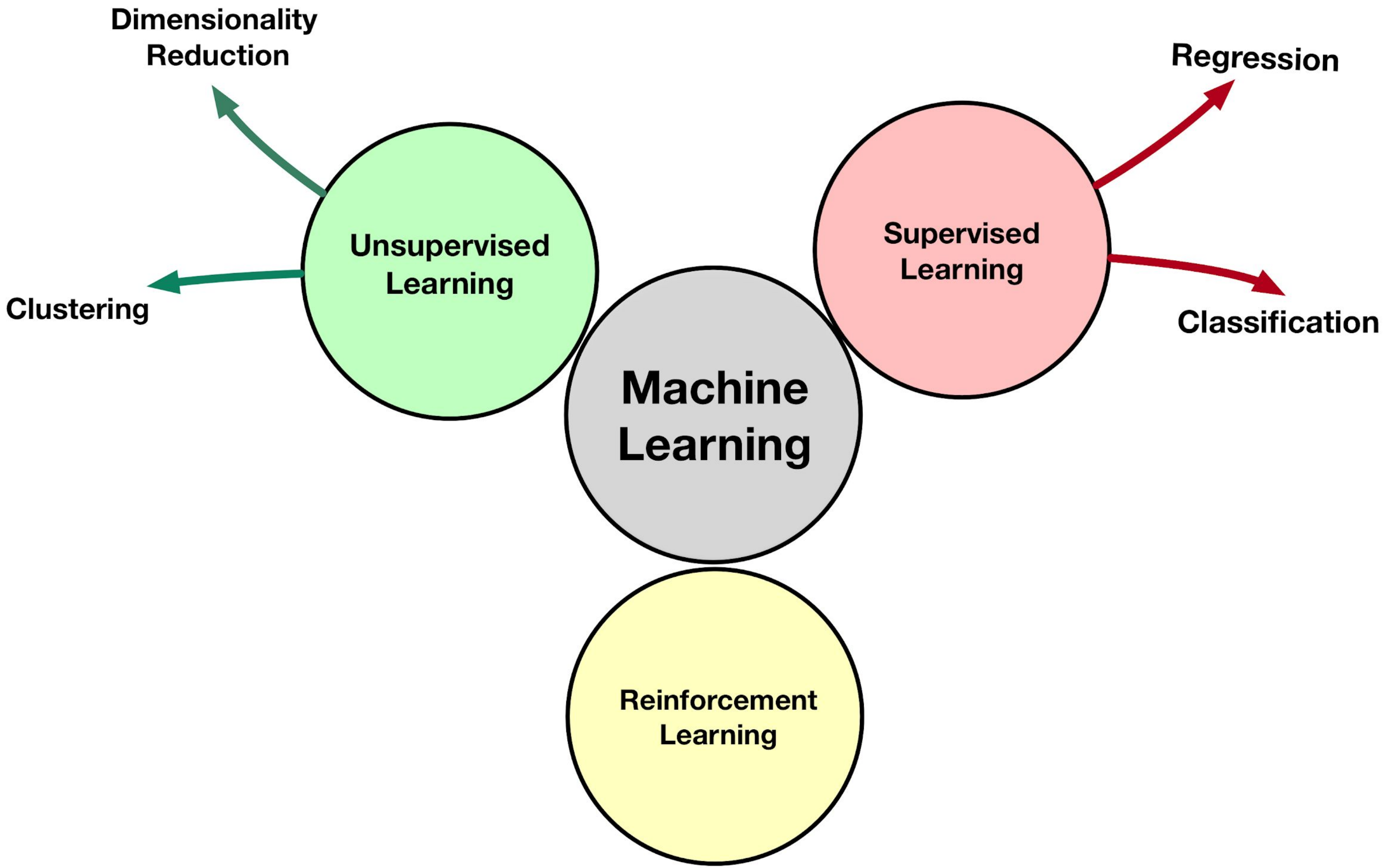
Supervised learning

Ali Madani
Farnoosh Khodakarami


Dimensionality
Reduction
Regression
Unsupervised
Learning
Supervised
Learning
Clustering
Classification
Machine
Learning
Reinforcement
Learning

Supervised vs Unsupervised Learning

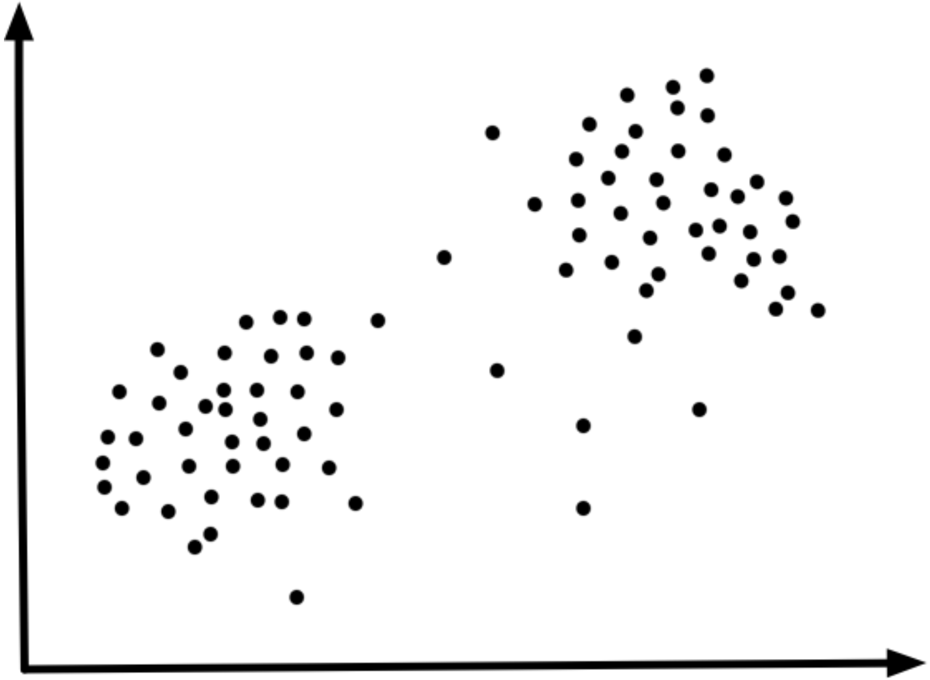
Unsupervised Learning

- **No Knowledge** of output
- data is **unlabeled**
- Self guided learning
- **Goal:** determine data patterns/grouping

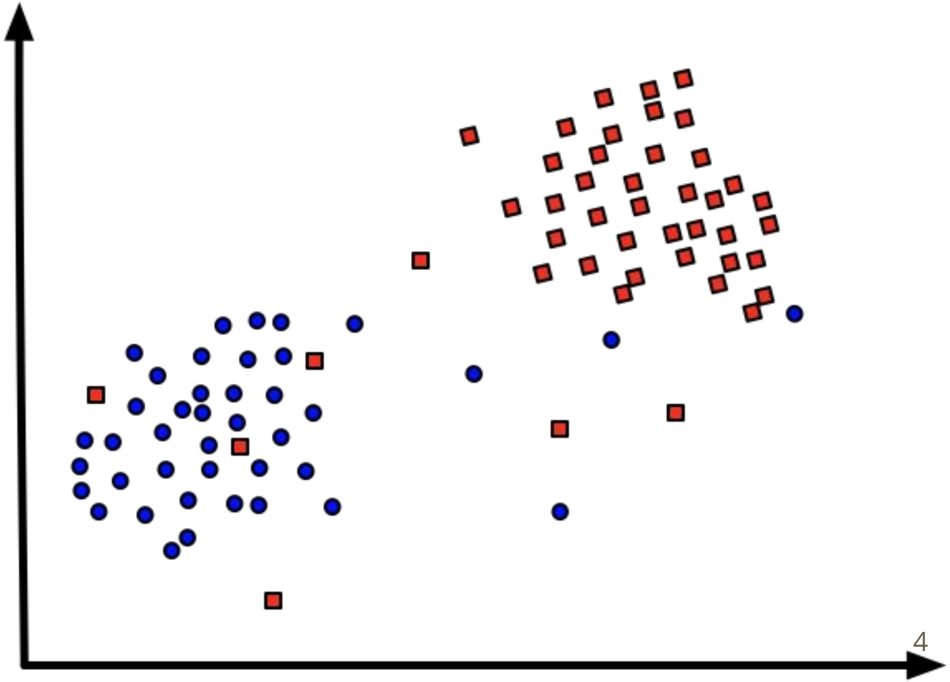
Supervised Learning

- **Knowledge** of output
- data is **labeled** with class or value
- **Goal:** predict value label or class label

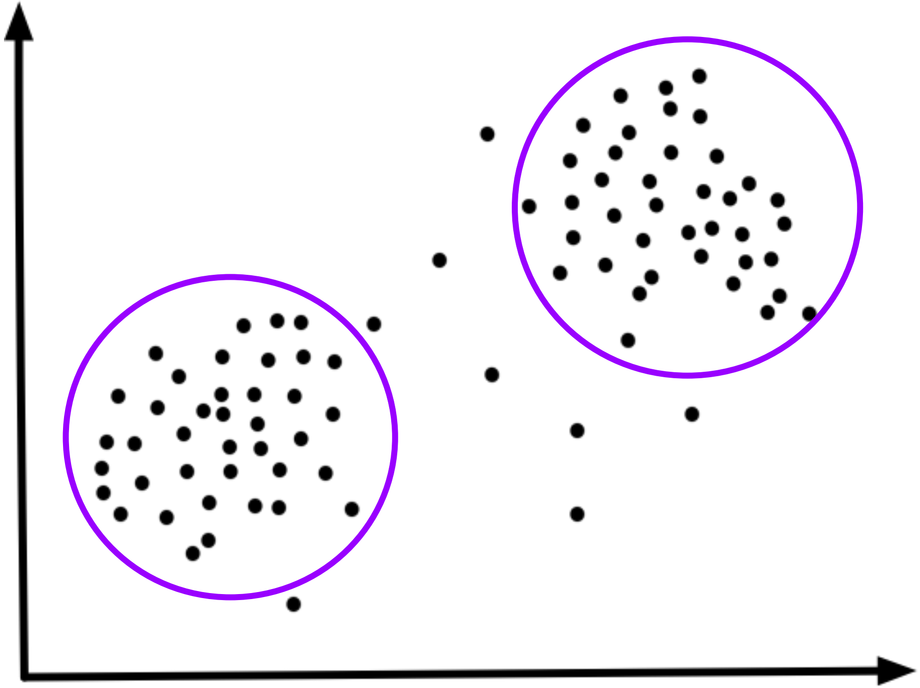
Unsupervised Learning



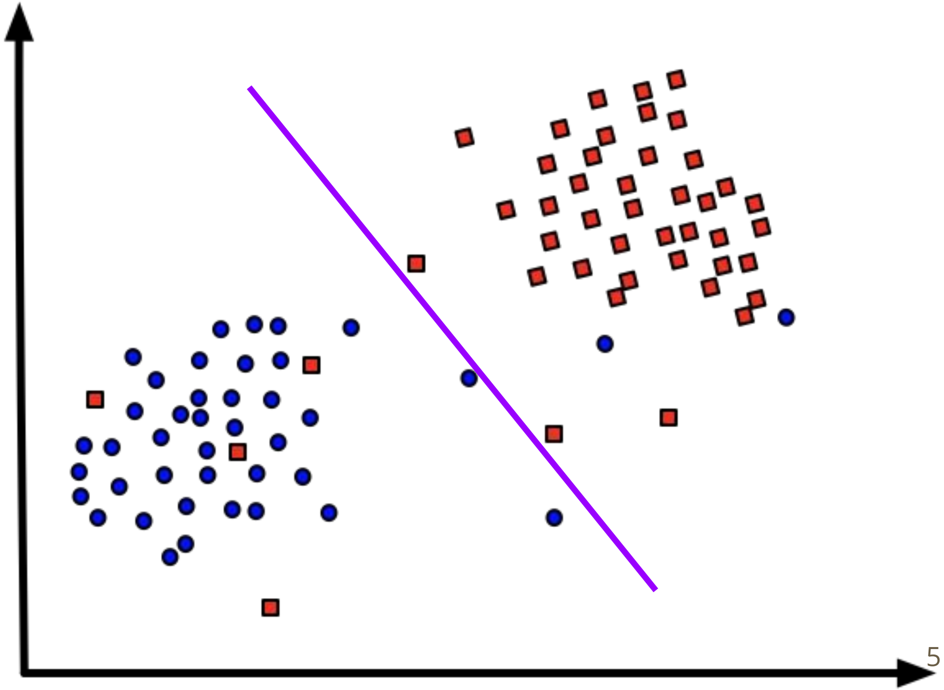
Supervised Learning

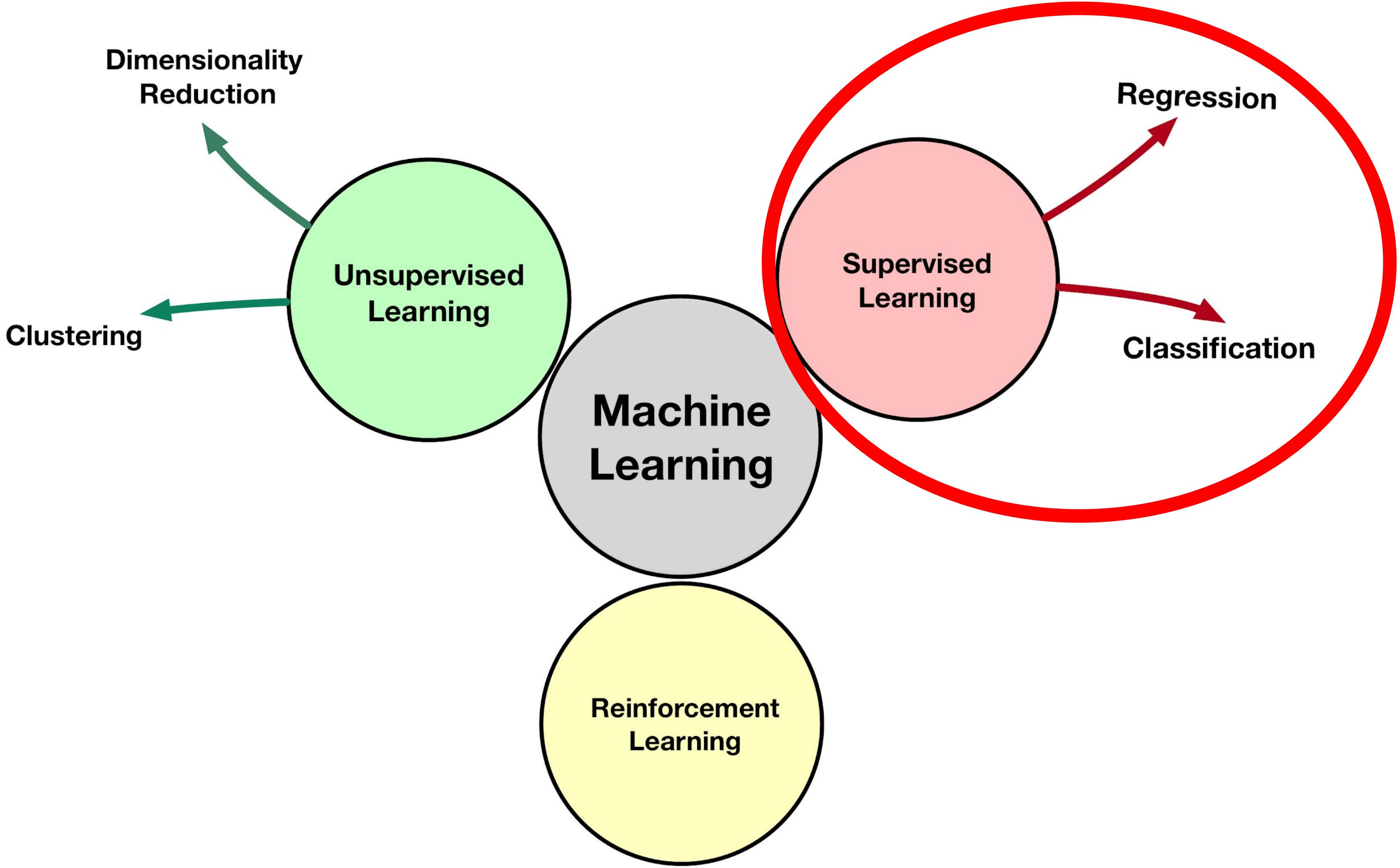


Unsupervised Learning



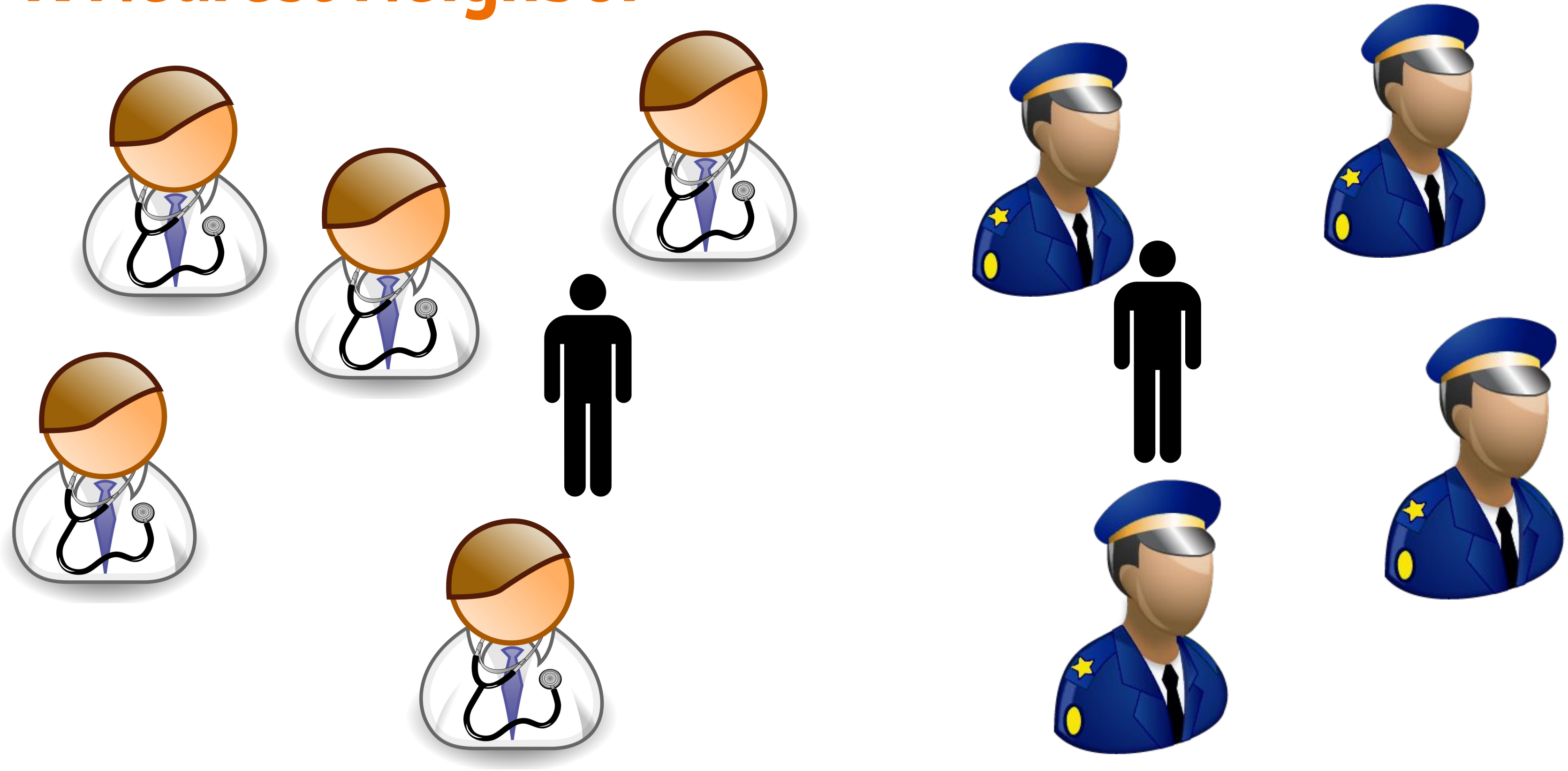
Supervised Learning




Dimensionality Reduction
Regression
Unsupervised Learning
Supervised Learning
Clustering
Classification
Machine Learning
Reinforcement Learning

Machine Learning Algorithms

K Nearest Neighbor



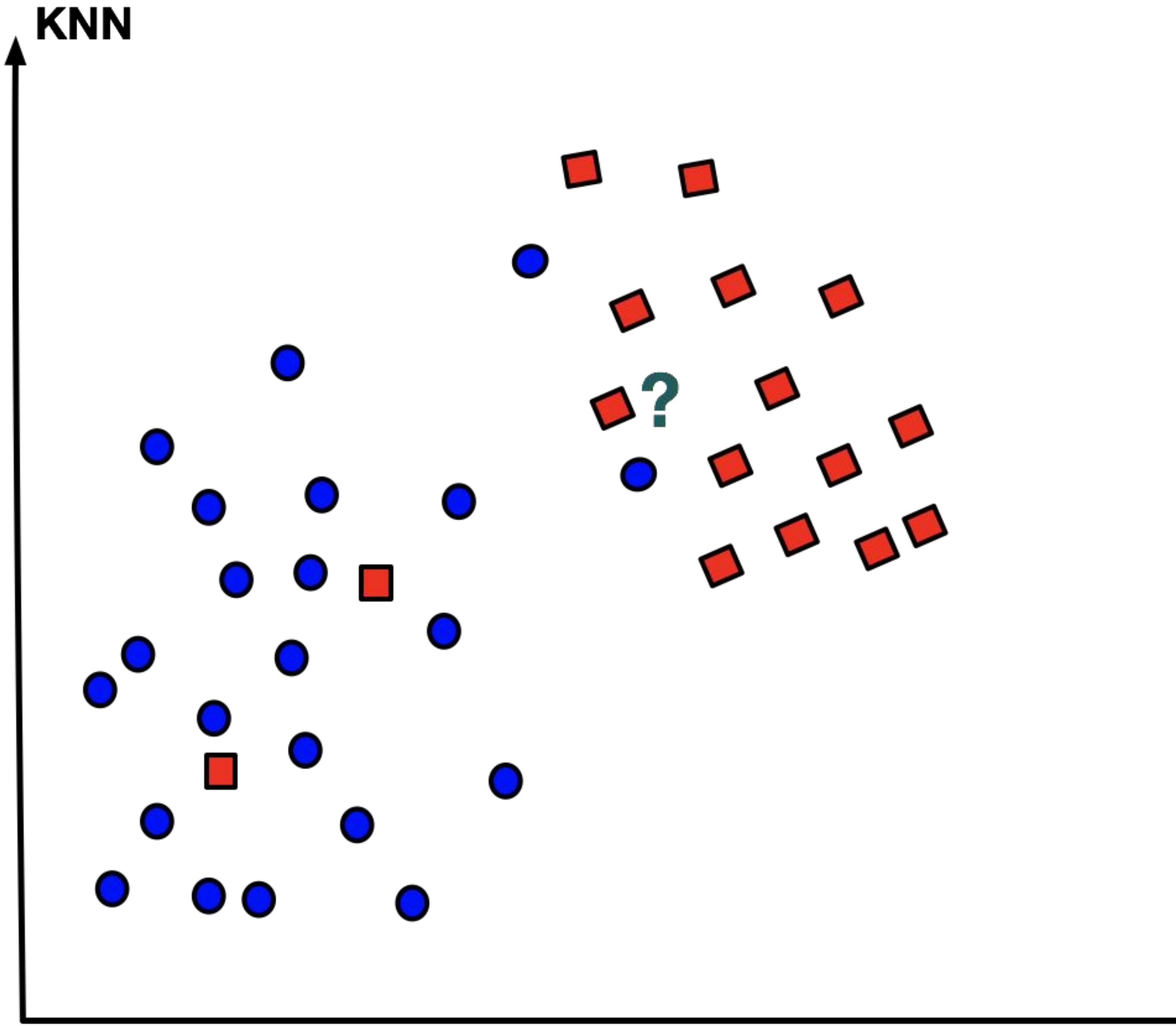
K Nearest Neighbor

- Can be used both for classification and regression.
- Uses **feature similarity** to predict values of any new data points.
- The output based on the majority vote (for classification)
- or mean (or median, for regression)

K Nearest Neighbor

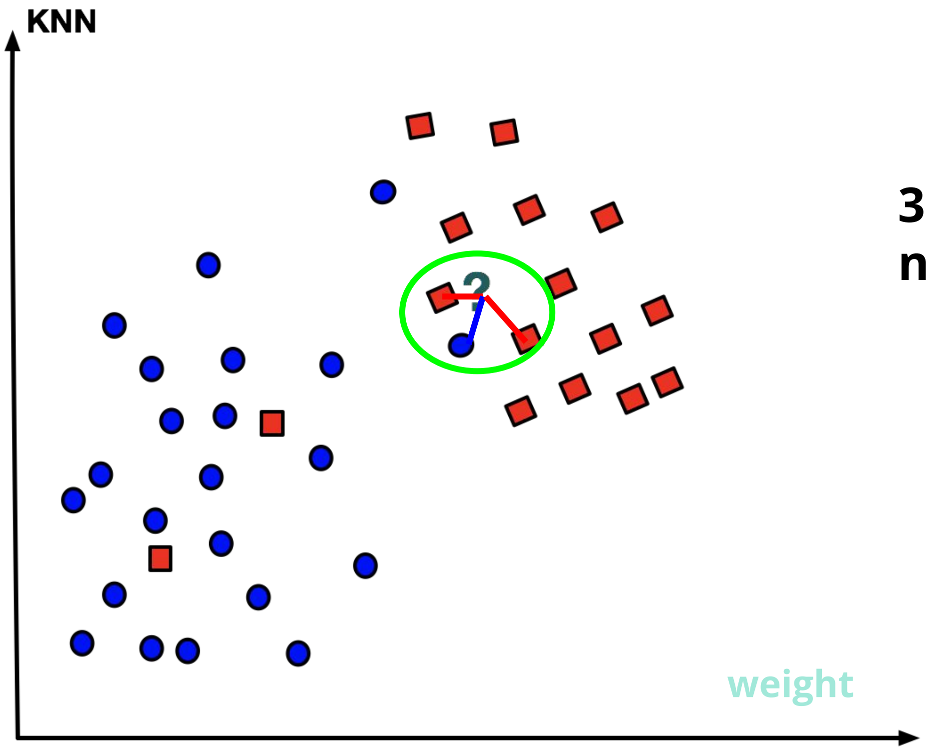
Pick a value k

Use x’s K-Nearest Neighbors to vote on what x’s label should be.



K Nearest Neighbor

hight

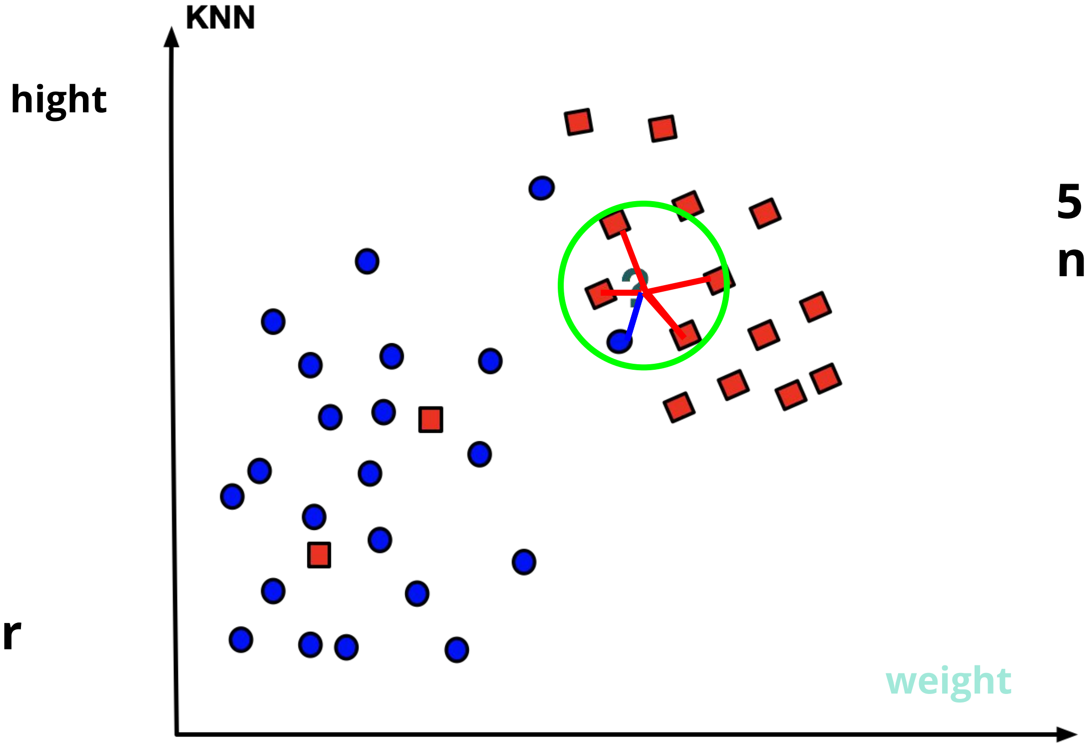


3 nearest neighbours

Gymnast

Basketball player

K Nearest Neighbor

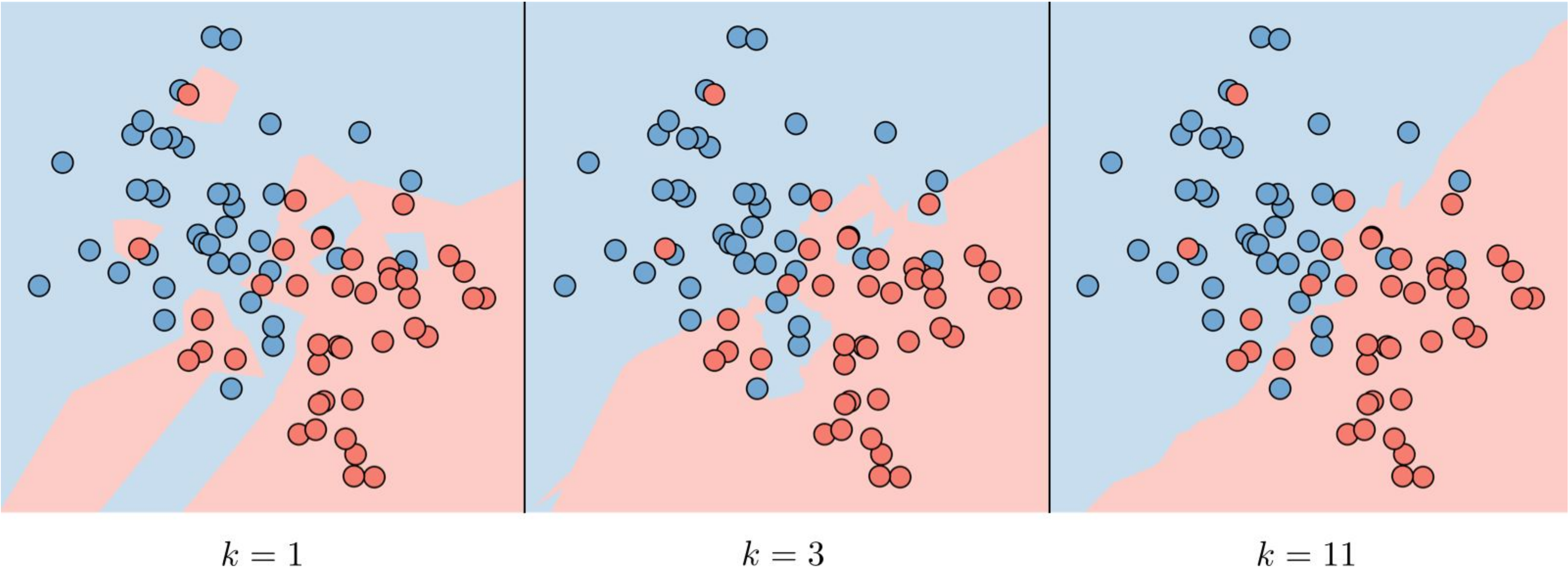


5 nearest neighbours

Gymnast

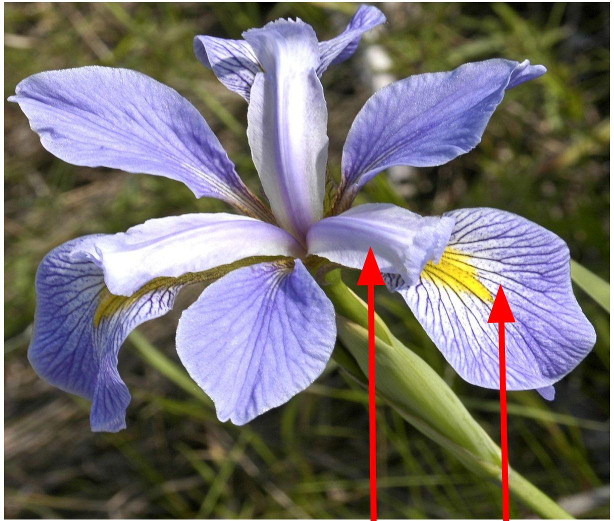
Basketball player

K Nearest Neighbor



https://stanford.edu/~shervine/teaching/cs-229/cheatsheet-supervised-learning

Iris DataSet



Iris virginica

Iris setosa

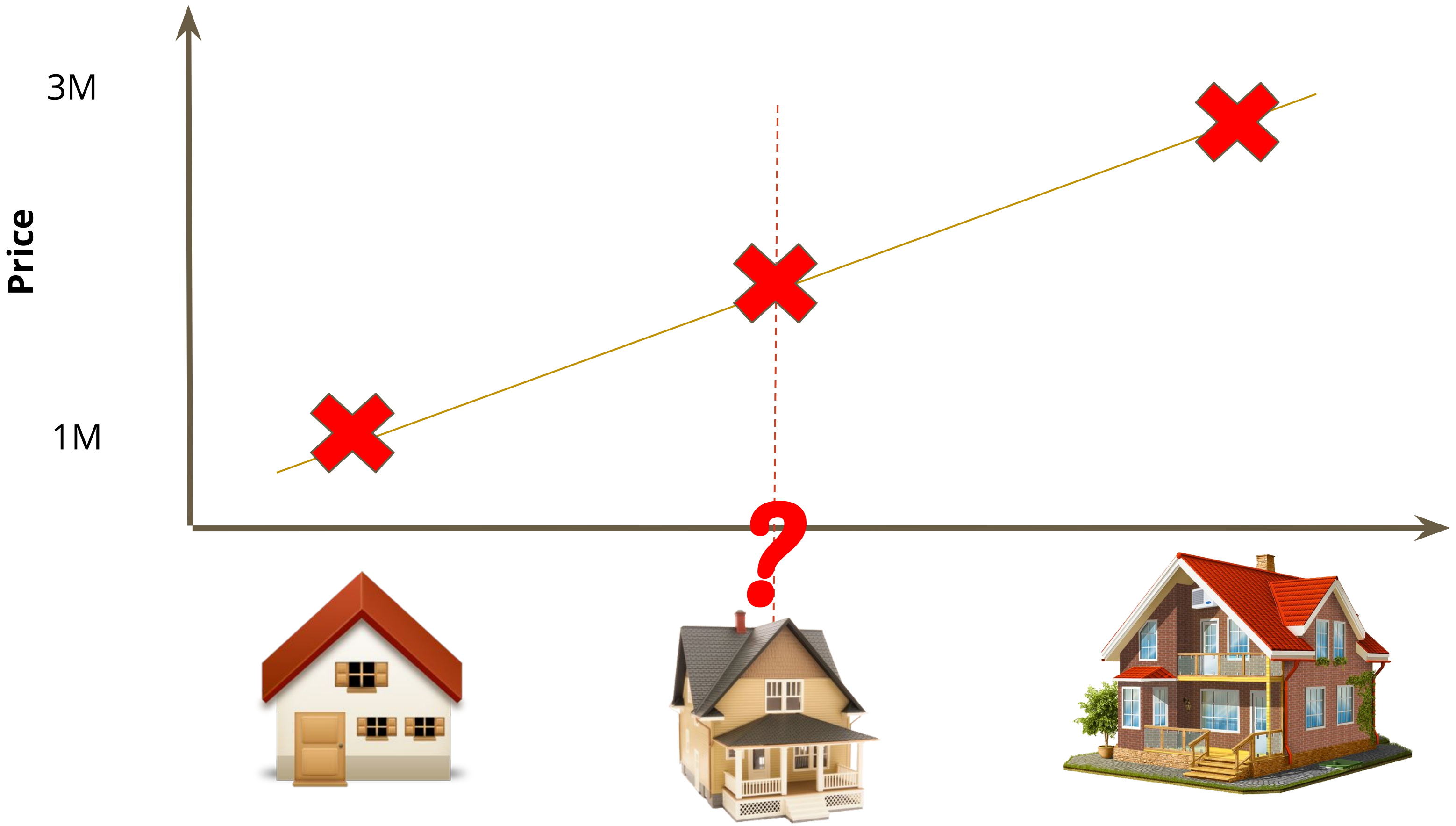
Iris versicolor

Sepal

Petal

https://ca.wikipedia.org/wiki/Conjunt_de_dades_de_flors_Iris

Linear Regression

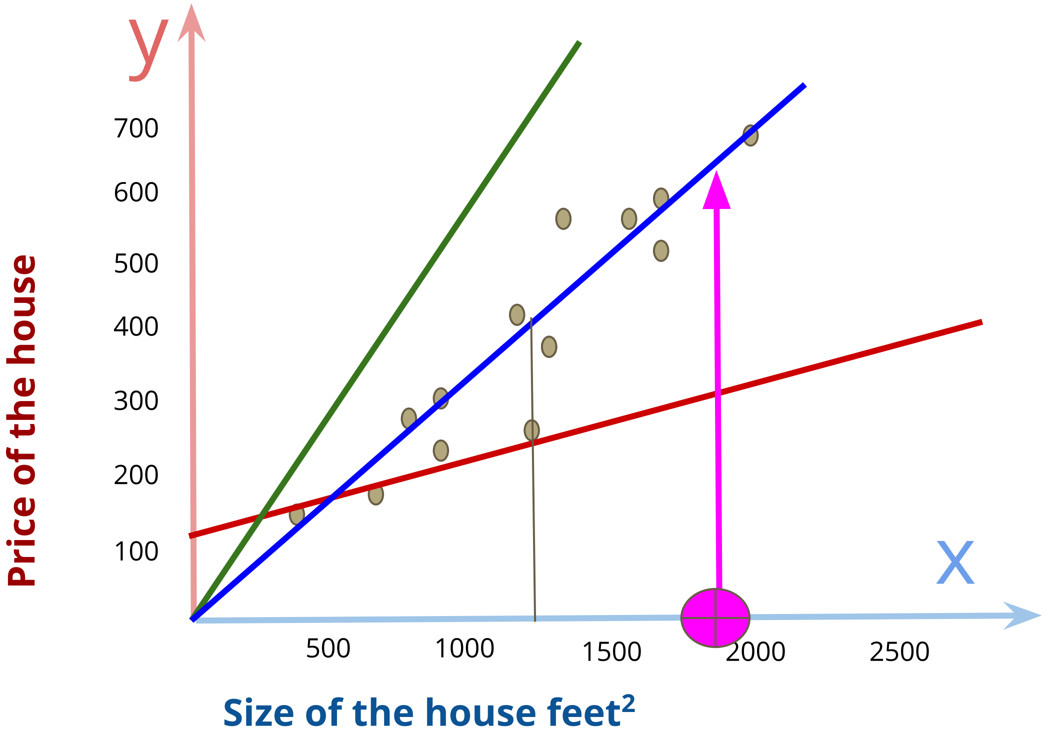

3M
Price
1M
?

Linear Regression

Linear regression is the simplest and most widely used statistical technique

A linear model expresses the target output value in terms of a sum of weighted input variables.

Linear Regression



$$f_i = w_0 x_i + b_0$$

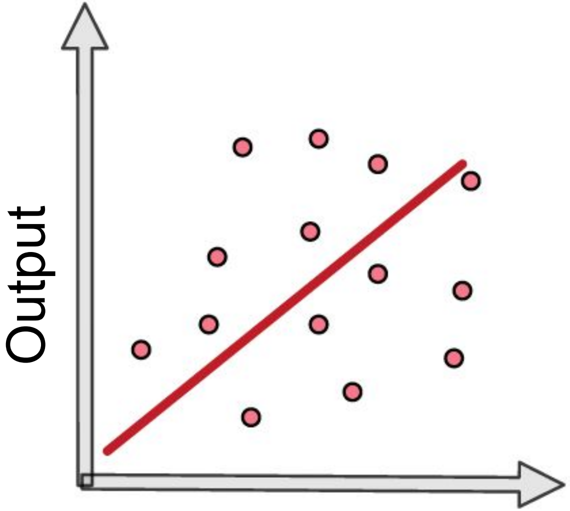
Mean squared error

$$MSE = \frac{1}{N}\sum_{i=1}^{N}(f_i - y_i)^2$$

where N is the number of data points, f_i the value returned by the model and y_i the actual value for data point i.

Univariate versus Multivariate Modeling

Univariate

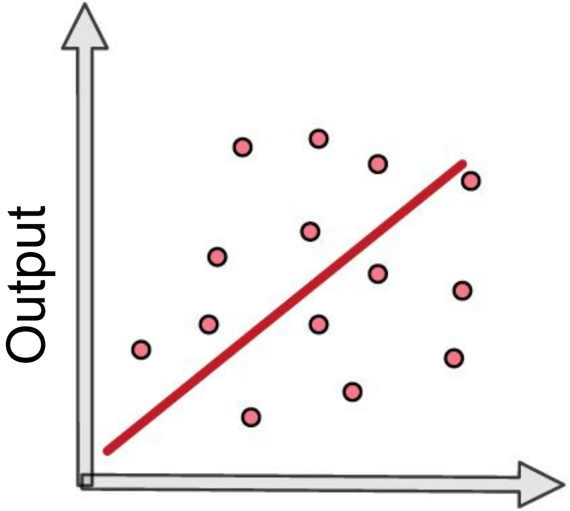


$f_i = w_0 x_{1i} + b_0$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (f_i - y_i)^2$$

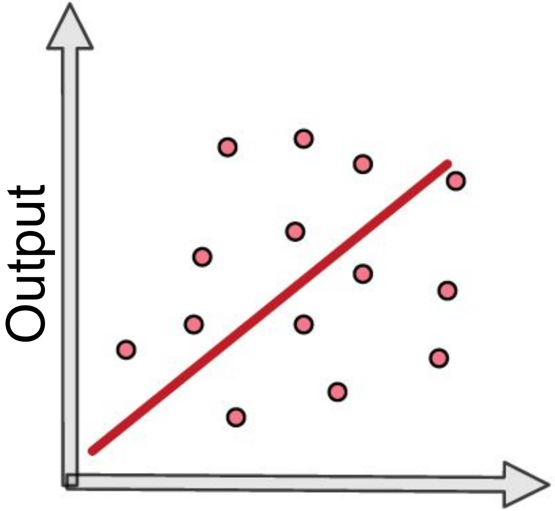
Univariate versus Multivariate Modeling

Univariate



$f_i = w_0 x_{1i} + b_0$

Multivariate



$f_i = w_0 x_{1i} + b_0$
$+ w_1 x_{2i} + b_0$

$+ \ldots\ldots$
$+ w_1 x_{mi} + b_0$

$$MSE = \frac{1}{N} \sum_{i=1}^{N} (f_i - y_i)^2$$

Diabetes

Ten baseline variables:

age, sex, body mass index, average blood pressure, and six blood serum measurements

n = 442 diabetes patients

Target value:

A quantitative measure of disease progression one year after baseline.

Overfitting

Overfitting: Good performance on the training data, poor generalization to other data.

Underfitting: Poor performance on the training data and poor generalization to other data

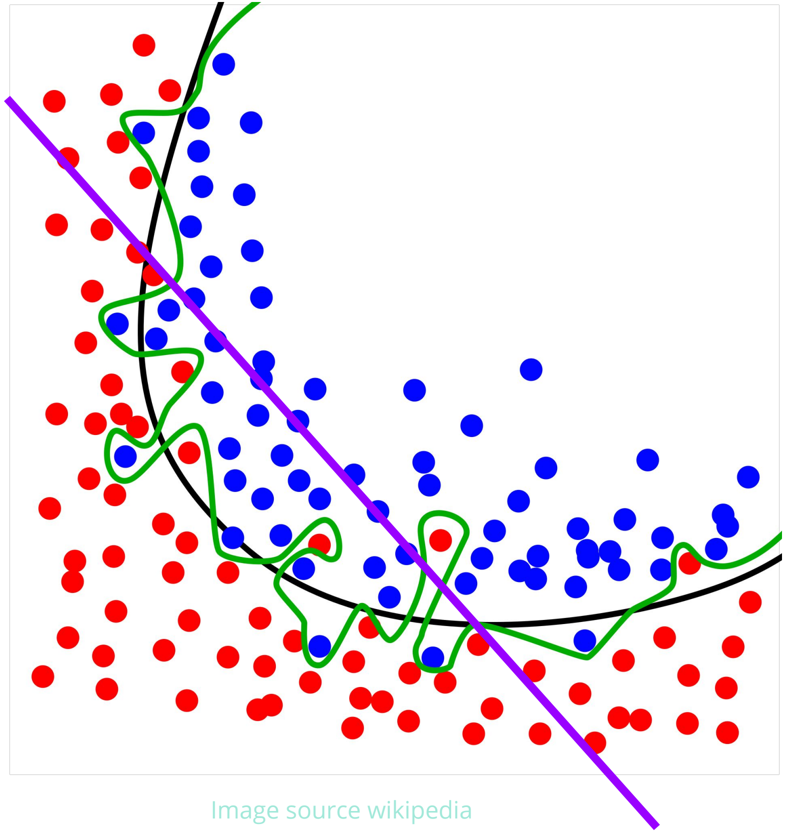
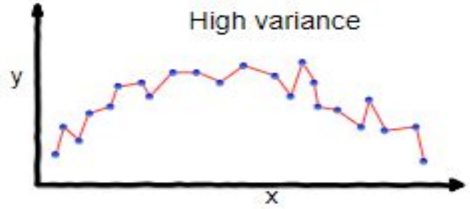
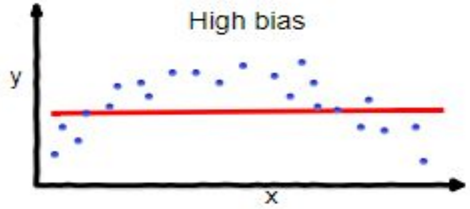


Image source wikipedia

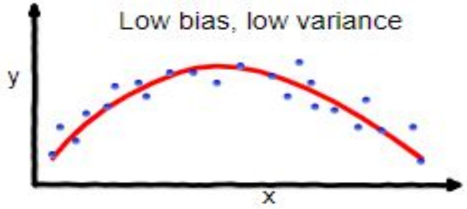
Bias–Variance Tradeoff



overfitting

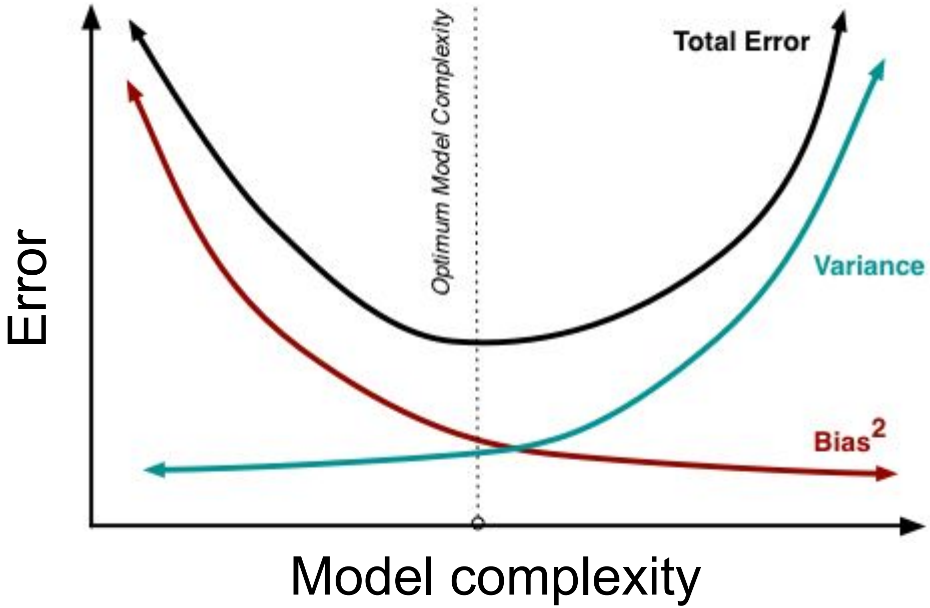


underfitting



Good balance

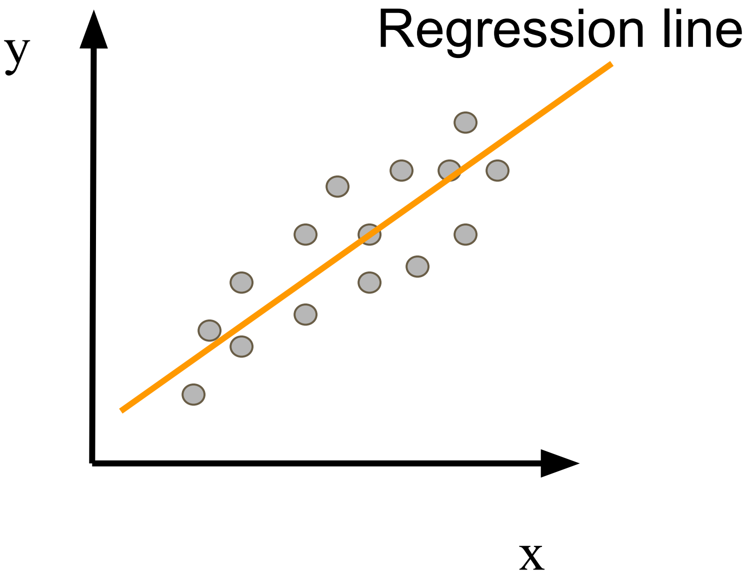
Bias–Variance Tradeoff



Logistic Regression

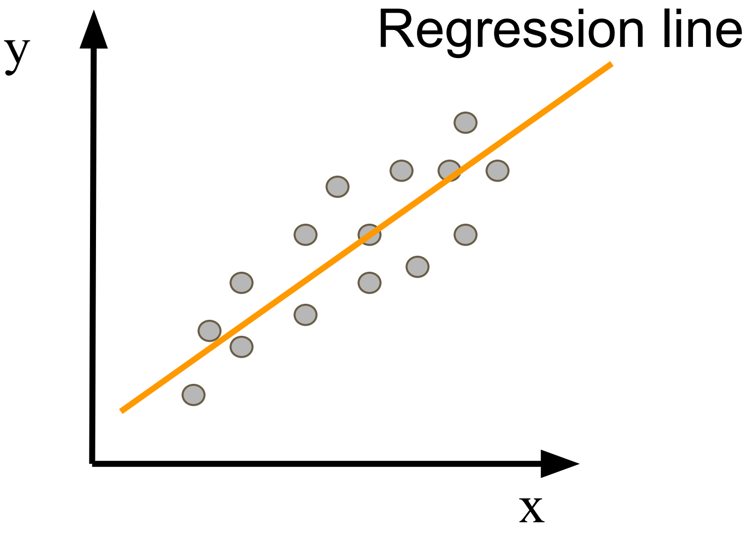
Linear versus Logistic Regression

Regression (linear regression)

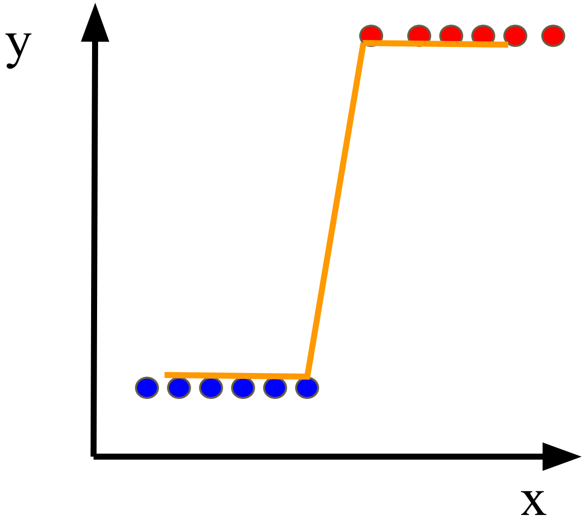


Linear versus Logistic Regression

Regression (linear regression)

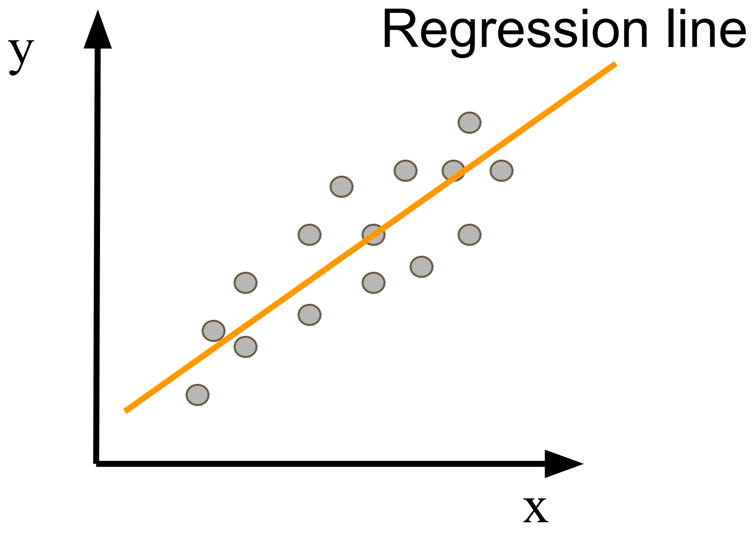


Classification (logistic regression)

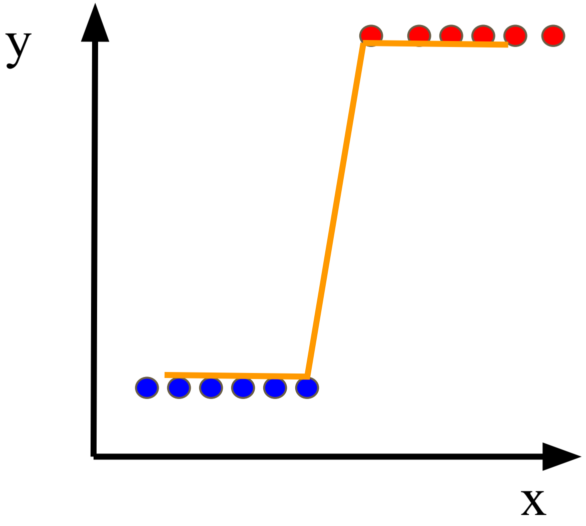


Linear versus Logistic Regression

Regression (linear regression)



Classification (logistic regression)



We need a smooth function that gives us this trend (Sigmoid Function)

Linear versus Logistic Regression

Linear regression

$$f_i = \Sigma_i w_i x_i + b_0$$

Logistic regression

$$f_i = \frac{1}{1+e^{\Sigma_i w_i x_i + b_0}}$$

W will be identified to minimize cost

$$\mathrm{Cost}(w)=\mathrm{function}(w,f_i,y_i)$$

Bayes rule and Naive Bayes classifier

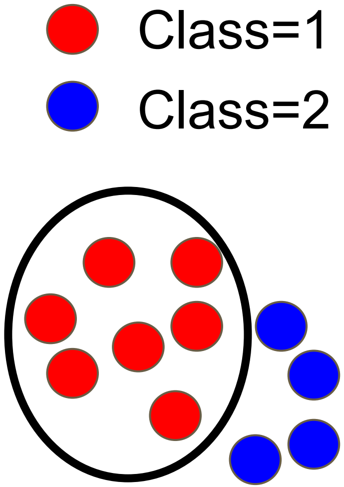
What we know when training a model

$$p(X_1 = \mathbf{x}_1 \mid Class = 1)$$

$$p(X_2 = \mathbf{x}_2 \mid Class = 1)$$

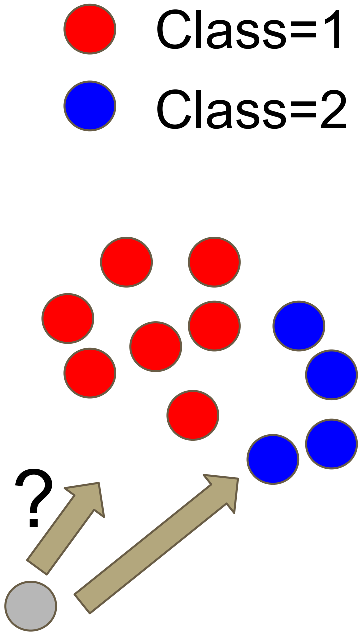
$$\vdots$$

$$p(X_m = \mathbf{x}_m \mid Class = 1)$$



What do we care about?

$$p(\mathrm{Class}=1 \mid X_1=x_1, X_2=x_2, \ldots, X_m=x_m) = ?$$



Bayes rule is useful to figure out the relationship

$$p(A|B)p(B)=p(B|A)p(A)$$

$$p(Class=1|X_1=x_1,X_2=x_2,\ldots,X_m=x_m) *$$

$$p(X_1=x_1,X_2=x_2,\ldots,X_m=x_m) =$$

$$p(X_1=x_1,X_2=x_2,\ldots,X_m=x_m|Class=1)p(Class=1)$$

The relationship looks complicated

$$\mathbf{WWW}*p(X_1=x_1, X_2=x_2, \ldots, X_m=x_m) =$$

$$p(X_1=x_1, X_2=x_2, \ldots, X_m=x_m \mid Class=1)\,p(Class=1)$$

WWW: What We Want

$p(Class=1)$ **: easy to calculate**

$$p(Class = i) = \frac{N_i}{\Sigma_i^C N_i}$$

Naive Bayes

Naive Bayes classifier is called ***Naive*** as it assumes each feature will independently contribute in prediction of a class for each data point

$$p(X1=x1, X2=x2, \ldots, Xm=xm) = p(X1=x1)\,p(X2=x2)\ldots p(Xm=xm)$$

$$p(X1=x1, X2=x2, \ldots, Xm=xm \mid Class=1) =$$
$$p(X1=x1 \mid Class=1)\,p(X2=x2 \mid Class=1)\ldots p(Xm=xm \mid Class=1)$$

Bayesian approach for problem solving

Bayesian versus frequentist

Frequentist:
- Variation of data and derived quantities in terms of fixed model parameters

Bayesian:
- Variation of beliefs about parameters in terms of fixed observed data

Note. The difference become clear in complicated problems like *Interpretation of uncertainty.*

37

Bayesian versus frequentist

Frequentist:

- If an experiment is repeated many times, in 95% of these cases the computed confidence interval will contain the true theta.

Note. In general, a frequentist 95% confidence interval is not 95% likely to contain the true value. This very common mistake is a Bayesian interpretation of a frequentist construct.

38

Bayesian versus frequentist

Frequentist:

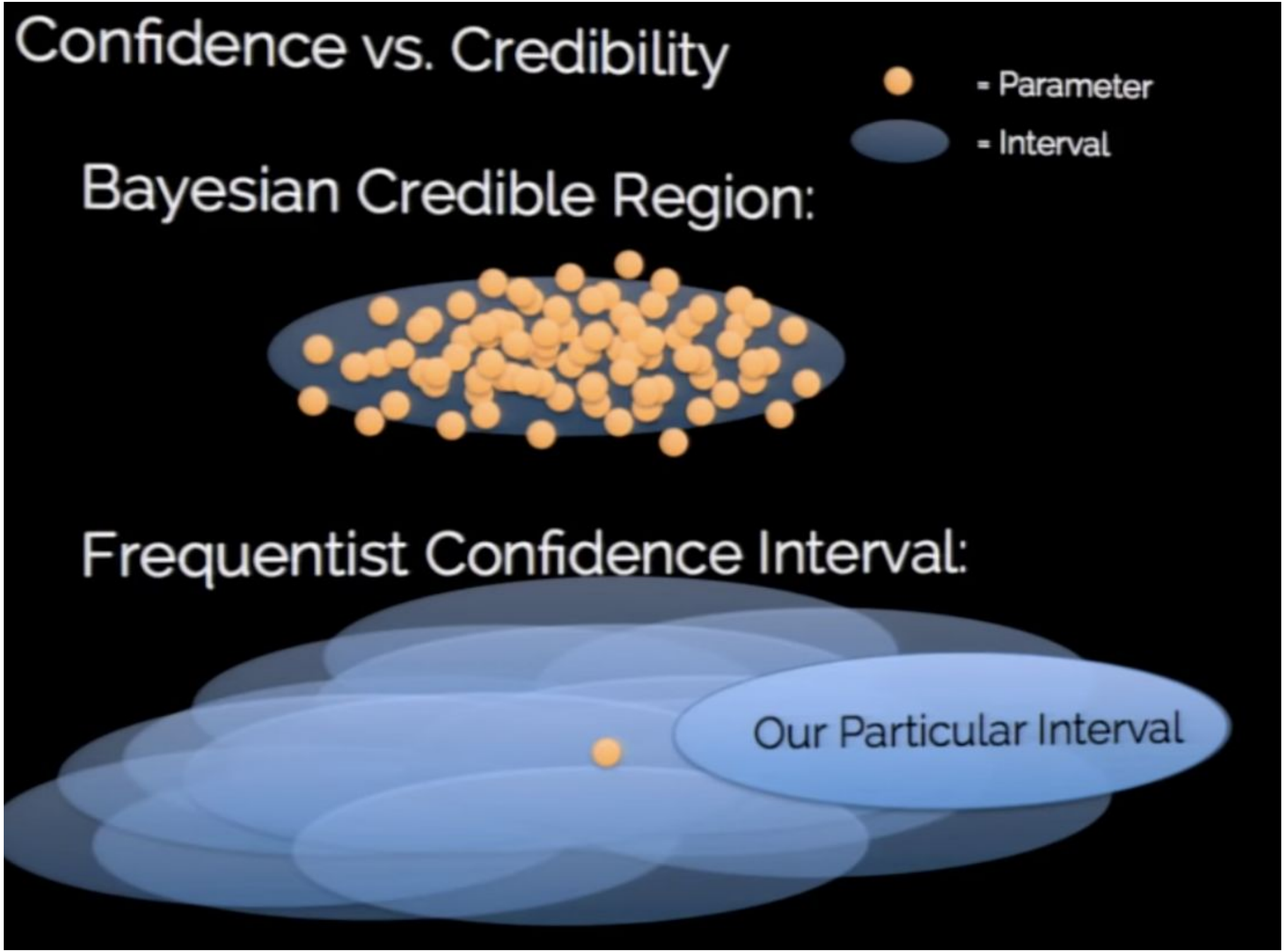
- If an experiment is repeated many times, in 95% of these cases the computed confidence interval will contain the true theta.

Note. In general, a frequentist 95% confidence interval is not 95% likely to contain the true value. This very common mistake is a Bayesian interpretation of a frequentist construct.

Bayesian:

- Given our observed data there is a 95% probability that the value of theta lies within the credible region

39



https://speakerdeck.com/jakevdp/

Conversation between a (frequentist)statistician and a scientist:

Statistician:

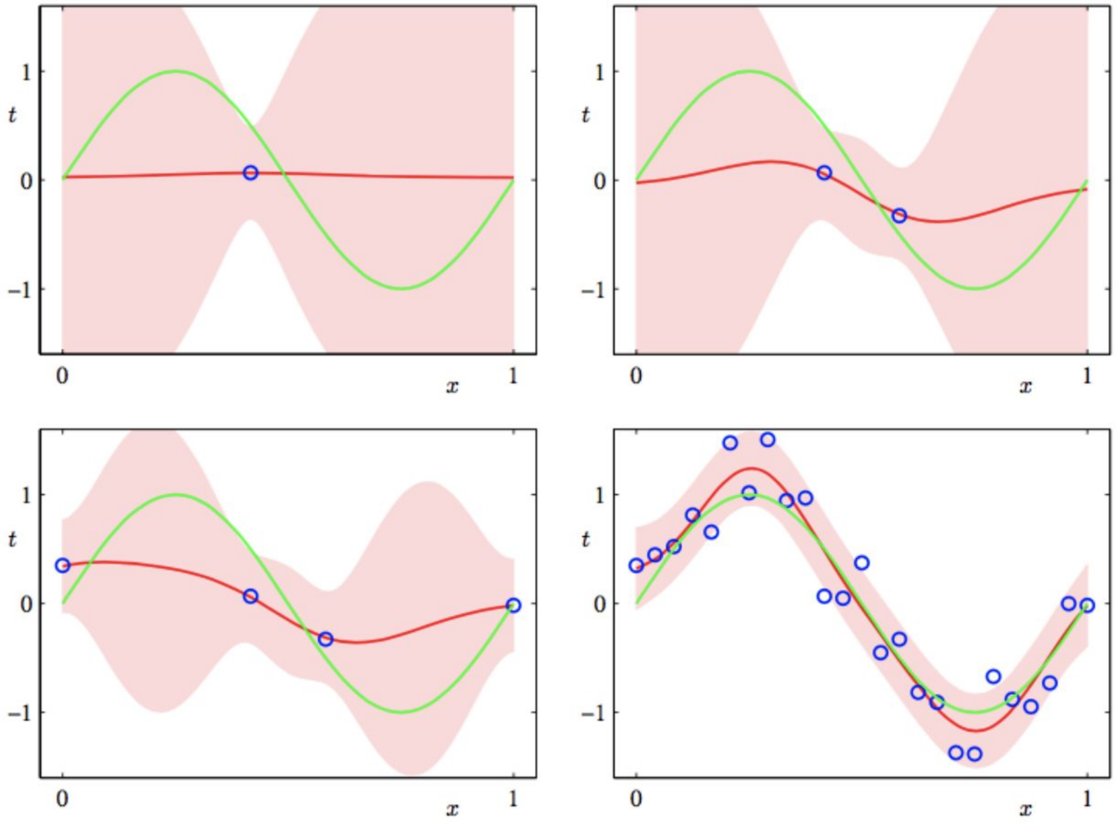
- 95% of such confidence intervals in repeated experiments will contain the true value. (Referring to chance in that context is meaningless. The 95% refers to the interval itself)
    - The long term limiting frequency of the procedure or constructing this interval ensures that 95% of the resulting ensemble of intervals contains the value.

Scientist:

- So there is a 95% chance that the value is in this interval?

41

Example in regression

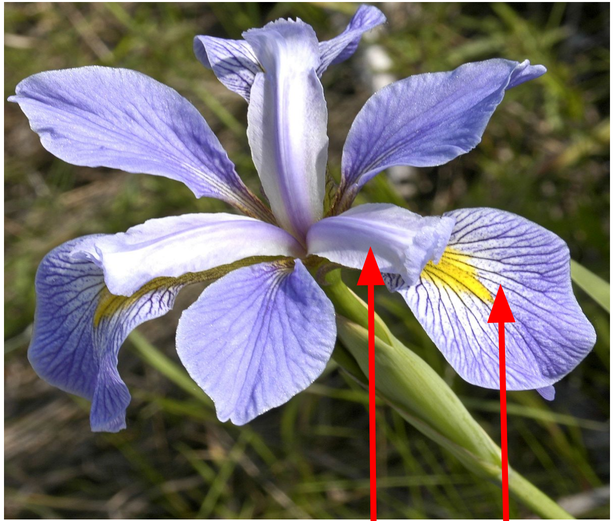


Pattern Recognition and machine learning, Bishop.

42

Thanks

Iris DataSet



Iris virginica

Sepal

Petal

Iris setosa

Iris versicolor

https://ca.wikipedia.org/wiki/Conjunt_de_dades_de_flors_Iris

Diabetes

Ten baseline variables:

age, sex, body mass index, average blood pressure, and six blood serum measurements

n = 442 diabetes patients

Target value:

A quantitative measure of disease progression one year after baseline.

Breast cancer dataset

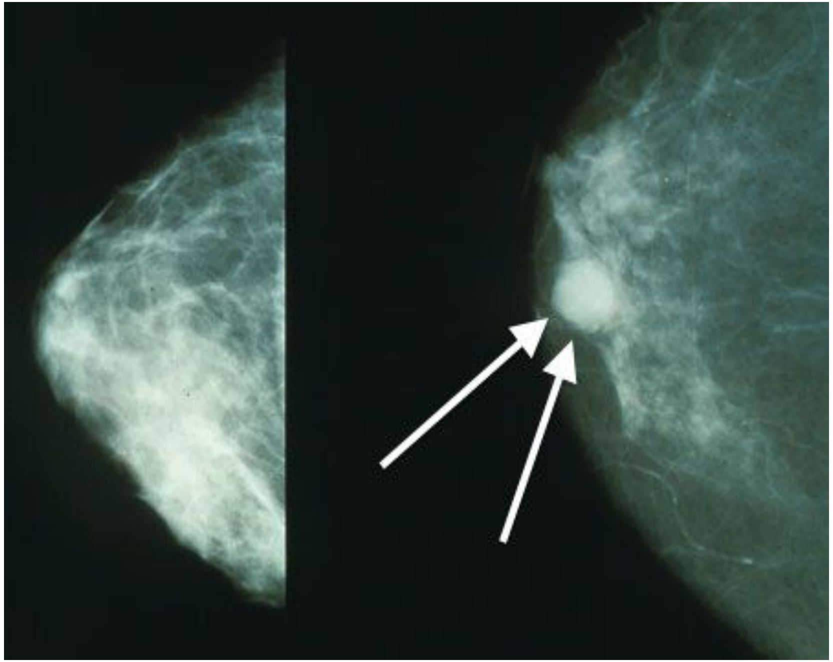
The breast cancer dataset is a classic and very easy binary classification dataset.

Features :

Computed from a digitized image of a fine needle aspirate (FNA) of a breast mass.

Target values:

Benign /Malignant



Extra useful information

Useful links

Installation instructions

- [scikit-learn]
- [IPython]

Data Sets

- [scikit-learn DataSet]

scikit-learn: machine learning in Python :

- https://scikit-learn.org/stable/

Useful cheat sheets:

- https://www.analyticsvidhya.com/blog/2017/02/top-28-cheat-sheets-for-machine-learning-data-science-probability-sql-big-data/