

Supervised Learning

Introduction

Diego Campos Sobrino



UNIVERSIDAD
POLITÉCNICA
DE YUCATÁN



1 Basic concepts

2 Data

3 Model

Artificial Intelligence

"The designing and building of intelligent agents that receive percepts from the environment and take actions that affect that environment"

— Peter Norvig & Stuart Russell

"AI is the new electricity"

— Andrew Ng

- Machine Learning grew out of work in AI.
- AI have been fueled greatly by Machine Learning.

Added capabilities for computers.

Database mining

- Large datasets from growth of automation/web, e.g., web click data, medical records, biology, engineering, etc.

Applications can't be programmed by hand

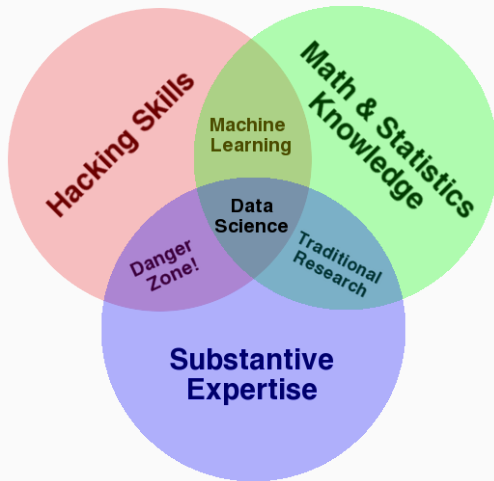
- Autonomous vehicles, handwriting recognition, natural language processing, computer vision.

Self-customizing programs

- Amazon and Netflix product recommendations.

The era of big data calls for automated methods of data analysis, which ML provides.

Conway Venn Diagram



Machine Learning

"Field of study that gives computers the ability to learn without being explicitly programmed."

— Arthur Samuel (1959)

Machine Learning

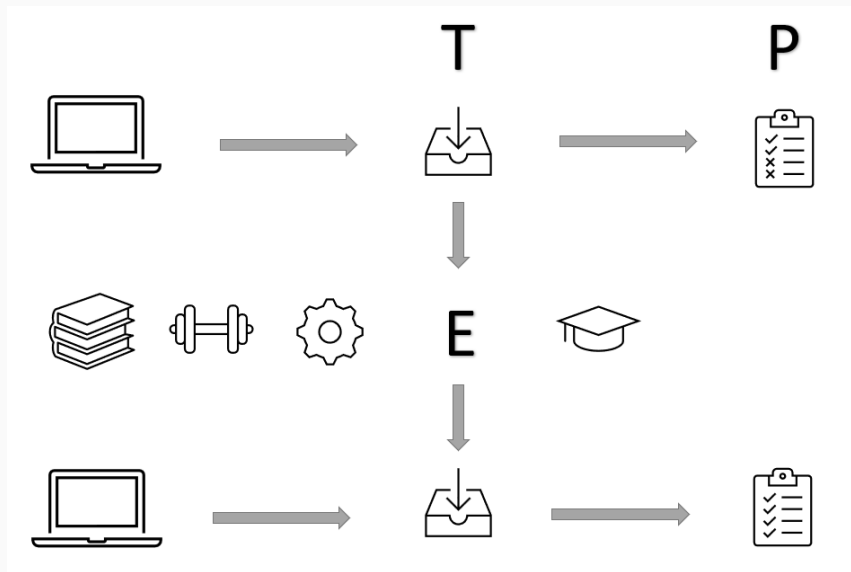
"Field of study that gives computers the ability to learn without being explicitly programmed."

— Arthur Samuel (1959)

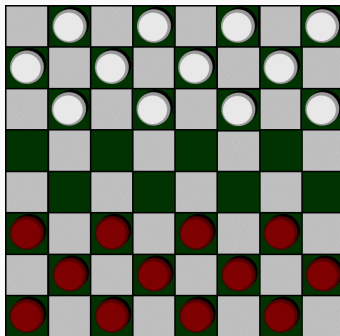
Machine Learning problem

"A computer is said to learn from experience E with respect to some task T and some performance measure P , if its performance on T , as measured by P , improves with experience E ."

— Tom Mitchell (1997)



- T = the task of playing checkers.
- E = the experience of playing many games of checkers.
- P = the probability that the program will win the next game.



Suppose your email program records which emails you do mark as spam, and based on that learns how to better filter spam.



What are T, E, and P?

- Keep record of the messages marked as spam
- Fraction of emails correctly classified as spam / not spam
- Classify emails as spam or not spam

T

- What the computer tries to learn.
- The process of learning is not the task.
- ML tasks are usually described in terms of how the ML system process an example.
- An example is a collection of features represented as a vector $\mathbf{x} \in \mathbb{R}^n$.

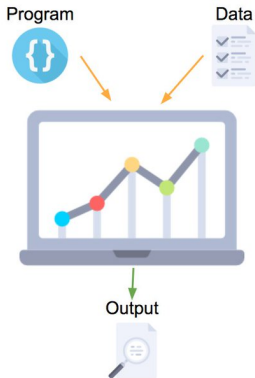
P

- Quantitative measure specific to the task *T*.
- Measures how good the system is at performing the task (specially for unseen data).
- Sometimes is not easy to choose or even to obtain the appropriate *P* measure.

E

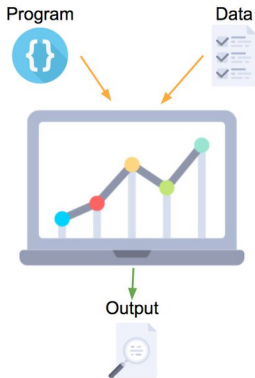
- Algorithm for learning the task.
- The algorithm experience the known data to learn from it.

Traditional Programming



@matthewfitch23

Traditional Programming

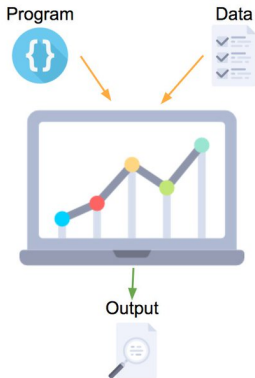


Machine Learning

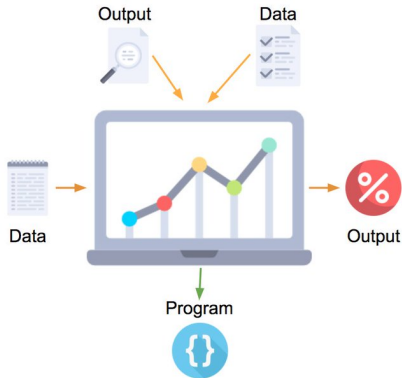


@matthewfitch23

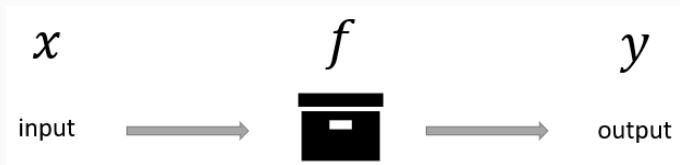
Traditional Programming

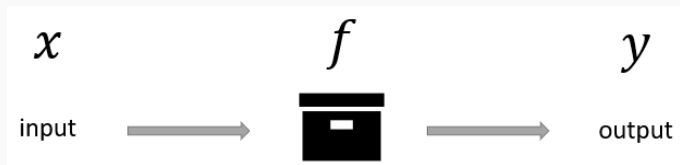


Machine Learning

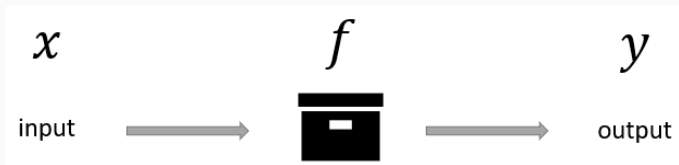


@matthewfitch23



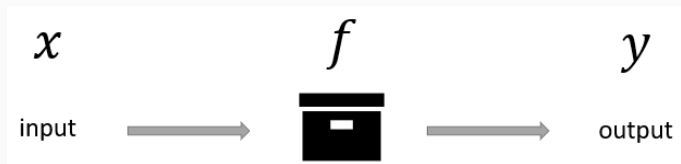


x	f	y
2	?	6
4		12
-5		-15
8		24
3		9
-1		-3

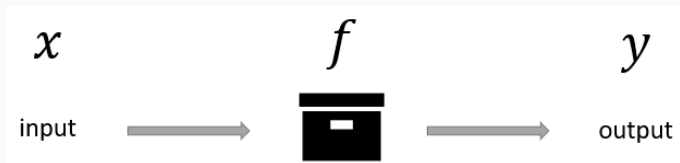


x	f	y
2	?	6
4		12
-5		-15
8		24
3		9
-1		-3

Once you identify f , what is the value of y if $x = 6$?

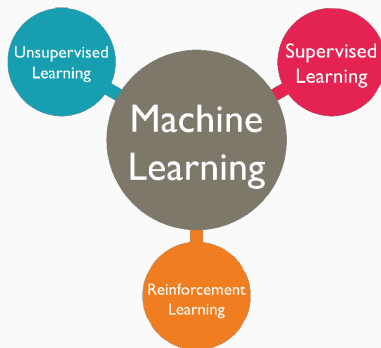


x	f	y
2	?	5.8
4		12.1
-5		-15.4
8		25
3		9
-1		-2.9

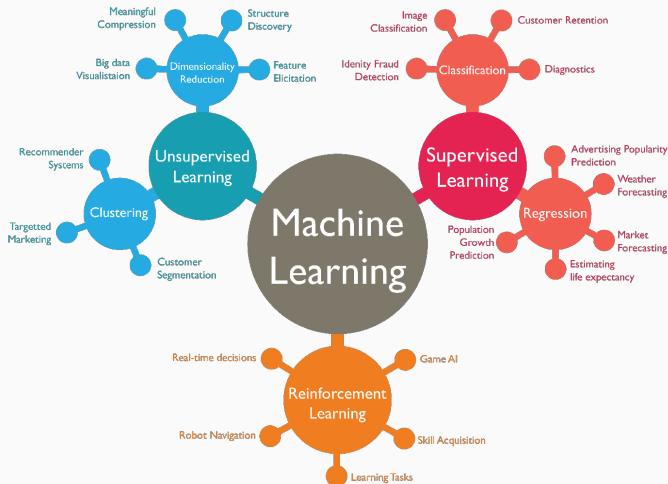


x	f	y
2	?	5.8
4		12.1
-5		-15.4
8		25
3		9
-1		-2.9

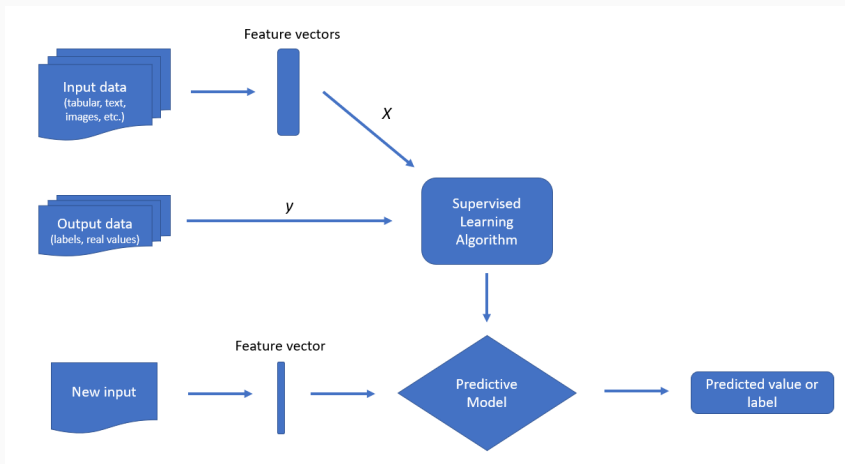
- Most of the time f is not obvious and can only be approximated.
- Input data X is usually multidimensional (not in this example).



<https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/>







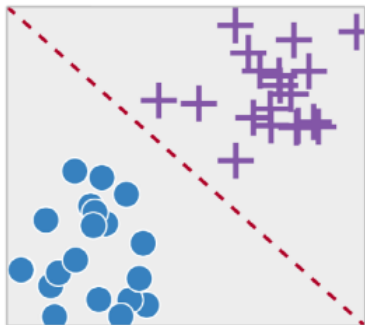
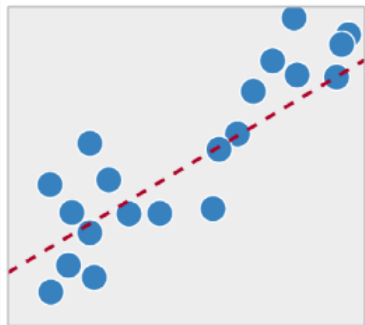
<https://www.techleer.com/articles/203-machine-learning-algorithm-backbone-of-emerging-technologies/>



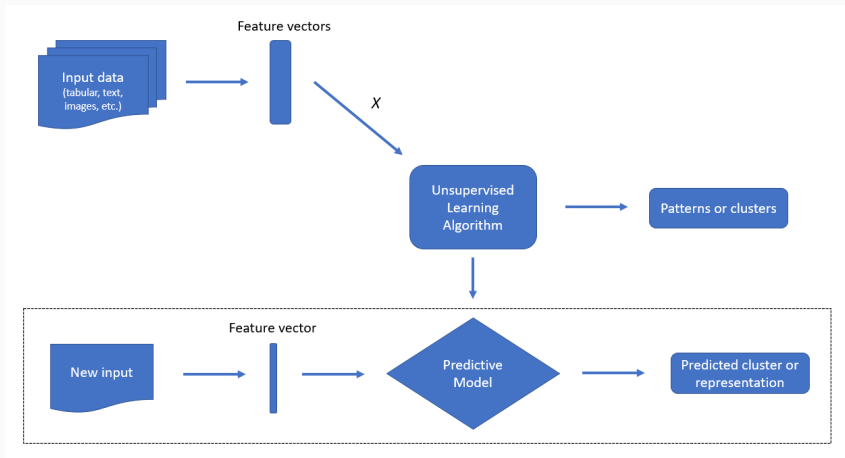
$$\mathbf{x} \rightarrow \mathbf{y}$$

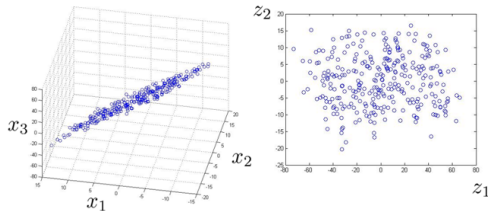
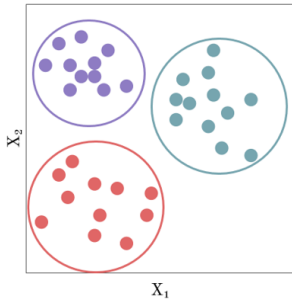
size of house (square feet)	# of bedrooms	price (1000\$)
523	1	115
645	1	150
708	2	210
1034	3	280
2290	4	355
2545	4	440

image	label
	cat
	not cat
	cat
	not cat



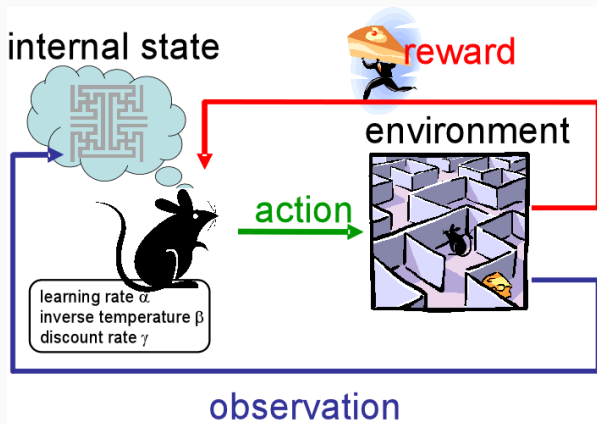
<https://scorecardstreet.wordpress.com/2015/12/09/is-machine-learning-the-new-epm-black/>





<https://laptrinhx.com/a-brief-introduction-to-unsupervised-learning-3814402973/>

A software agent performs actions inside an environment to maximize its long term reward.



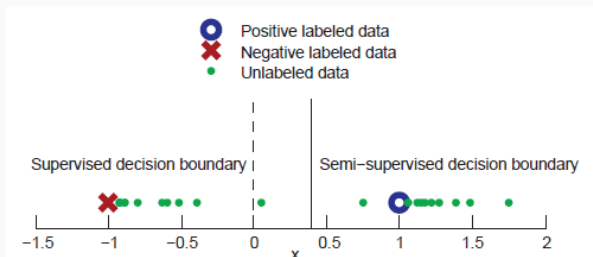
<https://becominghuman.ai/the-very-basics-of-reinforcement-learning-154f28a79071>

Machine learning approach that combines a small amount of labeled data with a large amount of unlabeled data during training.

- Acquisition of labeled data often is hard to obtain or expensive (i.e. requires a skilled human agent, repetition of physical experiments, etc.).
- Get unlabeled data often is relatively inexpensive.
- Can produce considerable improvement in performance without being too costly.
- Theoretically interesting as a model for human learning.

Machine learning approach that combines a small amount of labeled data with a large amount of unlabeled data during training.

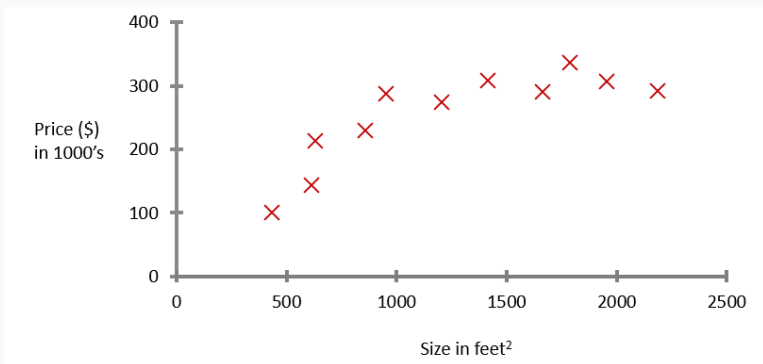
- Acquisition of labeled data often is hard to obtain or expensive (i.e. requires a skilled human agent, repetition of physical experiments, etc.).
- Get unlabeled data often is relatively inexpensive.
- Can produce considerable improvement in performance without being too costly.
- Theoretically interesting as a model for human learning.



- Use additional unlabeled data to better capture the shape of the data distribution and generalize better to new samples.

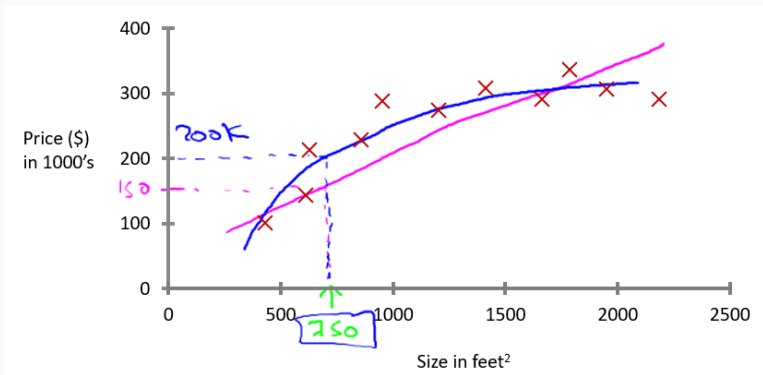
- Speech recognition: Accurate transcription by human expert can be extremely time consuming.
- Natural language parsing: Tree-banks are time consuming and require expertise of linguistics.
- Spam filtering: The bottleneck is an average user's patience to label a large amount of emails.
- Video surveillance: Manually labeling objects in a large number of video frames is tedious and time consuming.
- Protein 3D structure prediction: It can take months of expensive laboratory work by expert crystallographers to identify the 3D structure of a single protein.

Housing price prediction with one variable (size).



- Supervised Learning: "right" answers are given.
- Regression: predict continuous valued output.

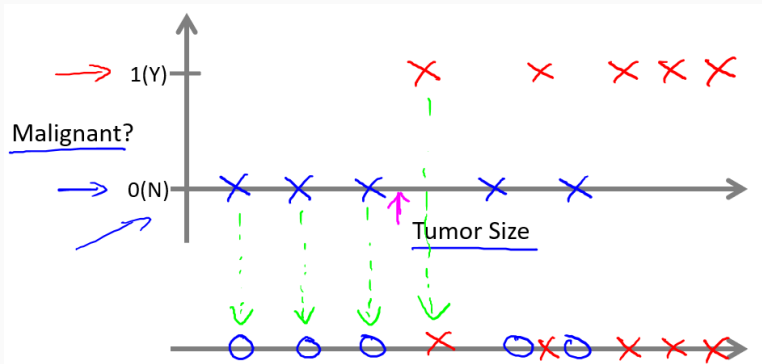
Housing price prediction with one variable (size).



- Supervised Learning: "right" answers are given.
- Regression: predict continuous valued output.

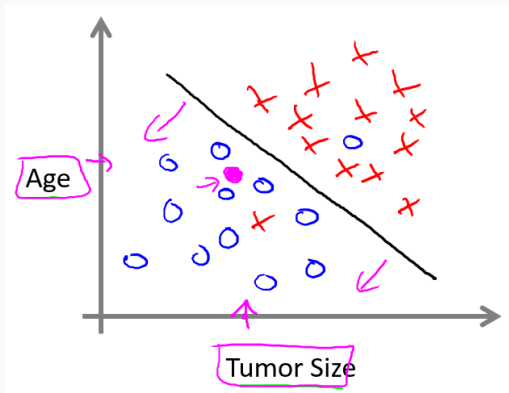
- Linear regression
- Ridge regression
- Lasso
- Elastic-Net
- Bayesian linear regression
- Decision Trees
- Support Vector Regression
- Neural Networks

Breast cancer (malignant, benign).



- Discrete value output (0 label for negative, 1 label for positive).

Breast cancer (malignant, benign).



- Discrete value output (0 label for negative, 1 label for positive).
- Include Age variable.

- K-Nearest Neighbors
- Logistic regression
- Naive Bayes
- Decision Trees
- Random Forests
- Support Vector Machines
- Neural Networks
- Linear Discriminant Analysis

Determine if it is a regression or classification problem:

1. You have a large inventory of identical items, and you want to predict how many of these items will sell over the next 3 months.
2. You'd like software to examine individual customer accounts, and for each account decide if it has been hacked.
3. Given a dataset of patients diagnosed as either having diabetes or not, learn to predict if a new patient have diabetes.
4. Predict the amount of rainfall over a region in mm for a particular day.
5. Given a database of customer data, automatically discover market segments and group customers into different market segments.
6. Predict whether stock price of a company will increase tomorrow.

- Explain the main differences between supervised and unsupervised learning.
- What is the key difference between regression and classification?
- What kind of problems are solved with Machine Learning?

1. Dataset

- Examples of $\mathbf{x} \rightarrow y$ mapping.
- Subdivided in *train*, *validation* or *development*, and *test*.

2. Model

- Primary function or procedure that accepts \mathbf{x} and returns \hat{y} .
- Algorithm learns model *parameters*.
- User defines model *hyperparameters*.

3. Cost function

- Measures how good the model is at predicting the reality (how close \hat{y} is from y).
- Depends on the task and data type.

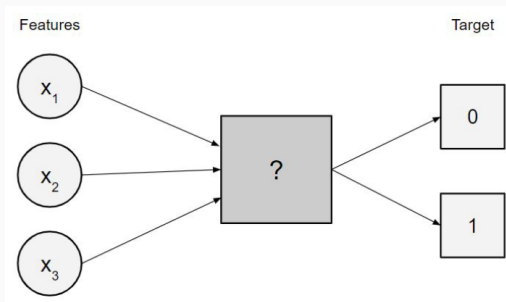
4. Optimization procedure

- Method used to minimize or maximize the cost function.
- Estimate the model parameters that make the model perform better.

Goodfellow, Bengio, & Courville

The data in Supervised Learning usually consists of observations on different aspects of objects:

- **Target:** the output variables / goal of prediction.
- **Features:** measurable properties that provide a description of the objects.



- We assume some kind of relationship between the features and the target, in a sense that the value of the target variable can be explained by a combination of the features.
- When we know the target y for every example x in the dataset D doesn't mean we learned f since we have not guarantee to know anything about f outside of D .
- If D is capable of telling something outside of D that was previously unknown then learning is happening.

- Features and target variables may be of different data types:
 - **Numerical:** values in \mathbb{R} .
 - **Integer:** values in \mathbb{Z} .
 - **Binary:** values in $\{0, 1\}$.
 - **Categorical:** values in $\{C_1, \dots, C_k\}$.
 - **Ordinal:** values in $\{C_1, \dots, C_k\}$ where classes have a natural order with unknown distances.
- Most learning algorithms can only deal with numerical features, but there are some exceptions (i.e. decision trees).

Input

- Numerical (Real, Ordinal)
- Categorical (Binary, Multi-class)
- Other (Text, Images, Video, etc.)

Depending on algorithm used sometimes we need to transform the input data (data preparation, pre-processing, feature engineering).

Algorithms by data shape

- Tabular (Classical ML, Artificial Neural Networks)
- Image (Convolutional NN)
- Text (Classical NLP, Recurrent NN)

Output

- Numerical (Regression)
- Categorical (Binary, Multi-class, Multi-label Classification)

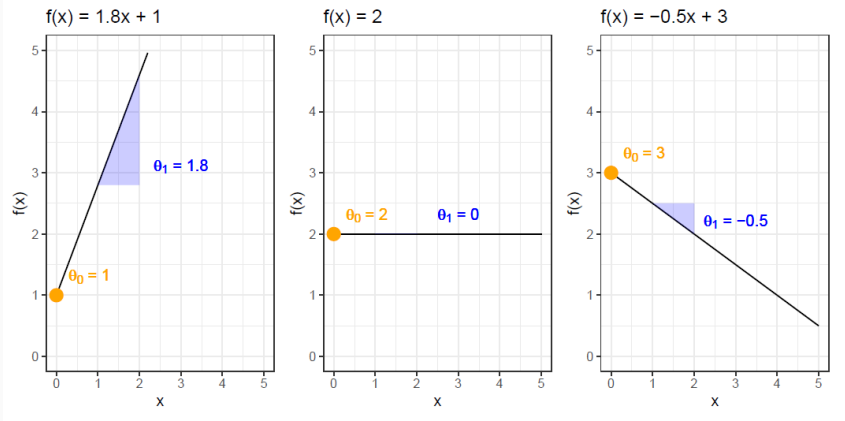
Another considerations

- A regression problem where input variables are ordered by time is called a time series forecasting problem.
- A classification algorithm may predict a continuous value, but in the form of a probability of a class label.
- A regression algorithm may predict a discrete value, but in the form of an integer quantity.
- A regression problem can be converted into a classification problem doing discretization (creating discrete ordered buckets).
- A classification problem can be converted into a regression problem transforming a label into a continuous range.

- A model (or hypothesis) is a function f that maps feature vectors \mathbf{x} to predicted target values y .
- f is meant to capture intrinsic patterns of the data with the assumption that these patterns hold true for all data drawn for the same probability distribution.
- Models can range from simple linear functions to very complex deep neural networks.
- Depending on the problem we can constrain f to be of a certain type of functions.
- The set of functions defining a specific model type is called a **hypothesis space** \mathcal{H} .
- All models within \mathcal{H} share a common structure that can be parameterized by a parameter vector θ .
- Finding the optimal model is equivalent to find the optimal set of parameter values, hence the utility of optimization methods in ML.

$$\mathcal{H} = \{f : f(\mathbf{x}) = \theta_0 + \theta_1 x_1\}$$

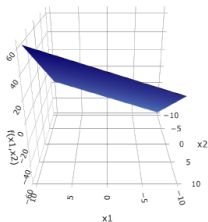
$$\theta \in \mathbb{R}^2$$



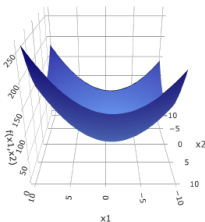
$$\mathcal{H} = \{f : f(\mathbf{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2\}$$

$$\theta \in \mathbb{R}^6$$

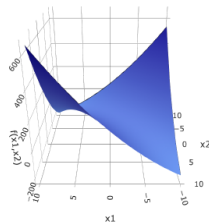
$$f(x) = 3 + 2x_1 + 4x_2$$



$$f(x) = 3 + 2x_1 + 4x_2 + 1x_1^2 + 1x_2^2$$



$$f(x) = 3 + 2x_1 + 4x_2 + 1x_1^2 + 1x_2^2 + 4x_1x_2$$



How many parameters have the following type of functions?
Can you write the corresponding hypothesis space?

1. 5-variable linear function passing through the origin.
2. 3-variable quadratic function.
3. Bivariate cubic function.

In supervised learning the focus is on probabilistic models of the form $p(y|\mathbf{x})$.

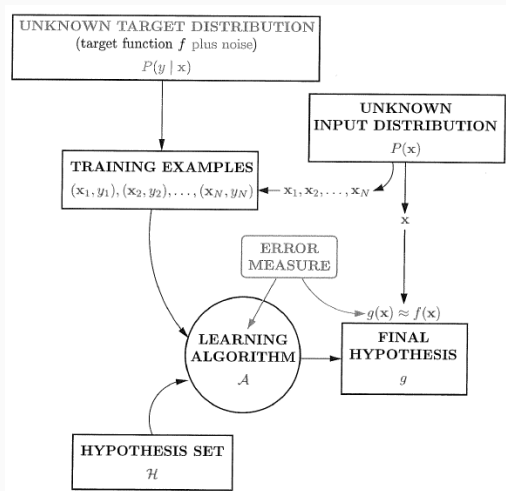
Parametric

- The model has a fixed number of parameters.
- Faster to use but with stronger assumptions about data distribution.
- Linear Regression, Logistic Regression, Naive Bayes, Neural Networks.

Non-parametric

- Model has no fixed number of parameters (can grow with the amount of data).
- More flexible but often intractable for large datasets.
- More risk of overfitting.
- K-Nearest Neighbors, Decision Trees.

- The algorithm for finding f is called a **learner** or **learning algorithm**.
- The learner also considers a vector λ of control settings called **hyperparameters**.
- The process goes like this:
 1. Learner has a defined model space \mathcal{H} .
 2. User passes data for training \mathcal{D}_{train} and control settings λ .
 3. Learner finds parameters so the model fits the data best.
 4. Optimal parameters $\hat{\theta}$ or function \hat{f} is returned for later usage with new data.



Abu-Mostafa. Learning from Data.

In Machine Learning we want to learn general patterns from the data, but how can be sure that we truly discovered a general pattern and not simply memorize the available data?

- Our goal is to discover patterns that capture regularities in the population from which our data was drawn.
- The problem is that we access only a finite sample of the data, so we run the risk that our discovered patterns turn out to not hold up when we collect more data.
- The phenomenon of fitting our available sample data more closely than we fit the real distribution of the data is called **overfitting**.
- The technique used to combat overfitting are called **regularization**.

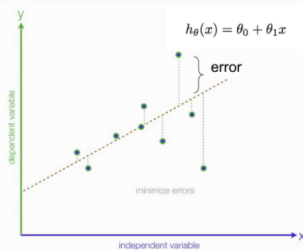
Model How can we know if our model is good?

Loss function

- Is a measurement of the difference between the actual and the predicted value of a single training example.

Cost function or Error function

- Helps us to understand the difference between the predicted values and the actual values as a whole.
- Usually is an average of the individual loss functions.
- Regression: Mean Error, Mean Absolute Error, Mean Squared Error.
- Classification: Binary Cross-entropy, Categorical Cross-entropy, Hinge Loss



$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

Hypothesis:

$$h_{\theta}(x) = \theta_0 + \theta_1 x$$

Parameters:

$$\theta_0, \theta_1$$

Cost Function:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^m (h_{\theta}(x^{(i)}) - y^{(i)})^2$$

Goal:

$$\underset{\theta_0, \theta_1}{\text{minimize}} J(\theta_0, \theta_1)$$

Training error

- Error of our model as calculated on the training set.

Generalization error

- Expectation of our model's error when applied to an infinite stream of additional data examples drawn from the same distribution as the training dataset.
- We can never know exactly this error exactly, so it is estimated by applying the model to an independent test set.

- Problem:
Classify the outcomes of coin tosses (0: heads, 1: tails).
- Training data:
{0, 1, 1, 1, 0, 1}.
- Model:
Predict always the majority class of our training data.
- Error function:
$$ME = \frac{1}{n} \sum (\hat{y} - y)$$
- What is the training error?
- Assuming the coin is fair what is the generalization error?
- What happens if we increase our training set?

- When we have simple models and abundant data, we expect the generalization error to resemble the training error.
- With more complex models and fewer examples, we expect the training error to go down but the gap with the generalization error to grow.
- Factors that influence model's ability to generalize:
 1. Number of tunable parameters (when large model are more susceptible to overfitting).
 2. Values taken by the parameters (when weights take a wider range of values models are more susceptible to overfitting).
 3. Number of training examples (with less data is easier to overfit, with more data is harder).
- The ability to fit a wide variety of functions is called model's **capacity**.
 - Low capacity may struggle to fit the data (underfitting).
 - High capacity can memorize properties of the training set (overfitting).

Signal

- The true underlying pattern that we wish the model to learn from the data.

Noise

- Refers to irrelevant information or randomness in a dataset.

"Learning is forgetting the details as much as it is remembering the important parts. Computers are the ultimate idiot savants: they can remember everything with no trouble at all, but that's not what we want them to do."

— Pedro Domingos

Two sources of error in a ML model.

Bias

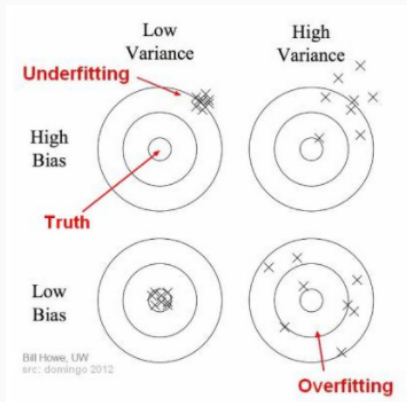
- Difference between your model's expected predictions and the true values.
- High bias cause an algorithm to miss relevant relations between features and target (underfitting).

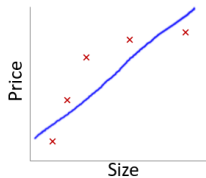
Variance

- Sensitivity to fluctuations in the sample data.
- High variance result from an algorithm modeling the random noise in the training data (overfitting).

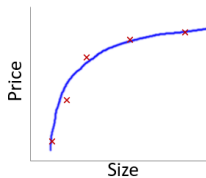
Tension between a model complex enough to capture the patterns in the data, but not so complex that we start to fit the noise.

- Too simple (high bias)
- Too complex (high variance)
- Increasing bias decreases variance
- Increasing variance decreases bias

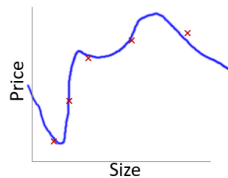




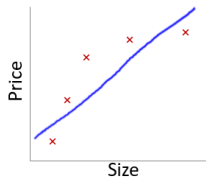
$$\theta_0 + \theta_1 x$$



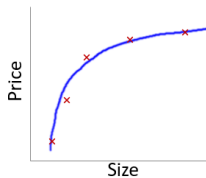
$$\theta_0 + \theta_1 x + \theta_2 x^2$$



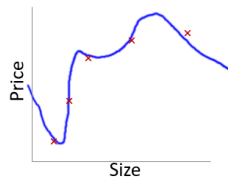
$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$



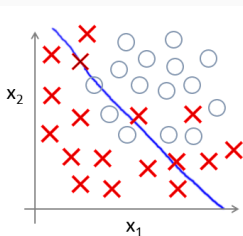
$$\theta_0 + \theta_1 x$$



$$\theta_0 + \theta_1 x + \theta_2 x^2$$

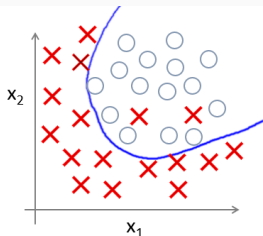


$$\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \theta_4 x^4$$

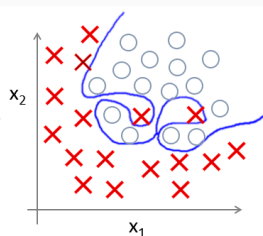


$$h_{\theta}(x) = g(\theta_0 + \theta_1 x_1 + \theta_2 x_2)$$

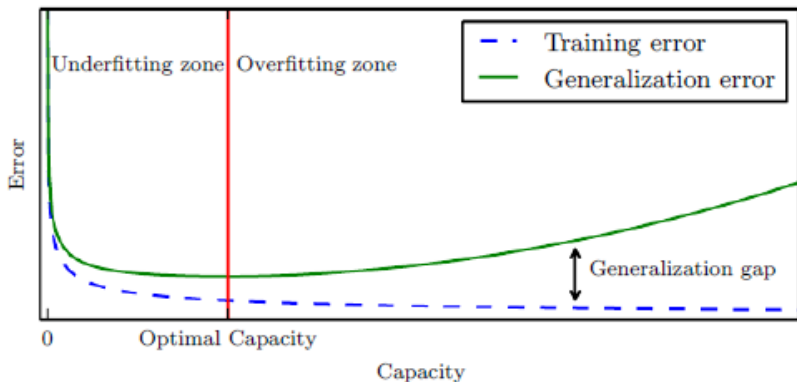
(g = sigmoid function)



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_1^2 + \theta_4 x_2^2 + \theta_5 x_1 x_2)$$



$$g(\theta_0 + \theta_1 x_1 + \theta_2 x_1^2 + \theta_3 x_1^2 x_2 + \theta_4 x_1^2 x_2^2 + \theta_5 x_1^2 x_2^3 + \theta_6 x_1^3 x_2 + \dots)$$



1. Reduce number of features.

- Manually select which features to keep.
- Model selection algorithm (later).

2. Regularization.

- Keep all the features, but reduce magnitude/values of parameters θ .
- Works well when we have a lot of features, each of them contributes a bit to predicting y .

Before an exam, a professor may hand out some practice problems with solutions (training set). The problems on the exam are the real test for your learning (test set).

Training set

- Data used for the model to learn.

Development or Validation set

- Data used to adjust hyperparameters of the model.

Test set

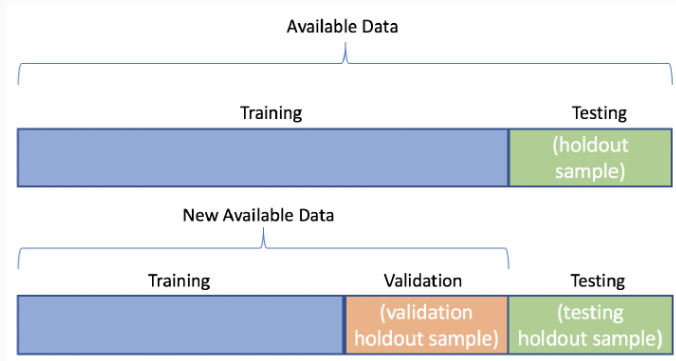
- To estimate the performance of the model with unseen data.

Traditionally:

- 70% train - 30% test
- 60% train - 20% val/dev - 20% test

With big data:

- 98% train - 1% val/dev - 1% test

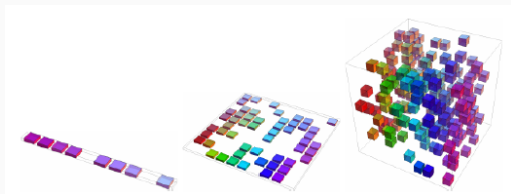


<https://datascience.stackexchange.com/questions/61467/clarification-on-train-test-and-val-and-how-to-use->

- Choose val-dev set size big enough to detect differences between models and algorithms.
- Choose test set size to obtain sufficient confidence in the performance of the predictive system.
- Choose val-dev and test sets to reflect data expected in the future in which is important to obtain good results.



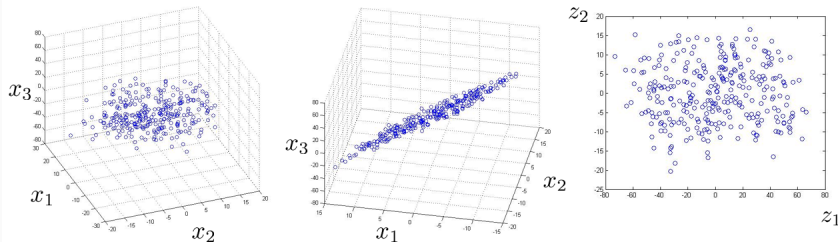
- Phenomenon that occurred when a ML problem become exceedingly difficult if the number of dimensions in the data is high.
- The number of possible distinct configurations of a set of variables increases exponentially as the number of variables increases.



- For d dimensions and v values to be distinguished along each axis we need $O(v^d)$ regions and examples.
- It is increasingly difficult to have examples for all regions, thus generalization suffers.

Transformation of data from high-dimensional space into a low-dimensional space that retains some meaningful properties of the original data.

- Feature selection: Try to find a subset of the input variables.
- Feature extraction: Transform the data from high to low dimensions.



Polynomial curve fitting

- Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*. Chapter 5.
- Murphy, K. *Machine Learning. A Probabilistic Perspective*. Chapter 1.
- Zhu, X.; Goldberg, A.B. *Introduction to Semi-Supervised Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning #6.
- Introduction to Machine Learning.
<https://introduction-to-machine-learning.netlify.app/>
- Machine Learning.
<https://www.coursera.org/learn/machine-learning>
- Difference Between Classification and Regression in Machine Learning.
<https://machinelearningmastery.com/classification-versus-regression-in-machine-learning/>
- The Ingredients of Machine Learning Algorithms.
<https://towardsdatascience.com/the-ingredients-of-machine-learning-algorithms-4d1ca9f5ceec>