Machine Learning - Simple Models

ImpactDeal 2022

What is Machine Learning?

Examples:

- **Recommenders**:
  - Try to predict users’ preferences, based on historical data.
- **Image recognition**:
  - Learn properties and relationships of pixels in images.
- **Fraud detection**:
  - classify if a transaction is fraudulent or not

What is Machine Learning?

	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	class
0	25.0	Private	226802.0	11th	7.0	Never-married	Machine-op-inspct	Own-child	Black	Male	0.0	0.0	40.0	United-States	<=50K
1	38.0	Private	89814.0	HS-grad	9.0	Married-civ-spouse	Farming-fishing	Husband	White	Male	0.0	0.0	50.0	United-States	<=50K
2	28.0	Local-gov	336951.0	Assoc-acdm	12.0	Married-civ-spouse	Protective-serv	Husband	White	Male	0.0	0.0	40.0	United-States	>50K
3	44.0	Private	160323.0	Some-college	10.0	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688.0	0.0	40.0	United-States	>50K
4	18.0	NaN	103497.0	Some-college	10.0	Never-married	NaN	Own-child	White	Female	0.0	0.0	30.0	United-States	<=50K

What is Machine Learning?

Number of rented bikes per day

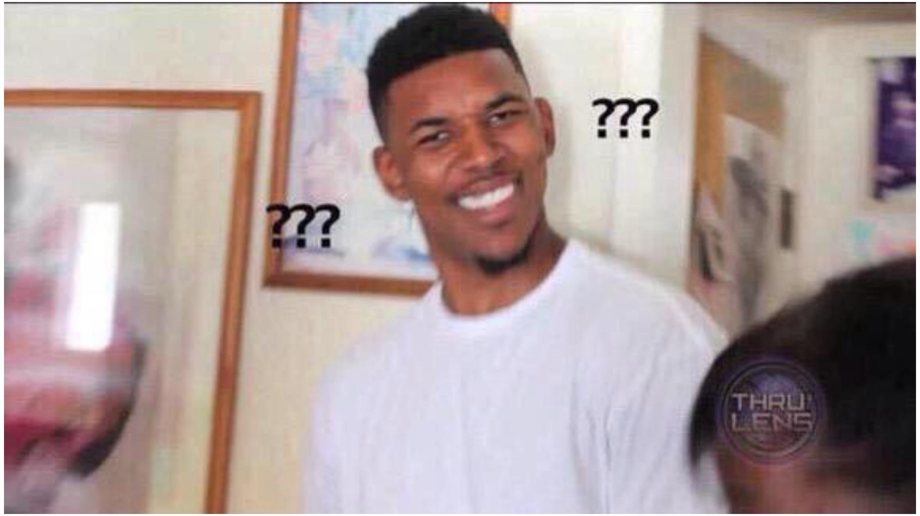
	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
726	727	2012-12-27	1	1	12	0	4	1	2	0.254167	0.226642	0.652917	0.350133	247	1867	2114
727	728	2012-12-28	1	1	12	0	5	1	2	0.253333	0.255046	0.590000	0.155471	644	2451	3095
728	729	2012-12-29	1	1	12	0	6	0	2	0.253333	0.242400	0.752917	0.124383	159	1182	1341
729	730	2012-12-30	1	1	12	0	0	0	1	0.255833	0.231700	0.483333	0.350754	364	1432	1796
730	731	2012-12-31	1	1	12	0	1	1	2	0.215833	0.223487	0.577500	0.154846	439	2290	2729

Model Evaluation

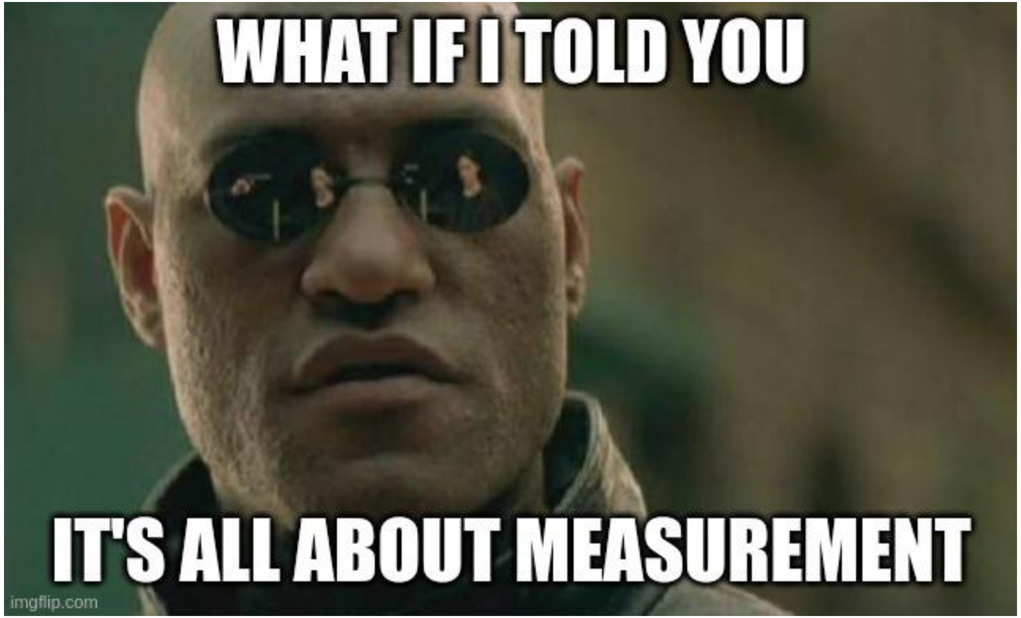
1. **Model Evaluation**
    a. Metrics and performances
    b. Generalization
2. Machine Learning Models
    a. Theory
    b. Linear Regression
    c. Logistic Regression
    d. Decision Tree

Metrics and Performances

Why starting the discussion from the end…?



Metrics and Performances

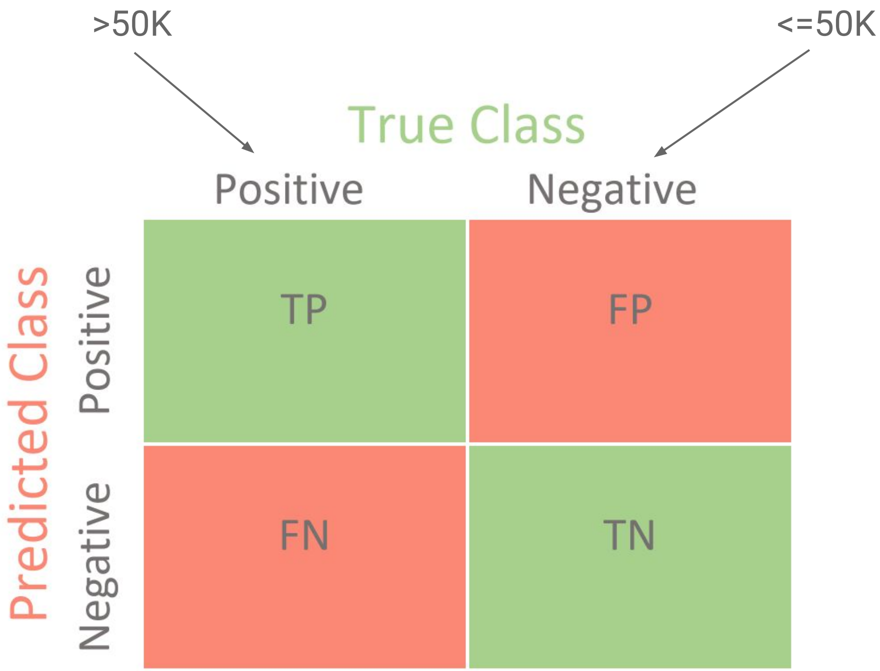


Metrics and Performances

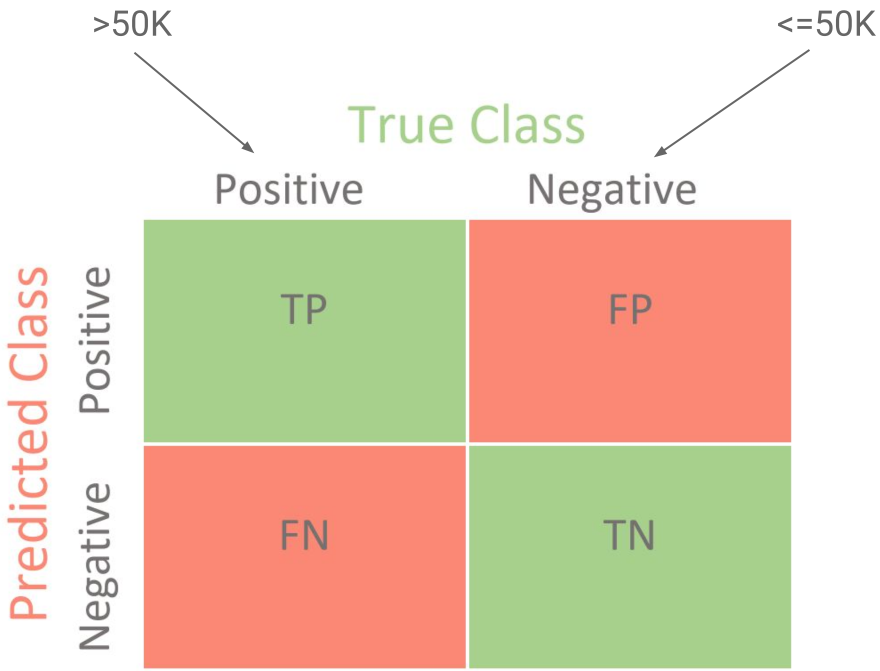
	age	workclass	fnlwgt	education	education-num	marital-status	occupation	relationship	race	sex	capital-gain	capital-loss	hours-per-week	native-country	class
0	25.0	Private	226802.0	11th	7.0	Never-married	Machine-op-inspct	Own-child	Black	Male	0.0	0.0	40.0	United-States	<=50K
1	38.0	Private	89814.0	HS-grad	9.0	Married-civ-spouse	Farming-fishing	Husband	White	Male	0.0	0.0	50.0	United-States	<=50K
2	28.0	Local-gov	336951.0	Assoc-acdm	12.0	Married-civ-spouse	Protective-serv	Husband	White	Male	0.0	0.0	40.0	United-States	>50K
3	44.0	Private	160323.0	Some-college	10.0	Married-civ-spouse	Machine-op-inspct	Husband	Black	Male	7688.0	0.0	40.0	United-States	>50K
4	18.0	NaN	103497.0	Some-college	10.0	Never-married	NaN	Own-child	White	Female	0.0	0.0	30.0	United-States	<=50K

What could be good measures of success for a (binary) classification problem?

Metrics and Performances

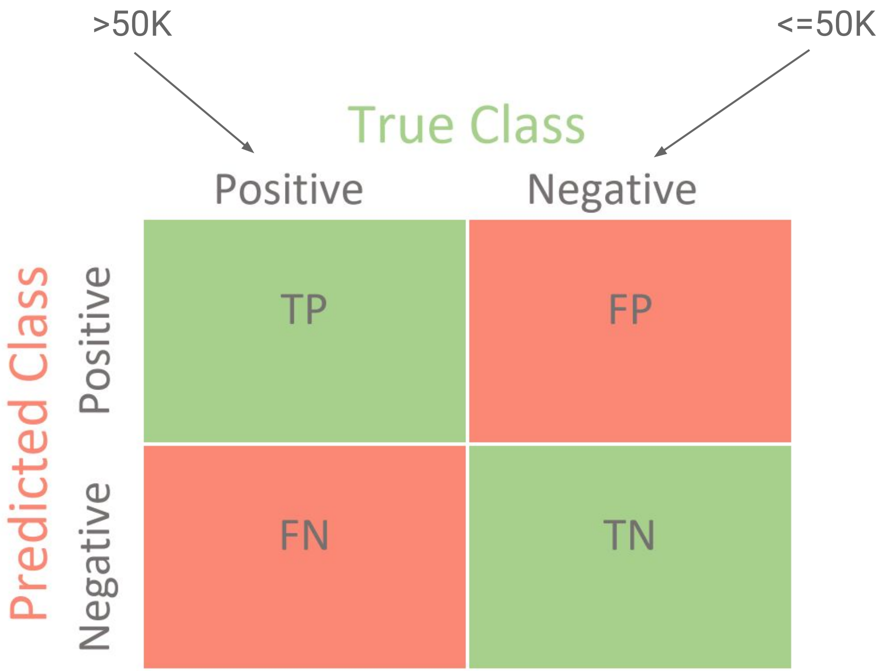


Metrics and Performances



Accuracy = (TP + TN) / all

Metrics and Performances



Accuracy = (TP + TN) / all

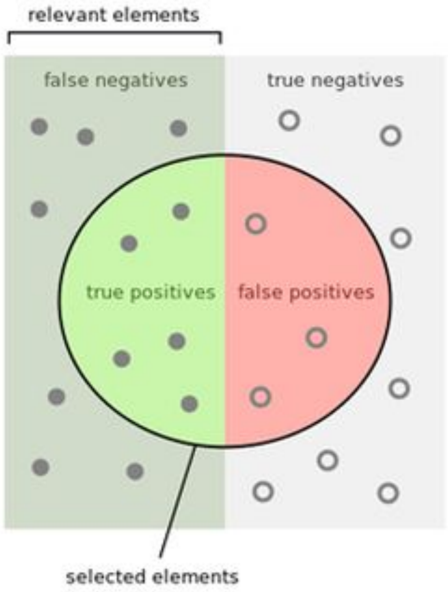
Precision = TP / (TP + FP)

Recall = TP / (TP + FN)

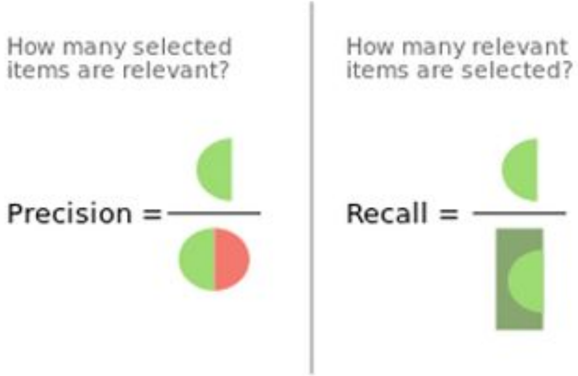
Metrics and Performances

>50K

<=50K



predicted >50K



Metrics and Performances

From Wikipedia, the free encyclopedia

Sources: [1][2][3][4][5][6][7][8] view · talk · edit

		Predicted condition			
	Total population $= P + N$	**Positive (PP)**	**Negative (PN)**	Informedness, bookmaker informedness (BM) $= TPR + TNR - 1$	Prevalence threshold (PT) $= \frac{\sqrt{TPR \times FPR} - FPR}{TPR - FPR}$
Actual condition	**Positive (P)**	**True positive (TP),** **hit**	**False negative (FN),** **type II error, miss,** **underestimation**	True positive rate (TPR), recall, sensitivity (SEN), probability of detection, hit rate, power $= \frac{TP}{P} = 1 - FNR$	False negative rate (FNR), miss rate $= \frac{FN}{P} = 1 - TPR$
	Negative (N)	**False positive (FP),** **type I error, false alarm,** **overestimation**	**True negative (TN),** **correct rejection**	False positive rate (FPR), probability of false alarm, fall-out $= \frac{FP}{N} = 1 - TNR$	True negative rate (TNR), specificity (SPC), selectivity $= \frac{TN}{N} = 1 - FPR$
	Prevalence $= \frac{P}{P + N}$	Positive predictive value (PPV), precision $= \frac{TP}{PP} = 1 - FDR$	False omission rate (FOR) $= \frac{FN}{PN} = 1 - NPV$	Positive likelihood ratio (LR+) $= \frac{TPR}{FPR}$	Negative likelihood ratio (LR−) $= \frac{FNR}{TNR}$
	Accuracy (ACC) $= \frac{TP + TN}{P + N}$	False discovery rate (FDR) $= \frac{FP}{PP} = 1 - PPV$	Negative predictive value (NPV) $= \frac{TN}{PN} = 1 - FOR$	Markedness (MK), deltaP (Δp) $= PPV + NPV - 1$	Diagnostic odds ratio (DOR) $= \frac{LR+}{LR-}$
	Balanced accuracy (BA) $= \frac{TPR + TNR}{2}$	F_1 score $= \frac{2 PPV \times TPR}{PPV + TPR} = \frac{2TP}{2TP + FP + FN}$	Fowlkes–Mallows index (FM) $= \sqrt{PPV \times TPR}$	Matthews correlation coefficient (MCC) $= \sqrt{TPR \times TNR \times PPV \times NPV} - \sqrt{FNR \times FPR \times FOR \times FDR}$	Threat score (TS), critical success index (CSI), Jaccard index $= \frac{TP}{TP + FN + FP}$

Metrics and Performances

Different use case may prioritize different metrics:

1. Predict if a patient is at risk of developing a serious medical problem.
2. Predict a good day based on weather conditions to launch a satellite.
3. Predict if a financial transaction is fraudulent.
4. Classify email as spam.

Recall

True positive rate (TPR), recall, sensitivity (SEN),
probability of detection, hit rate, power
$= \frac{TP}{P} = 1 - FNR$

Precision

Positive predictive value (PPV),
precision
$= \frac{TP}{PP} = 1 - FDR$

Metrics and Performances

Number of rented bikes per day

	instant	dteday	season	yr	mnth	holiday	weekday	workingday	weathersit	temp	atemp	hum	windspeed	casual	registered	cnt
0	1	2011-01-01	1	0	1	0	6	0	2	0.344167	0.363625	0.805833	0.160446	331	654	985
1	2	2011-01-02	1	0	1	0	0	0	2	0.363478	0.353739	0.696087	0.248539	131	670	801
2	3	2011-01-03	1	0	1	0	1	1	1	0.196364	0.189405	0.437273	0.248309	120	1229	1349
3	4	2011-01-04	1	0	1	0	2	1	1	0.200000	0.212122	0.590435	0.160296	108	1454	1562
4	5	2011-01-05	1	0	1	0	3	1	1	0.226957	0.229270	0.436957	0.186900	82	1518	1600
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
726	727	2012-12-27	1	1	12	0	4	1	2	0.254167	0.226642	0.652917	0.350133	247	1867	2114
727	728	2012-12-28	1	1	12	0	5	1	2	0.253333	0.255046	0.590000	0.155471	644	2451	3095
728	729	2012-12-29	1	1	12	0	6	0	2	0.253333	0.242400	0.752917	0.124383	159	1182	1341
729	730	2012-12-30	1	1	12	0	0	0	1	0.255833	0.231700	0.483333	0.350754	364	1432	1796
730	731	2012-12-31	1	1	12	0	1	1	2	0.215833	0.223487	0.577500	0.154846	439	2290	2729

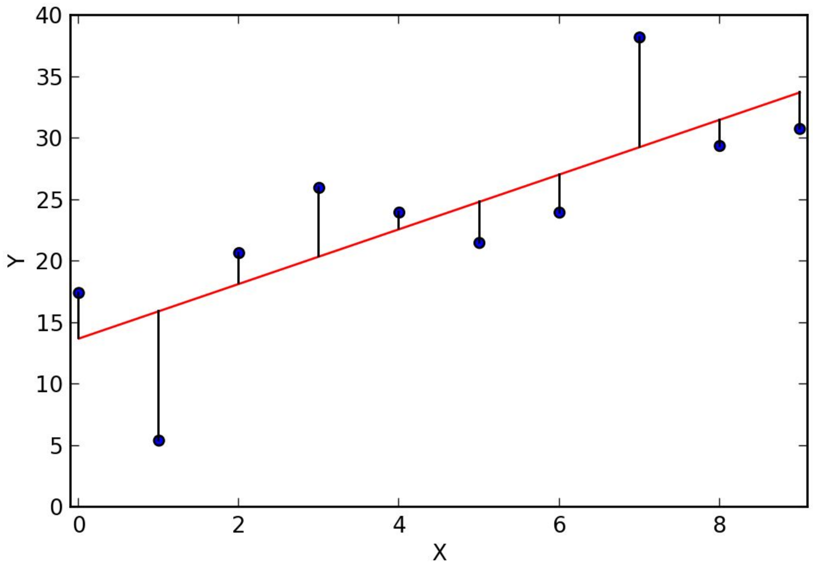
What could be good measures of success for a regression problem?

Metrics and Performances

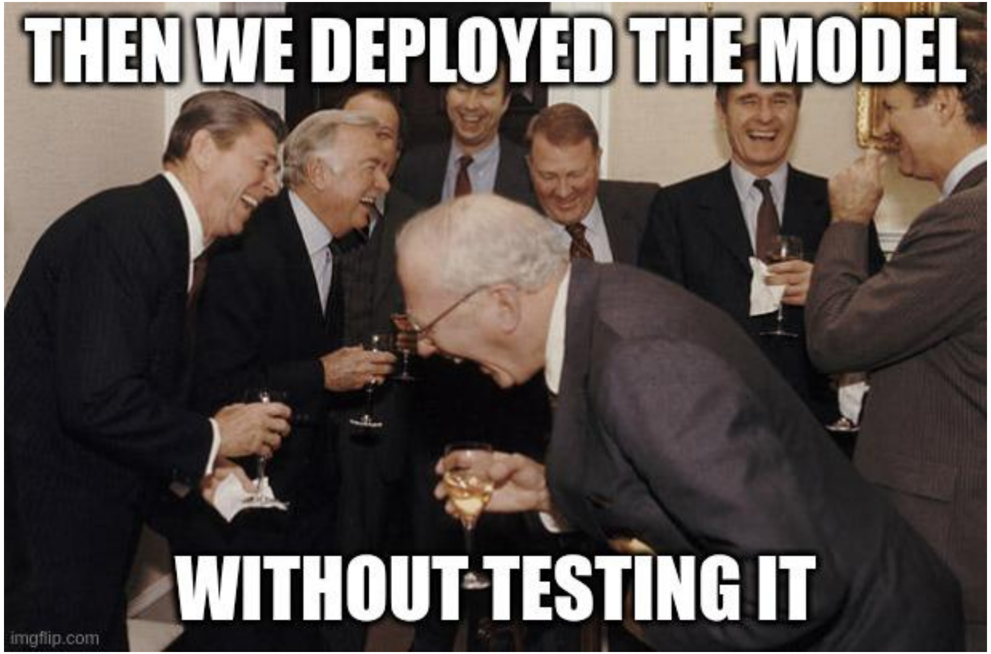
Measure of distance between predictions and actual values.

- Mean Squared Error
- Mean Absolute Error

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



Generalization



Generalization

Generalization Error

How good is the algorithm on previously unseen data?

Generalization

Generalization Error

How good is the algorithm on previously unseen data?

But how to estimate errors on “future” data?

Trick: split the data in two parts.

- Train the model on the first one,
- Test the model on the second one.

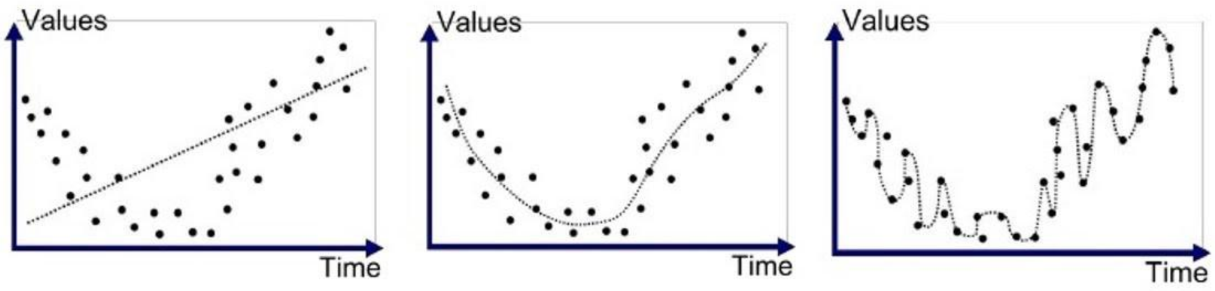
Generalization

Underfit

The model performs poorly on the training data (and on the test data).

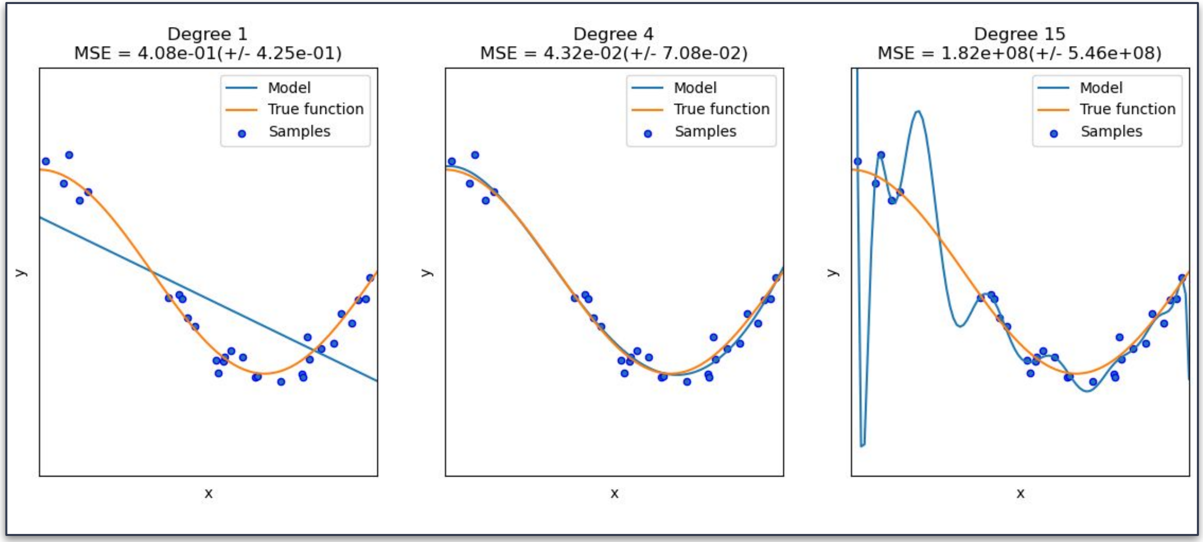
Overfit

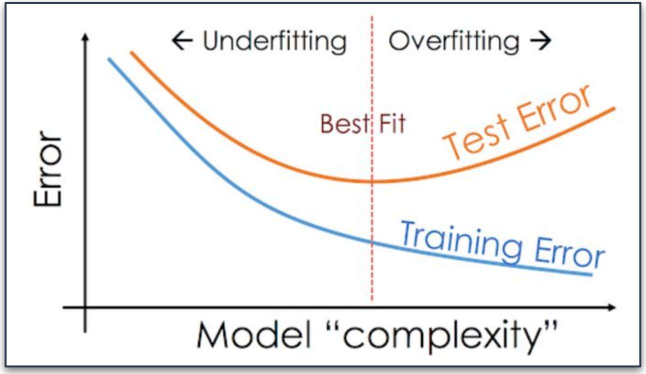
The model performs well on the training data but poorly on the test data.



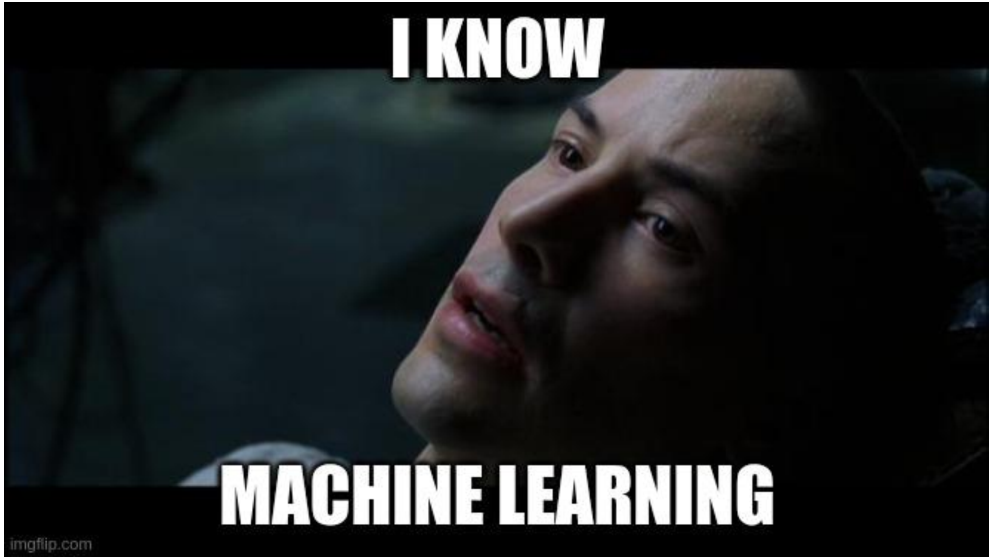
Underfitted Good Fit/Robust Overfitted

Generalization





Generalization



Not yet… but we are getting closer!

Machine Learning Models

Machine Learning Models

Census data

Income class

Properties

EPC Rating

$$Y = f(X)$$

Output = f(inputs)

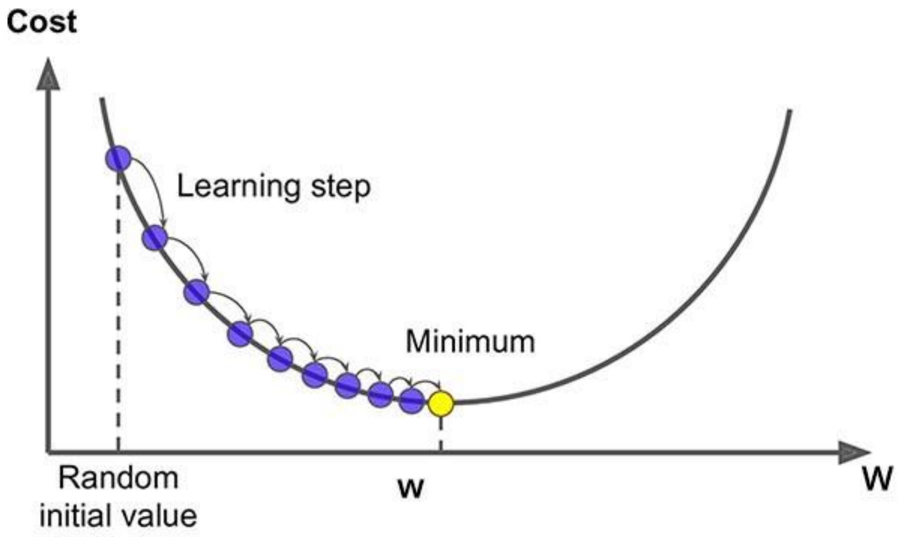
Target = f(features)

Machine Learning Models

Machine learning algorithms are techniques for estimating the target function (f) to predict the output variable (Y) given input variables (X).

But how...?

Usually trying to **minimize a loss function**.



Machine Learning Models

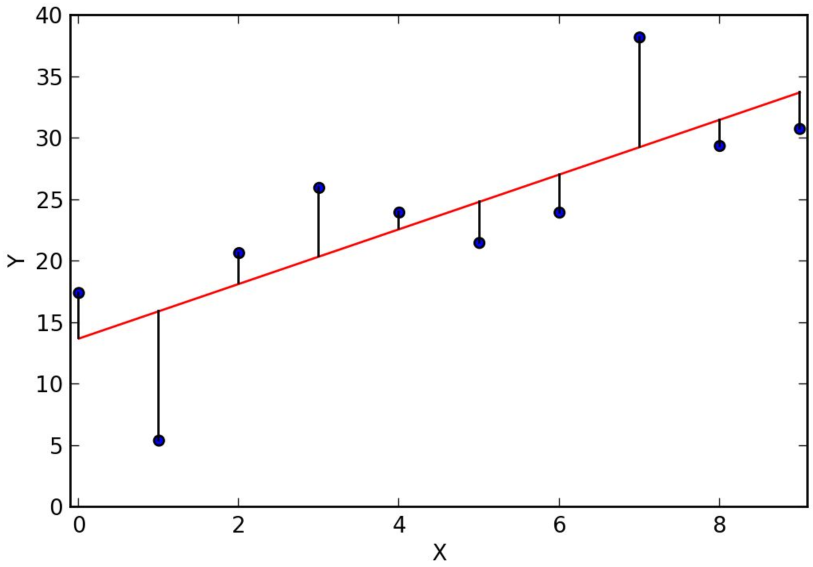
Linear Regression

$$f(x) = \beta_0 + \beta_1 x$$

The algorithm assigns to (β_0, β_1) the values that minimize the **loss function**:

$$\mathcal{L} = \sum_{i=1}^{N} (y_i - f(x_i))^2$$

Does it reminds you about MSE?



Machine Learning Models

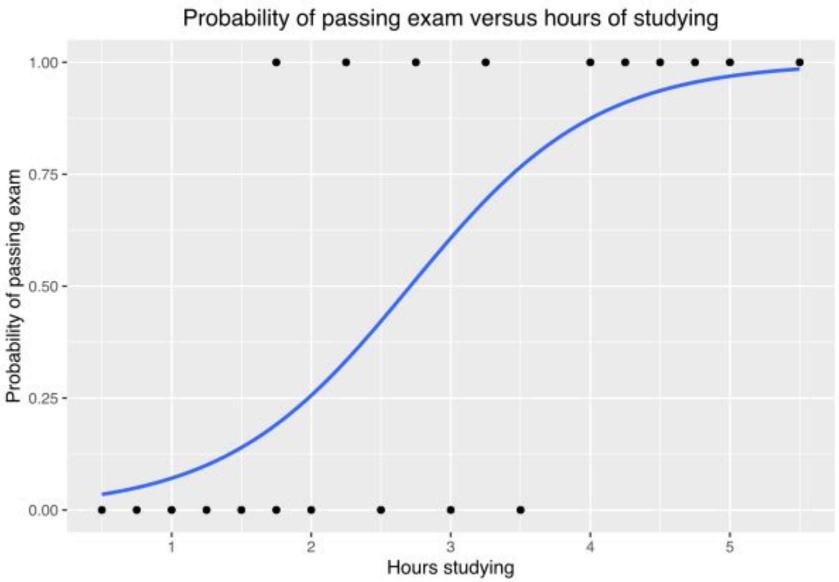
Logistic Regression

It is a binary classification model, where f has the form

$$z = \beta_0 + \beta_1 x$$
$$f(z) = \frac{1}{1+e^{-z}}$$

Again, the parameters are the ones that minimize a loss function:

$$\mathcal{L} = \sum_{n=1}^{N} \Big[y_n \ln(f(x_n)) + (1-y_n) \ln(1-f(x_n)) \Big]$$

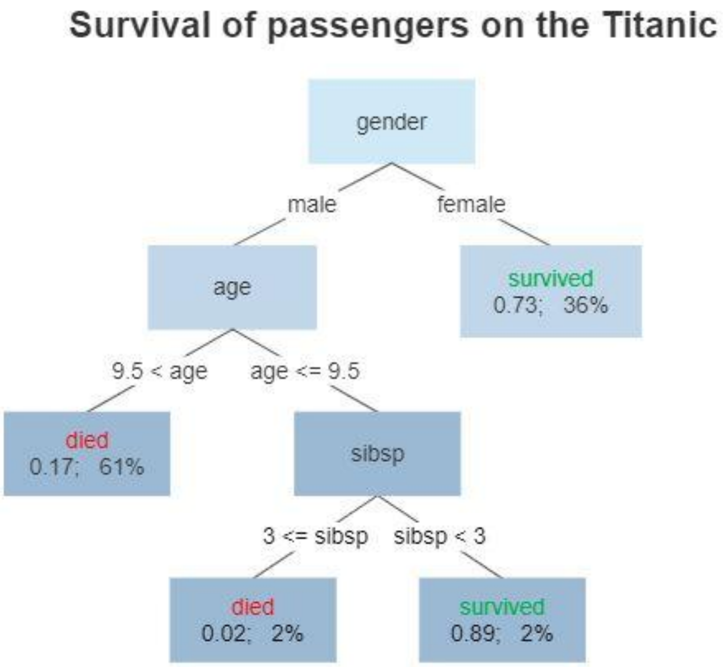


Machine Learning Models

Decision Tree

It is a **non-parametric** model for regression or classification.

Trees are built recursively splitting the feature space. It is a greedy algorithm: different variables and split points are tested, then at each step the chosen one is the one that minimizes a **loss function**.



Machine Learning Models

Summarizing, machine learning models:

- are algorithms that try to estimate a function that maps input variables to output variables.
- They usually do that with numerical optimization of an loss function.