \mathbf{w})$$

Substituting now with the sigmoid, it becomes:

$$L(\mathbf{w}) = \sum_j -y^j \cdot ln\ \sigma(\mathbf{w}^T x^j) - (1 - y^j) \cdot ln\ (1 - \sigma(\mathbf{w}^T x^j))$$

This loss does not have a closed-form solution, but is concave
-> Can be optimized with Gradient Descent

Assessing Model Performance - Metrics

It is critical to use quantitative metrics to evaluate machine learning models

- The loss function is not enough
- Different metrics provide different perspectives of models' performance!

Classification → Accuracy/Precision/Recall/F1-score, ROC Curves ,etc.

Regression → Root Mean Squared Error, Mean Absolute Error, etc.

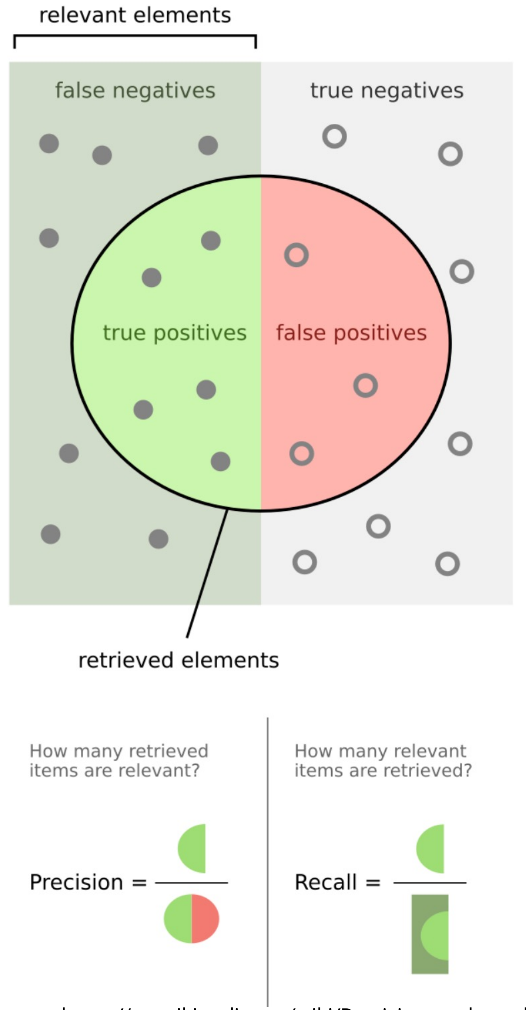
Assessing Model Performance - Metrics

$$\text{Accuracy} = \frac{correct\ classifications}{total\ classifications} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}$$

Assessing Model Performance - Metrics

$$Precision = \frac{True\ Positives}{Predicted\ Positives} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}$$

$$Recall/Sensitivity = \frac{True\ Positives}{Actual\ Positives} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$



Source: https://en.wikipedia.org/wiki/Precision_and_recall

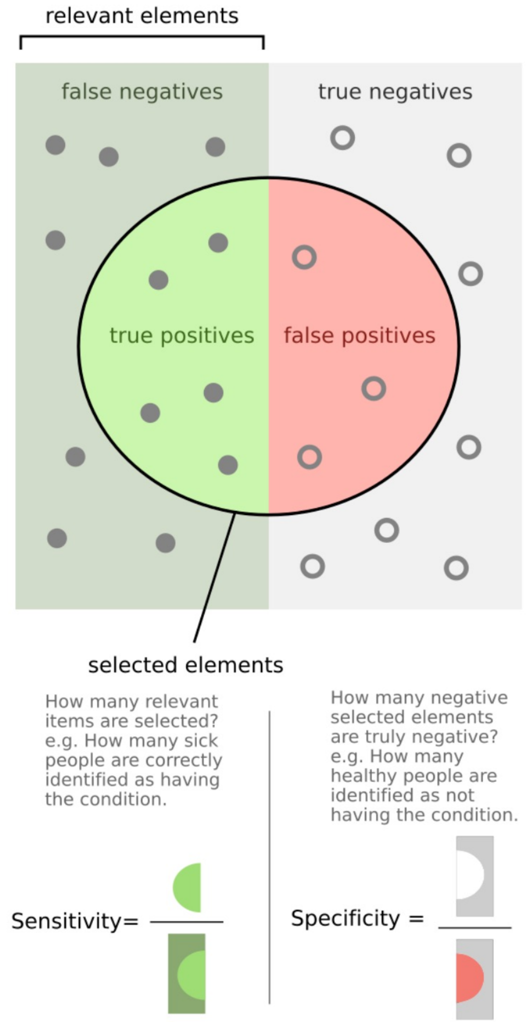
Assessing Model Performance - Metrics

$$\underbrace{Recall/Sensitivity}_{} = \frac{True\ Positives}{Total\ Positives} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

True **Positive** Rate (TPR)

$$\underbrace{Specificity}_{} = \frac{True\ Negatives}{Total\ Negatives} = \frac{TN}{TN + \mathrm{FP}}$$

True **Negative** Rate (TNR)



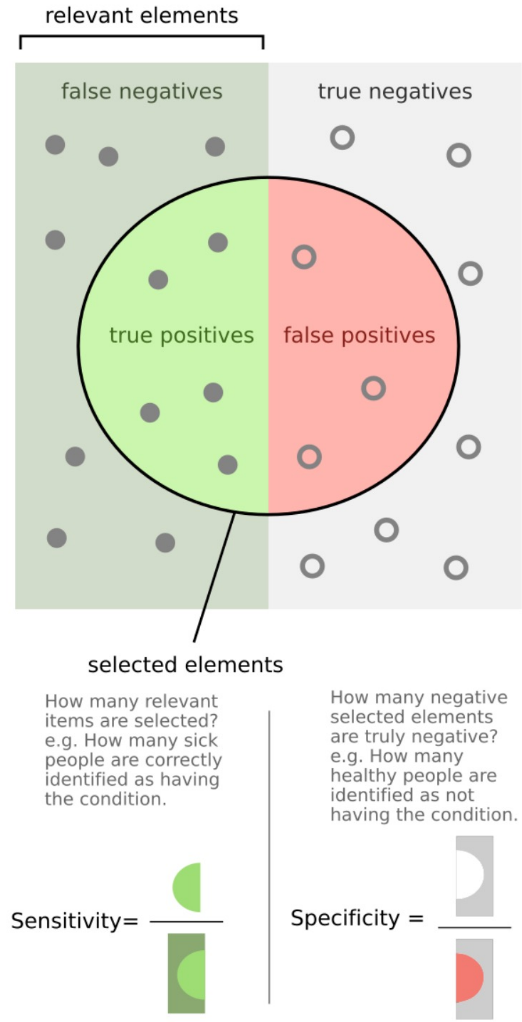
Source: https://en.wikipedia.org/wiki/Sensitivity_and_specificity

Assessing Model Performance - Metrics

F-score - Harmonic mean between precision and recall:

$$F_1 = \frac{1}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall}$$

Aggregates (symmetrically) information from both metrics.



CMU Portugal Academy

Assessing Model Performance - Metrics

Confusion Matrix

Provides a more detailed view of a classification model performance

		Predicted condition	
	Total population = P + N	Positive (PP)	Negative (PN)
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Why multiple perspectives?

Example 1:

- Imbalanced dataset, 99% of examples are Spam, 1% are not.
- Our model always predicts Spam. What is the accuracy?

It is okay if a Spam email goes to our inbox -> Low Sensitivity
It is not okay if a Non-spam email gets filtered! -> High Specificity

Precision!

Why multiple perspectives?

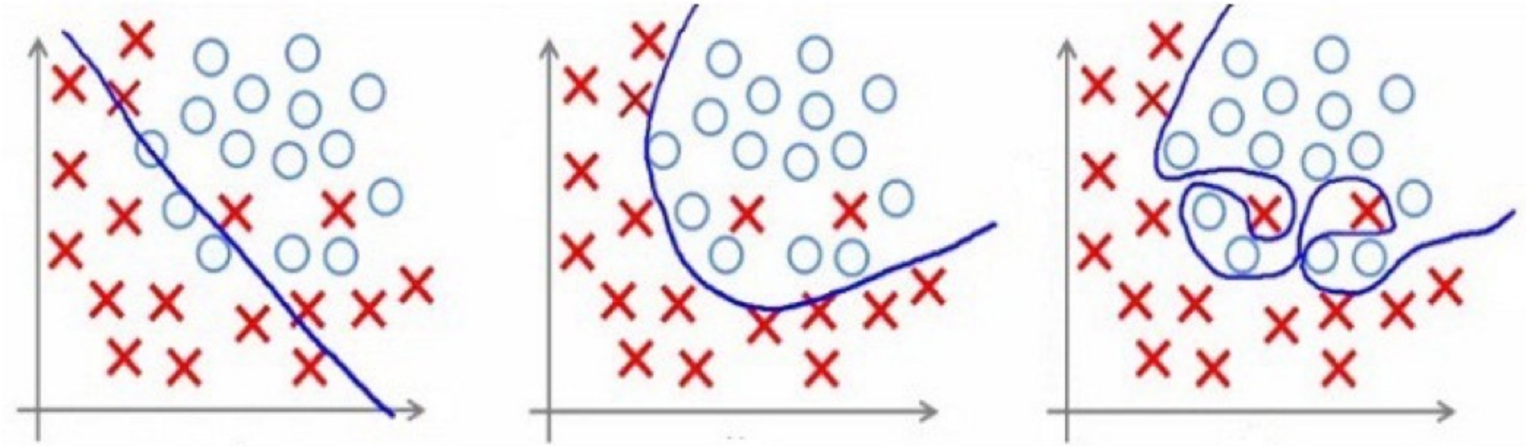
Example 2:

- Automatic Border Control Checking

Terrorists shouldn’t go through -> High Sensitivity
False alarms (False positives) are not a big deal -> Low Specificity

CMU Portugal Academy

Overfitting vs. Underfitting



Underfitting

Too simple to explain variance

High Loss, High Bias

Appropriate

Overfitting

Low generalization

Low Loss, High Variance

Hands-On Session!

[Course Shared Folder]

CMU Portugal Academy

CMU Portugal
Advanced Training Program
Foundations of Data Science

DAVID SEMEDO

RAFAEL FERREIRA

NOVA SCHOOL OF SCIENCE AND TECHNOLOGY

Carnegie Mellon Portugal · TÉCNICO LISBOA · Ciências ULisboa · NOVA School of Science & Technology

Email: rah.ferreira@campus.fct.unl.pt

Sources

These slides contain adapted materials from the following sources:

- Stefan Lee, Introduction to Machine Learning, Virginia Tech
- Dive into Deep Learning Book.