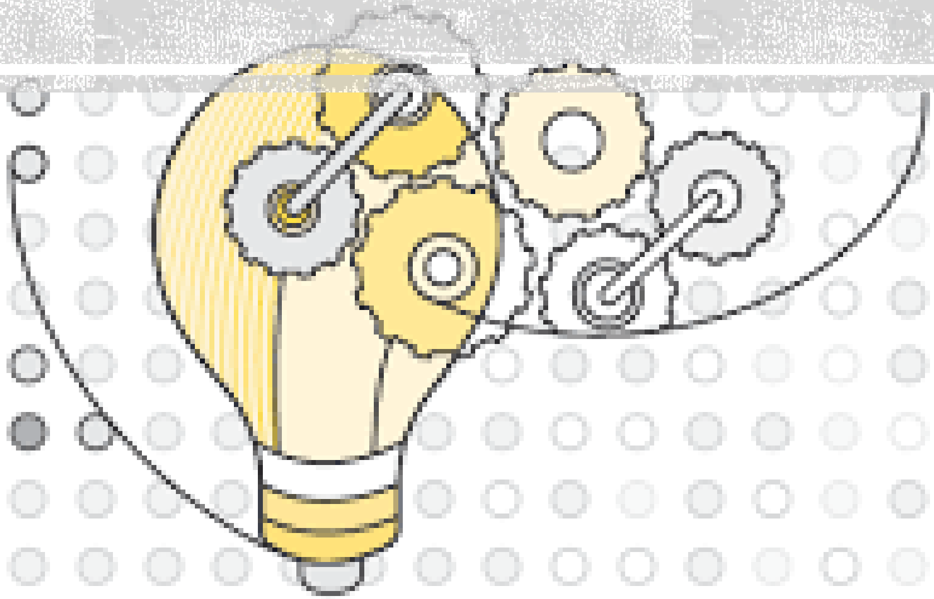
INTRODUCTION



- Class webpage:
  `http://people.sutd.edu.sg/~nengli_lim/teaching/`
- Textbook:
  Pattern Recognition and Machine Learning, Bishop 2006
- Grading:
  - HW: 50% (5 x 10%), Midterm: 20%, Final: 30%
  - Participation bonus/penalty: -2% to 2% of final score

ACKNOWLEDGEMENTS

- MIT 6.036 Introduction to Machine Learning
- SUTD 50.007 Machine Learning (Alex Binder)
- Stanford CS229 Machine Learning

WHAT IS MACHINE LEARNING?

Hard-Coded

Trained

Giving computers the ability to learn without being explicitly programmed – Arthur Samuel (1959)

TYPES OF MACHINE LEARNING

Supervised Learning

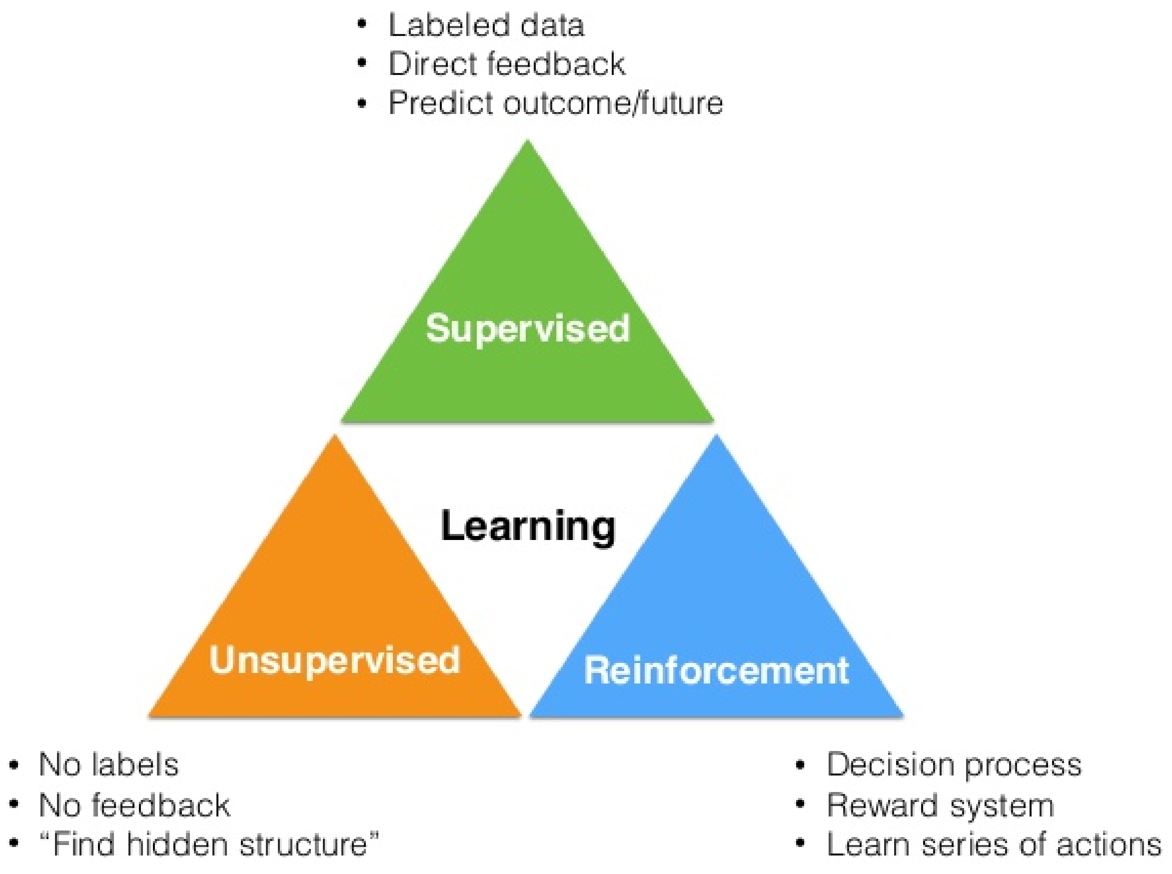
TYPES OF MACHINE LEARNING

Unsupervised Learning

TYPES OF MACHINE LEARNING

Reinforcement Learning

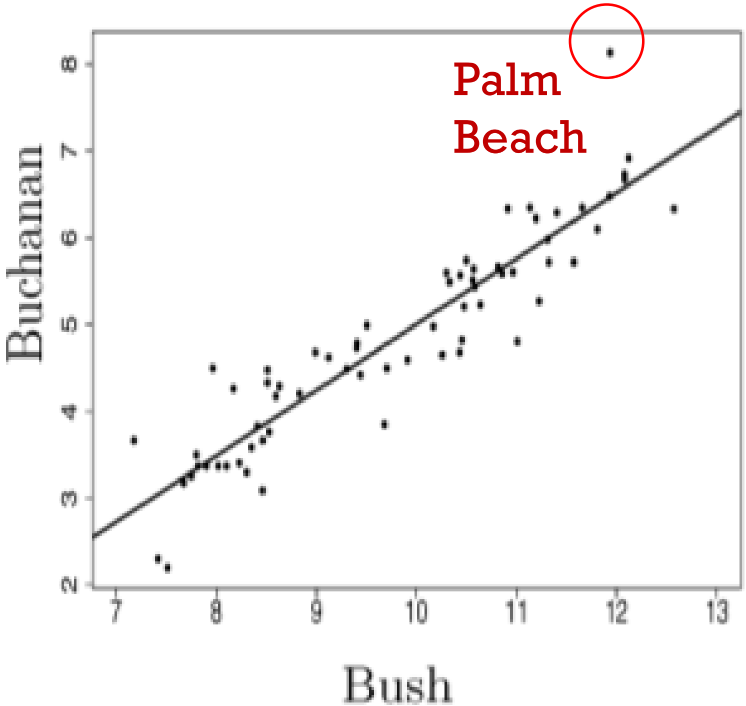
Machine Learning



SUPERVISED LEARNING

Learning a function
$y = f(x)$
$x \in \mathbb{R}$
$y \in \mathbb{R}$

Regression (Linear)



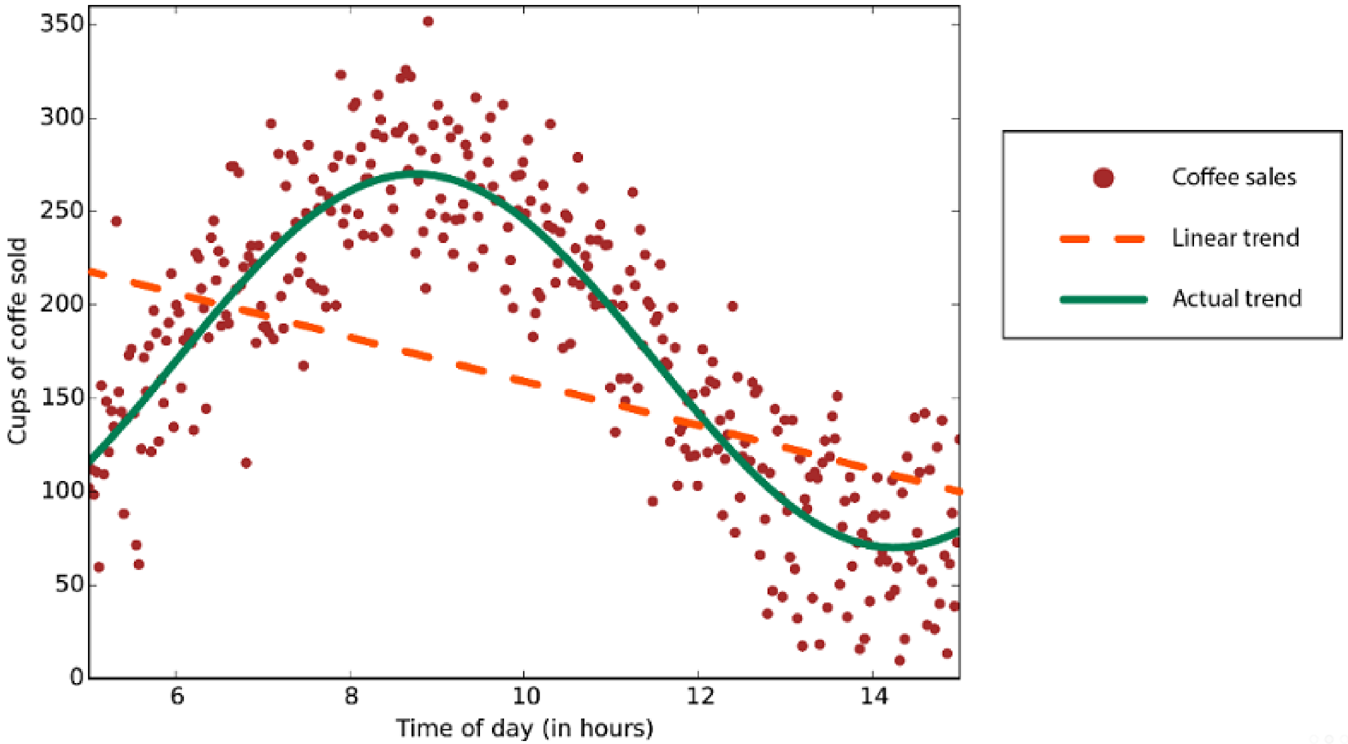
2000 USA Presidential Elections.

Votes for Buchanan and Bush in cities of Florida on a log scale.

https://en.wikipedia.org/wiki/United_States_presidential_election_in_Florida,_2000
http://andrys.com/flballot.html

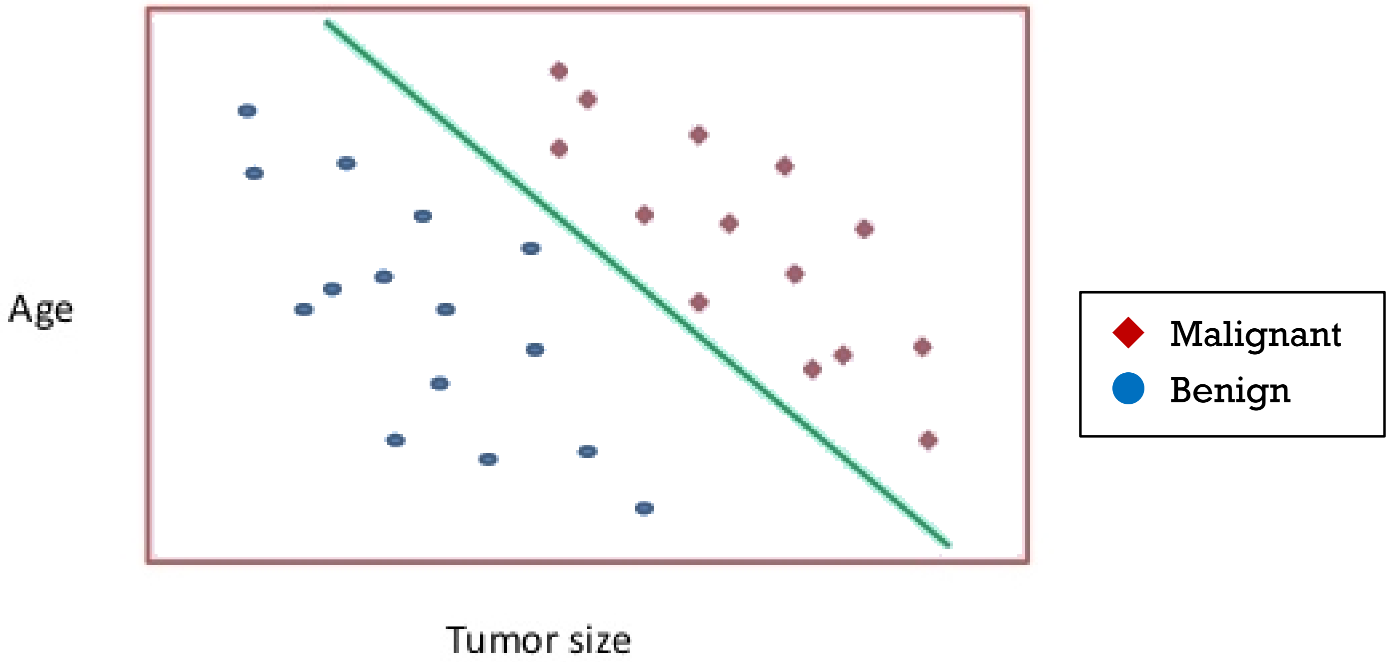
SUPERVISED LEARNING

Regression (Non-linear)



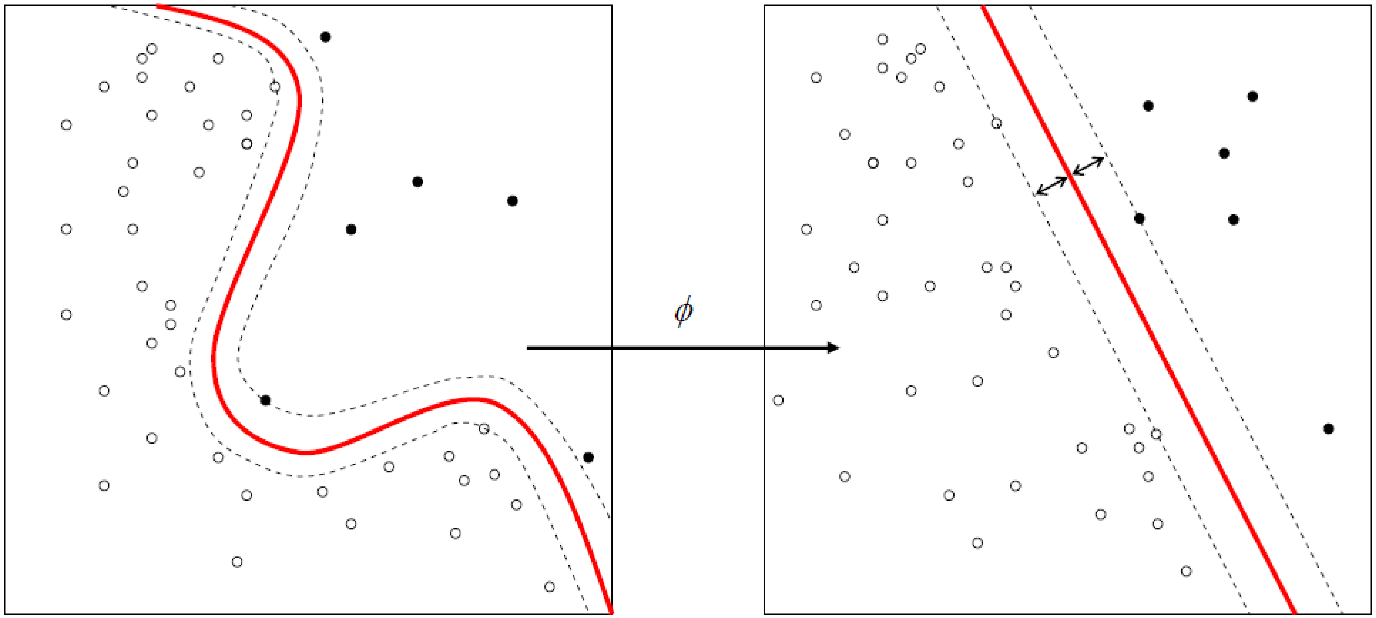
SUPERVISED LEARNING

Classification (Linear)



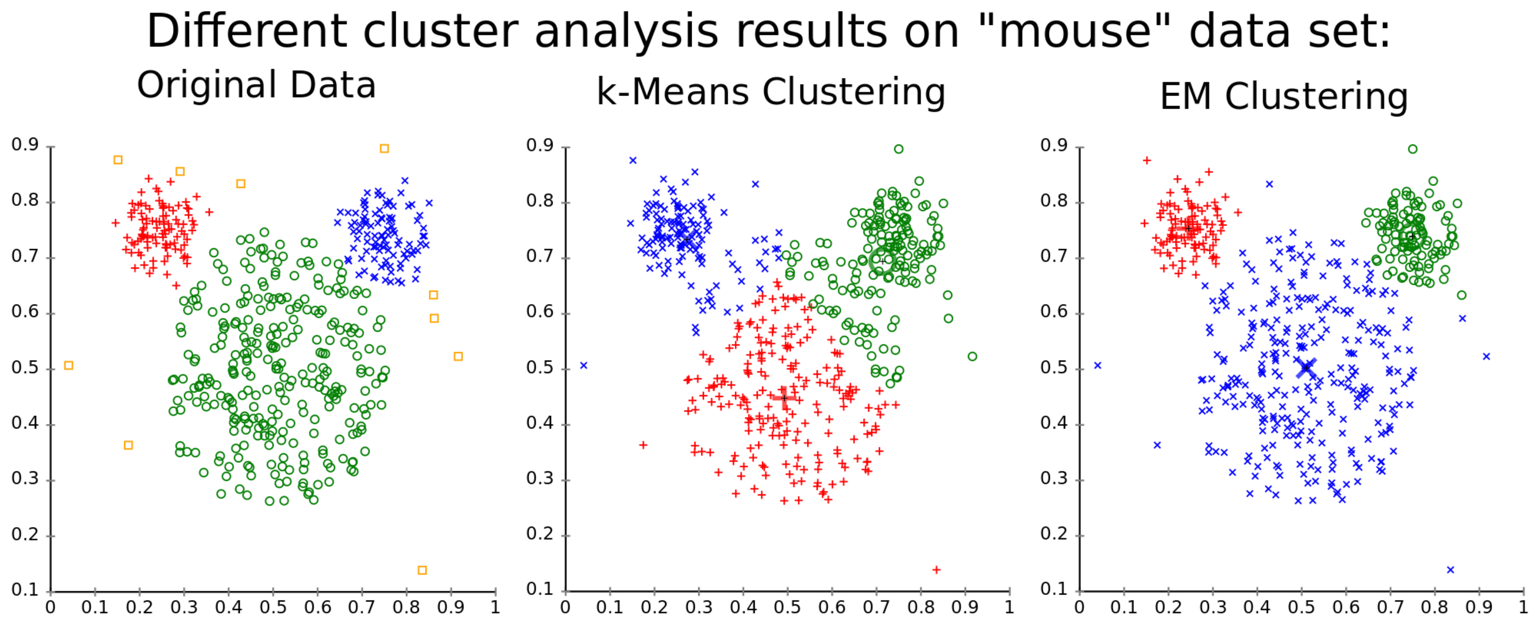
SUPERVISED LEARNING

Classification (Non-linear)



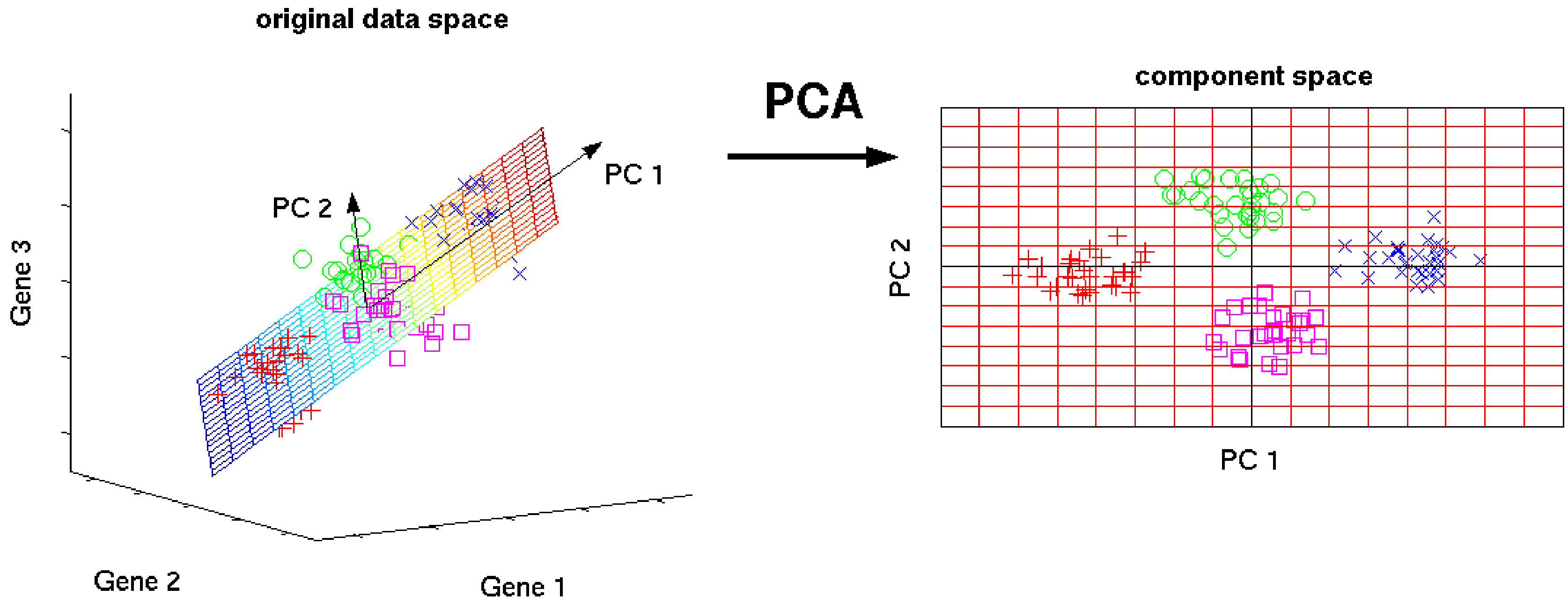
UNSUPERVISED LEARNING

Clustering



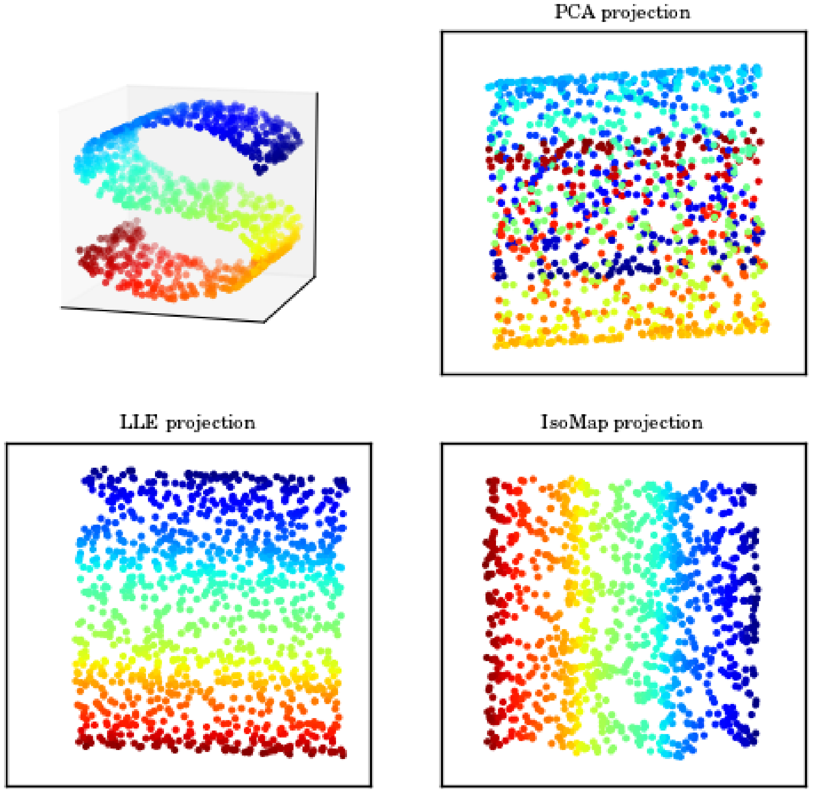
UNSUPERVISED LEARNING

Dimensionality Reduction: Subspace Learning

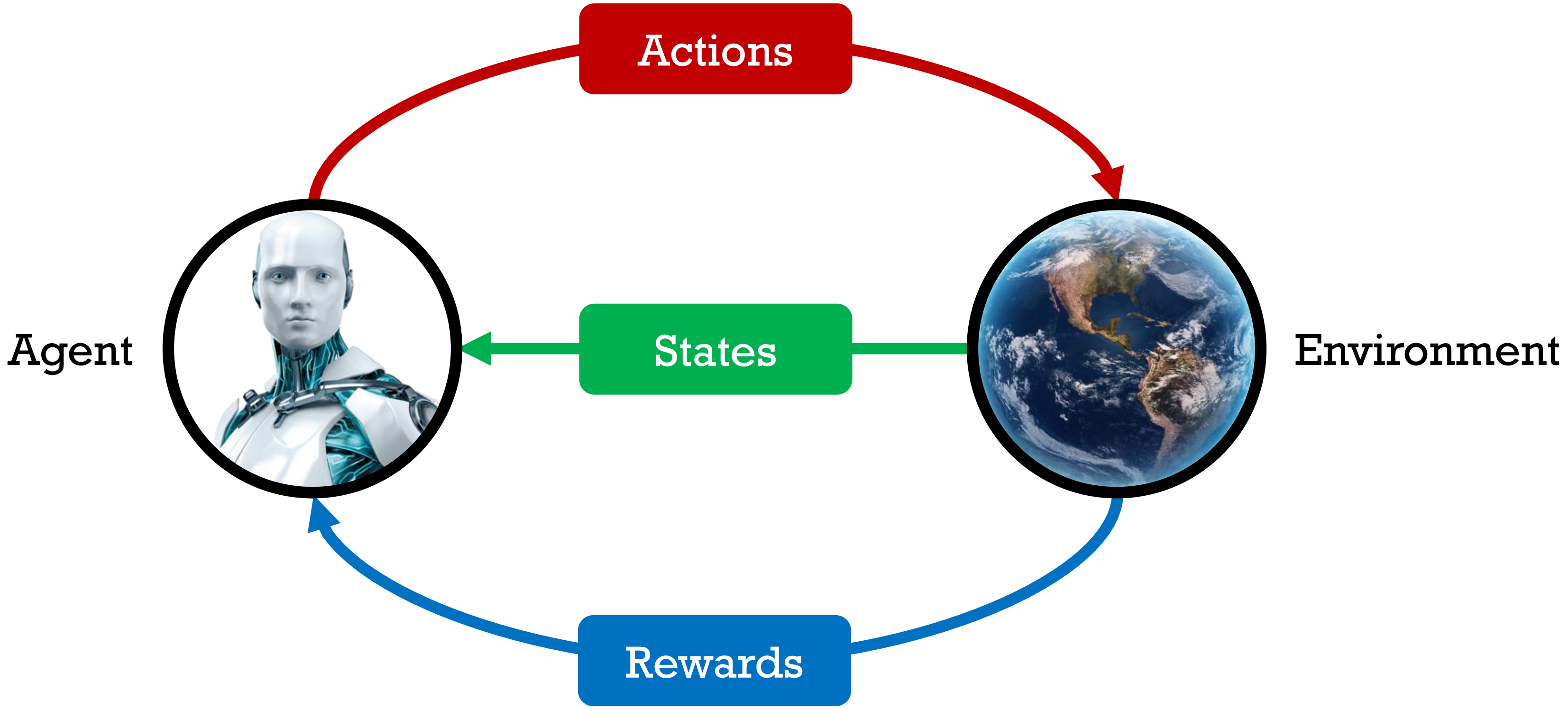


UNSUPERVISED LEARNING

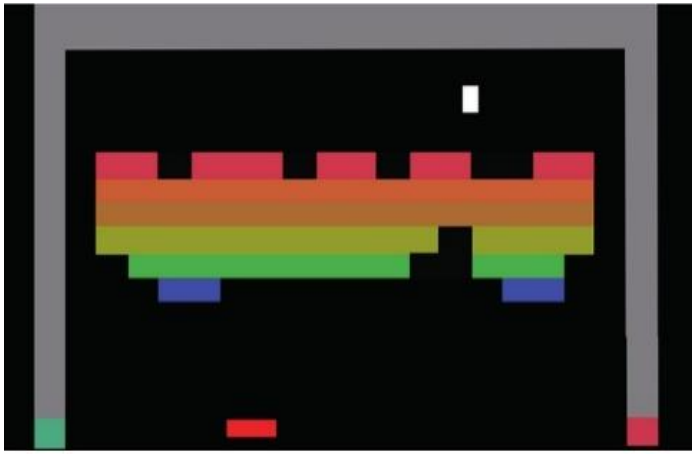
Dimensionality Reduction: Manifold Learning

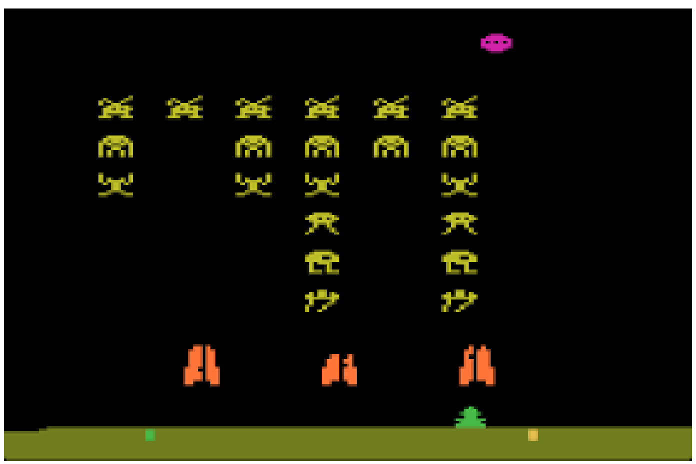


REINFORCEMENT LEARNING



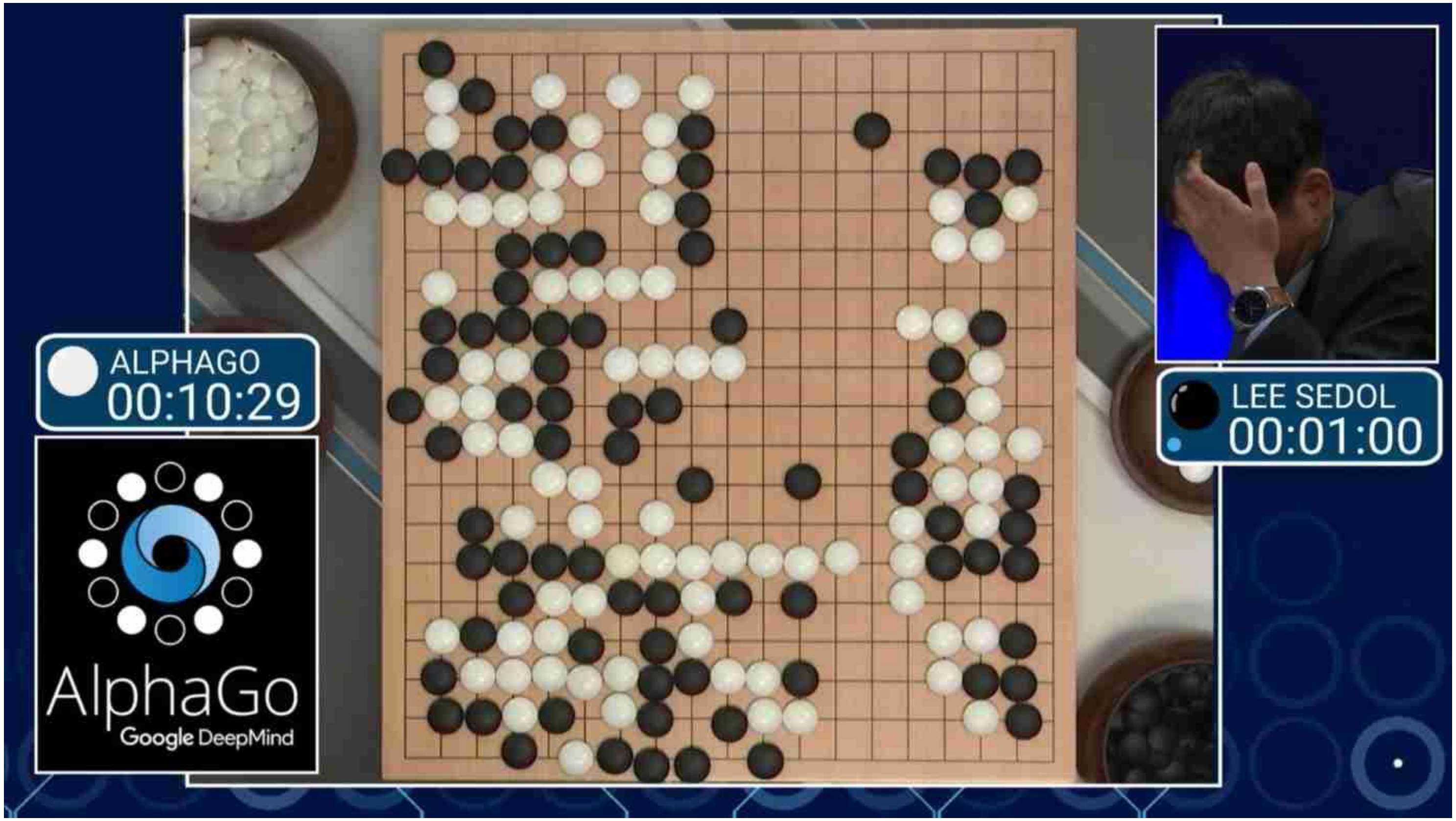
ATARI GAMES





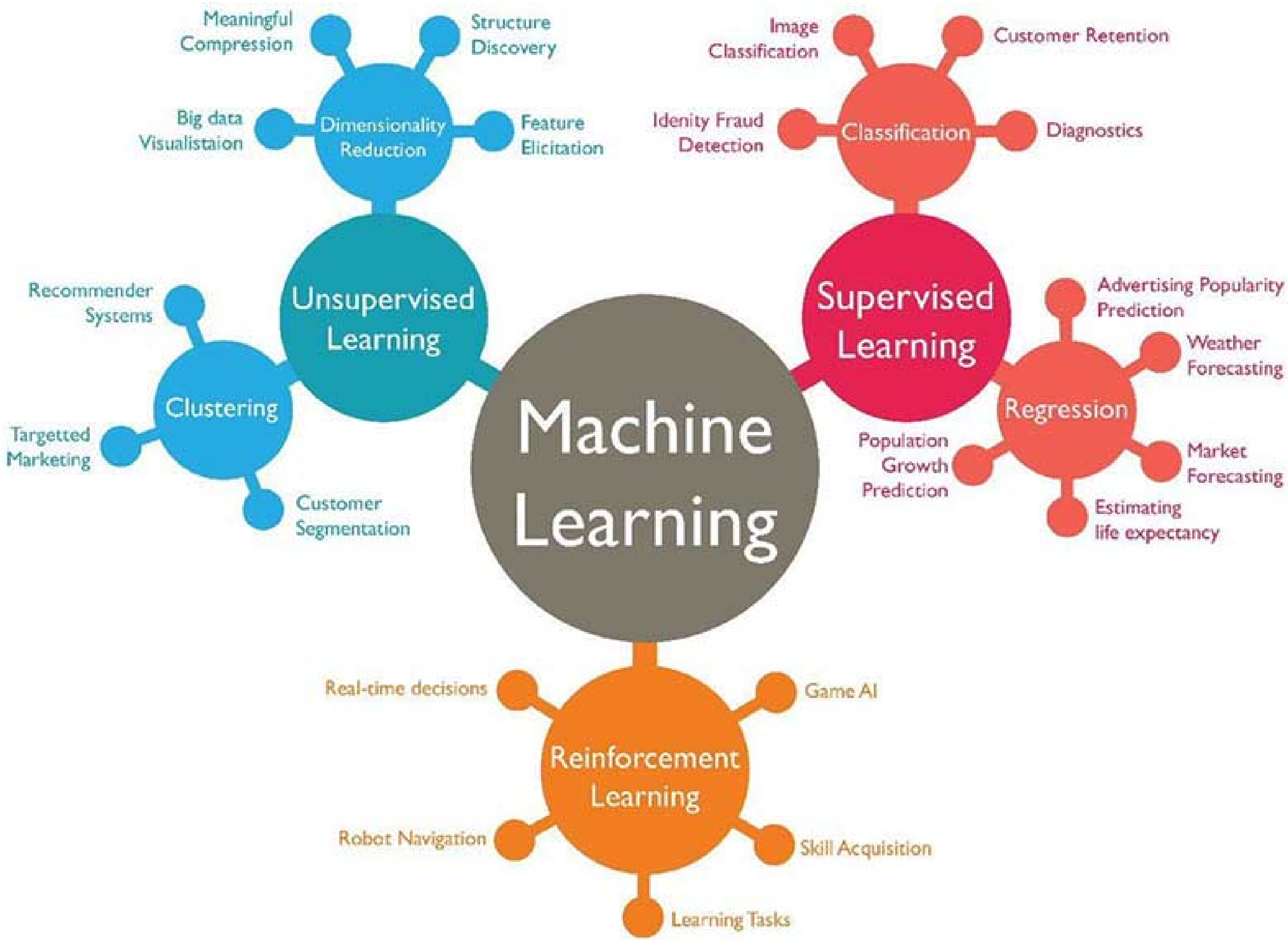
https://www.youtube.com/watch?v=rbsqaJwpu6A

ALPHAGO

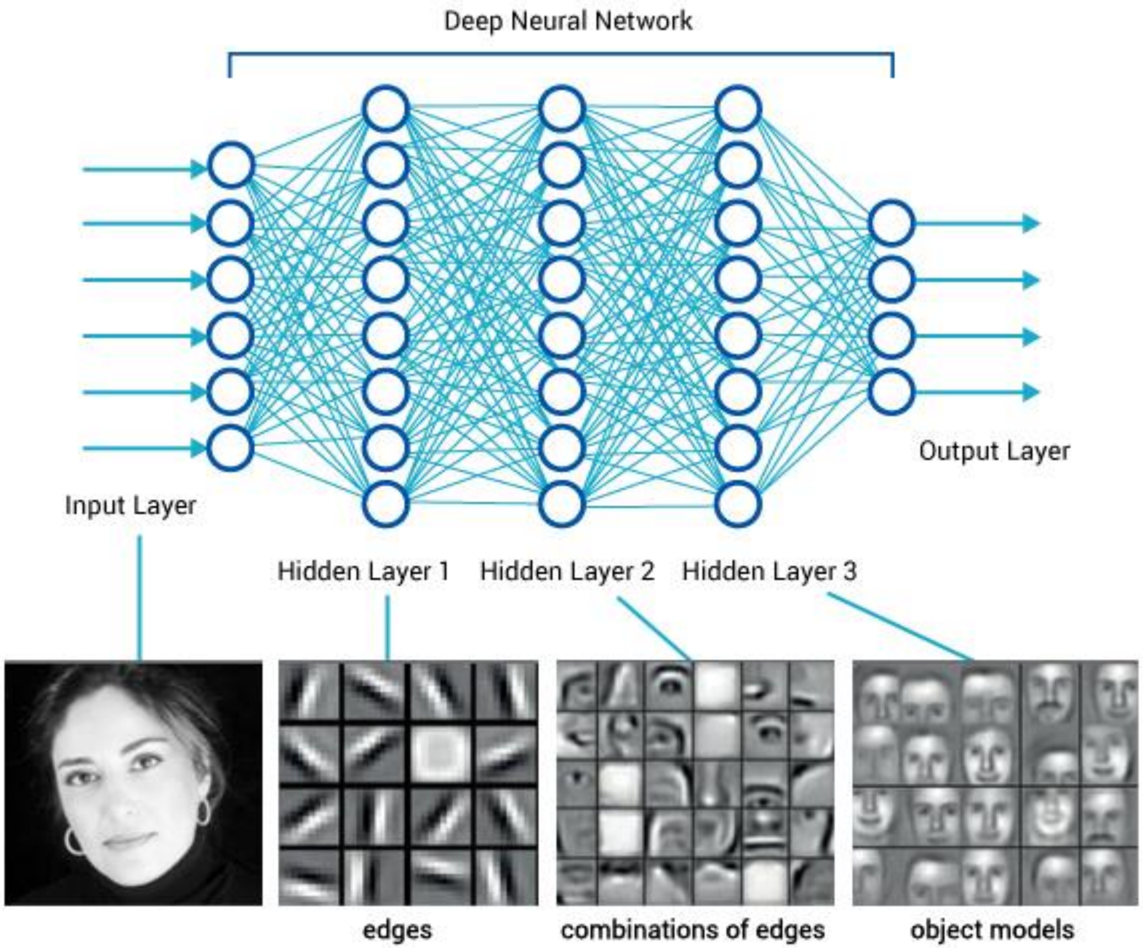


https://www.youtube.com/watch?v=SUbqykXVx0A

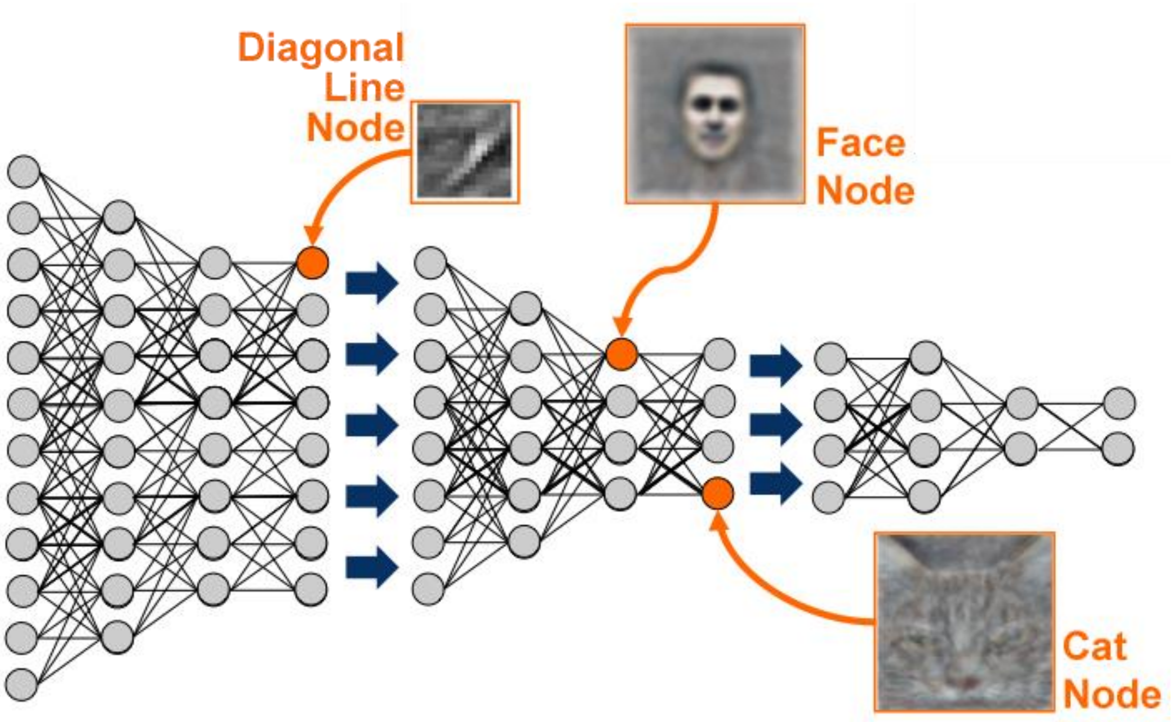
Machine Learning

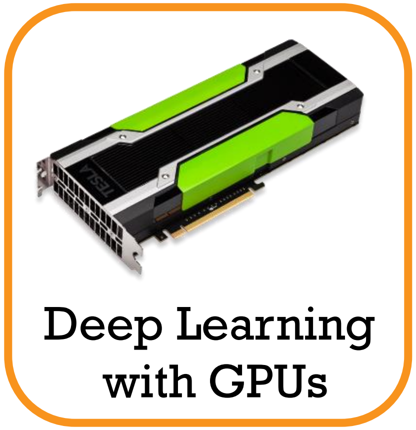


DEEP LEARNING



GOOGLE CAT VIDEOS





5 min break

Linear algebra review

Let u and v be vectors in $\mathbb{R}^d$. The inner-product is defined as

$$\langle u, v\rangle = \left\langle \begin{bmatrix} u_1 \\ \vdots \\ u_d \end{bmatrix}, \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} \right\rangle = \sum_{i=1}^{d} u_i v_d = \left[u_1, \cdots, u_d\right] \begin{bmatrix} v_1 \\ \vdots \\ v_d \end{bmatrix} = u^T v$$

Recall the properties of the inner product:

(i) $\langle u, v\rangle = \langle v, u\rangle$ (symmetry)

(ii) $\langle au_1 + bu_2, v\rangle = a\langle u_1, v\rangle + b\langle u_2, v\rangle$ (linearity)

Now let $A = \{a_{ij}\}$ be a d x d matrix. We have

$$\begin{aligned}
\langle u, Av \rangle &= \sum_{i=1}^{d} u_i (Av)_i \\
&= \sum_{i=1}^{d} u_i \sum_{j=1}^{d} a_{ij} v_j \\
&= \sum_{i=1}^{d} a_{ij} u_i \sum_{j=1}^{d} v_j \\
&= \left\langle A^T u, v \right\rangle,
\end{aligned}$$

where A^T is the transpose, or adjoint, of matrix A.

If $A^T = A^{-1}$, then A is an orthogonal matrix, i.e. $A^T = A^{-1}$ if and only if its columns are orthonormal vectors.

Example

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^T = \begin{bmatrix} 0 & 1 \\ -1 & 0 \end{bmatrix} = \begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}^{-1}$$

If $A^T = A$, then we say that A is self-adjoint, or symmetric.

Example

Hessian matrix of a (2-variable) function $f(x, y)$:

$$\nabla^2 f = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix}$$

Covariance matrices

Definition

The covariance matrix of a random vector $[X_1, \ldots, X_d]^T$ is a d x d matrix whose $(i,j)^{th}$ entry is the covariance $cov(X_i, X_j) = \mathbb{E}\left[(X_i - \mathbb{E}[X_i])(X_j - \mathbb{E}[X_j])\right]$.

Example

Covariance matrix of a mean-zero random vector $[X, Y]^T$; i.e. $\mathbb{E}[X] = 0 = \mathbb{E}[Y]$:

$$\begin{bmatrix} \mathbb{E}\left[X^2\right] & \mathbb{E}[XY] \\ \mathbb{E}[YX] & \mathbb{E}\left[Y^2\right] \end{bmatrix}$$

Gaussian probability density function

One dimension:

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

Multi-variate case ($x \in \mathbb{R}^d$):

$$p(x) = \frac{1}{\sqrt{\det A}\left(\sqrt{2\pi}\right)^d} e^{-\frac{1}{2}\left\langle x-\mu, A^{-1}(x-\mu)\right\rangle},$$

where A is the covariance matrix.

Theorem

Let A be a self-adjoint matrix. Then there exists an orthonormal eigenbasis $(v_1, \ldots, v_d)$ with eigenvalues $\lambda_1, \ldots, \lambda_d$ such that

$$A = \begin{bmatrix} | & \cdots & | \\ v_1 & \ddots & v_d \\ | & \cdots & | \end{bmatrix} \begin{bmatrix} \lambda_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \lambda_d \end{bmatrix} \begin{bmatrix} | & \cdots & | \\ v_1 & \ddots & v_d \\ | & \cdots & | \end{bmatrix}^{-1}.$$

First-order approximation

$$f(x+h) \approx f(x) + \langle \nabla f(x), h \rangle$$

Example

In two dimensions,

$$f(x_1 + h_1, x_2 + h_2) \approx f(x_1, x_2) + \left\langle \begin{bmatrix} \frac{\partial f}{\partial e_1}(x_1, x_2) \\ \frac{\partial f}{\partial e_2}(x_1, x_2) \end{bmatrix}, \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \right\rangle$$

Gradient ascent

What direction should h take to maximally increase f?

Theorem (Cauchy-Schwarz inequality)

$|\langle u, v\rangle| \leq \|u\| \|v\|$, *and both sides are equal if and only if* $v = \lambda u$ *for some* $\lambda \in \mathbb{R}$.

Note that if $v = \lambda u$, we have

$$\begin{aligned}\langle u, v\rangle &= \langle u, \lambda u\rangle \\ &= \lambda \langle u, u\rangle = \lambda \|u\|^2,\end{aligned}$$

so we get maximum increase if $h = \lambda \nabla f$ and $\lambda > 0$ (and similarly maximum decrease if $\lambda < 0$).

Second-order approximation

$$f(x+h) \approx f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \left\langle h, \nabla^2 f(x) h \right\rangle$$

Example

In two dimensions,

$$f(x_1 + h_1, x_2 + h_2) \approx f(x_1, x_2) + \left\langle \begin{bmatrix} \frac{\partial f}{\partial e_1}(x_1, x_2) \\ \frac{\partial f}{\partial e_2}(x_1, x_2) \end{bmatrix}, \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \right\rangle$$
$$+ \frac{1}{2} \left\langle \begin{bmatrix} h_1 \\ h_2 \end{bmatrix}, \begin{bmatrix} \frac{\partial^2 f}{\partial e_1^2}(x_1, x_2) & \frac{\partial^2 f}{\partial e_1 \partial e_2}(x_1, x_2) \\ \frac{\partial^2 f}{\partial e_2 \partial e_1}(x_1, x_2) & \frac{\partial^2 f}{\partial e_2^2}(x_1, x_2) \end{bmatrix} \begin{bmatrix} h_1 \\ h_2 \end{bmatrix} \right\rangle$$

Newton's method

Let $\tilde{f}(x) := f(x) + \langle \nabla f(x), h \rangle + \frac{1}{2} \left\langle h, \nabla^2 f(x) h \right\rangle$. To find the best h to maximize (minimize) $\tilde{f}$, we take the gradient with respect to h and set to 0:

$$\nabla_h \tilde{f}(x) = \nabla f(x) + \nabla^2 f(x) h = 0,$$

which implies that

$$h = \left(\nabla^2 f(x)\right)^{-1} (-\nabla f(x)) \, .$$

- In summary, we optimize at each time step by

$$x_{n+1} = x_n + \alpha h,$$

where $h = \pm \nabla f(x_n)$ in the case of gradient descent (ascent), and $h = \left(\nabla^2 f(x)\right)^{-1} (-\nabla f(x))$ in Newton's method.

- Newton's method converges faster with fewer iterations, but each iteration is computationally more expensive as one has to invert the Hessian matrix.

Gentle introduction to tensors

We encountered the quadratic form

$$\langle x, Ax \rangle$$

several times already; now we will discuss its significance in terms of tensors.

Definition

A k-tensor $T : \underbrace{\mathbb{R}^d \times \cdots \times \mathbb{R}^d}_{k \text{ times}} \to \mathbb{R}$ is a scalar function which is linear in each component.

Example

Given a vector u, the function

$$x \mapsto \langle u, x \rangle$$

is a 1-tensor.

Example

Given a matrix A, the function

$$(x, y) \mapsto \langle x, Ay \rangle$$

is a 2-tensor.

Example

The determinant of a $d \times d$ matrix, as a function of the d columns (or rows), is a d-tensor.

Theorem

All 1-tensors $T(x)$ can be written as

$$\langle u_T, x \rangle$$

for some vector u_T, and all 2-tensors $S(x, y)$ can be expressed as

$$\langle x, A_S y \rangle$$

for some matrix A_S.