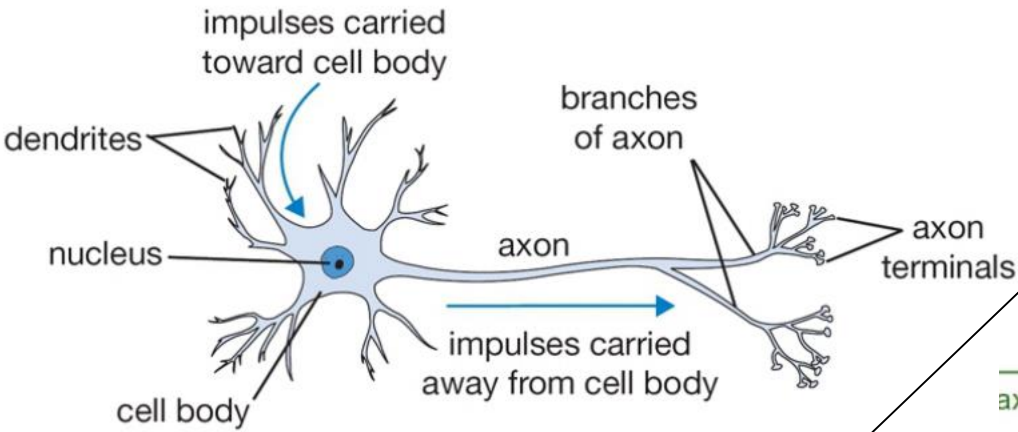
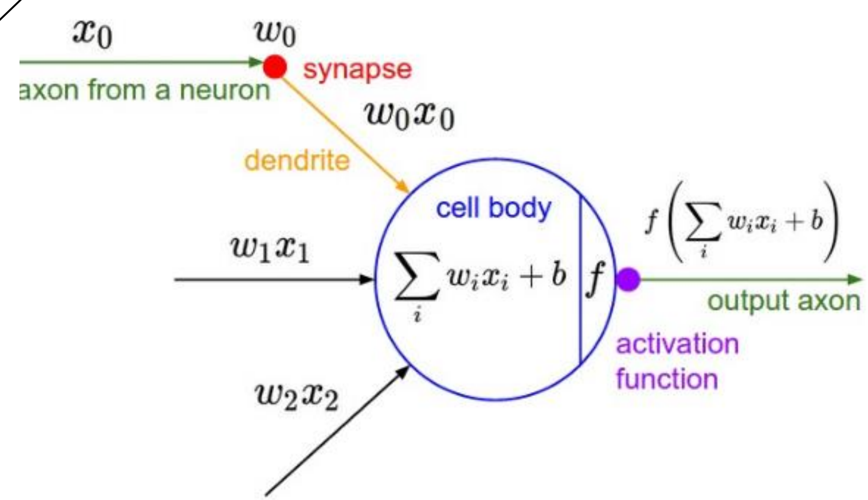
COGS 181, Fall 2017

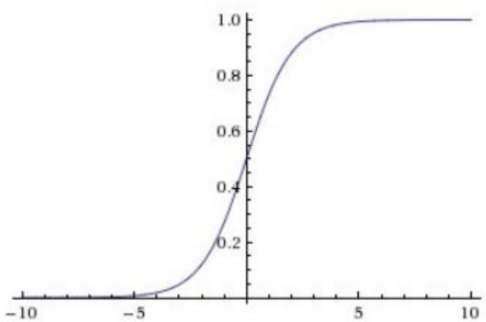
Neural Networks and Deep Learning

Lecture 2: Vector Calculus

Perceptron







Sigmoid function

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

Fei-fei and Karpathy

Mathematical representation for features

$$S = \{(\mathbf{x}_i), i = 1..n\} \qquad \mathbf{x}_i = (x_{i1}, ..., x_{im})$$

age	male or female	weight (lb)	height (cm)
$x_{11} = 22$	$x_{12} = M$	$x_{13} = 160$	$x_{14} = 180$
$x_{21} = 51$	$x_{22} = M$	$x_{23} = 190$	$x_{24} = 175$
$x_{31} = 43$	$x_{32} = F$	$x_{33} = 120$	$x_{34} = 165$

Gender variable: $x_{i2} \in \{Male, Female\}$?

$x_{i2} = 0$, if Male

$x_{i2} = 1$, if Female

Mathematical representation for features

$$S = \{(\mathbf{x}_i), i = 1..n\} \qquad \mathbf{x}_i = (x_{i1}, ..., x_{im})$$

What if it is a city: $x_{i2} \in \{LosAngeles, SanDiego, Irvine\}$?

We ue a coding strategy by expanding the features.
For N number of possible states, we expand the features into N-dimenstional.

One-hot encoding:

	coded values
Los Angeles	1, 0, 0
San Diego	0, 1, 0
Irvine	0, 0, 1

Pros: we can naturally deal with any type of input (can associate confidence directly).

Cons: the feature dimension has become much larger.

Input matrix

$$S = \{\mathbf{x}_i, i = 1..n\} \qquad \mathbf{x}_i = (x_{i1}, ..., x_{im})$$

age	male or female	weight (lb)	height (cm)
$x_{11} = 22$	$x_{12} = M$	$x_{13} = 160$	$x_{14} = 180$
$x_{21} = 51$	$x_{22} = M$	$x_{23} = 190$	$x_{24} = 175$
$x_{31} = 43$	$x_{32} = F$	$x_{33} = 120$	$x_{34} = 165$

If we write each sample as a row vector:

$$\mathbf{x}_1 = (22, 1, 0, 160, 180)$$

$$X = \begin{pmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \\ \mathbf{x}_3 \end{pmatrix} \qquad X \in R^{n \times m}$$

$$X = \begin{pmatrix} 22 & 1 & 0 & 160 & 180 \\ 51 & 1 & 0 & 190 & 175 \\ 43 & 0 & 1 & 120 & 165 \end{pmatrix}$$

Input matrix

$$S = \{\mathbf{x}_i, i = 1..n\} \qquad \mathbf{x}_i = (x_{i1}, ..., x_{im})^T$$

age	male or female	weight (lb)	height (cm)
$x_{11} = 22$	$x_{12} = M$	$x_{13} = 160$	$x_{14} = 180$
$x_{21} = 51$	$x_{22} = M$	$x_{23} = 190$	$x_{24} = 175$
$x_{31} = 43$	$x_{32} = F$	$x_{33} = 120$	$x_{34} = 165$

More often we write each sample as a COLUMN vector:

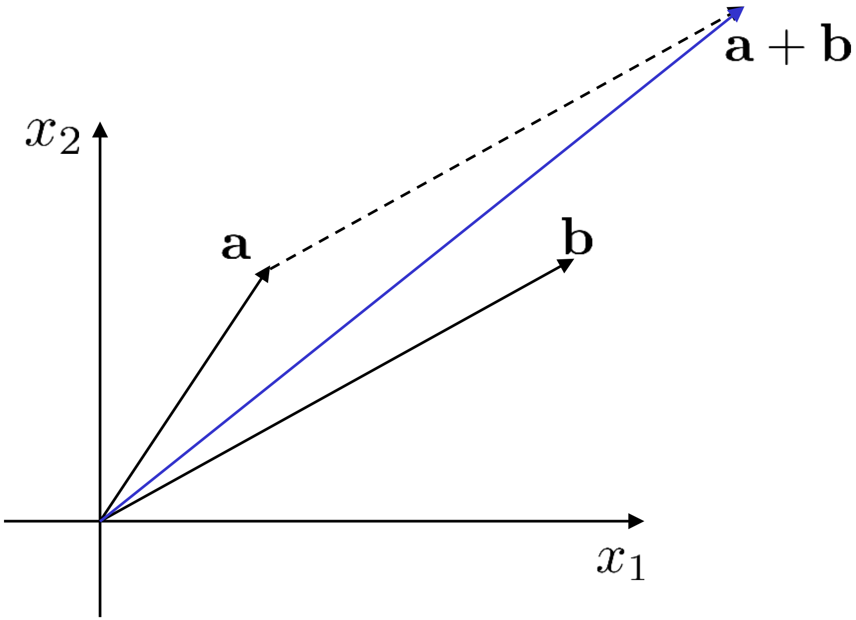
$$\mathbf{x}_1 = \begin{pmatrix} 22 \\ 1 \\ 0 \\ 160 \\ 180 \end{pmatrix}$$

$$X = (\mathbf{x}_1, \mathbf{x}_2, \mathbf{x}_3) \qquad X \in R^{m \times n}$$

$$X = \begin{pmatrix} 22 & 51 & 43 \\ 1 & 1 & 0 \\ 0 & 0 & 1 \\ 160 & 190 & 120 \\ 180 & 175 & 165 \end{pmatrix}$$

Vector

Addition:



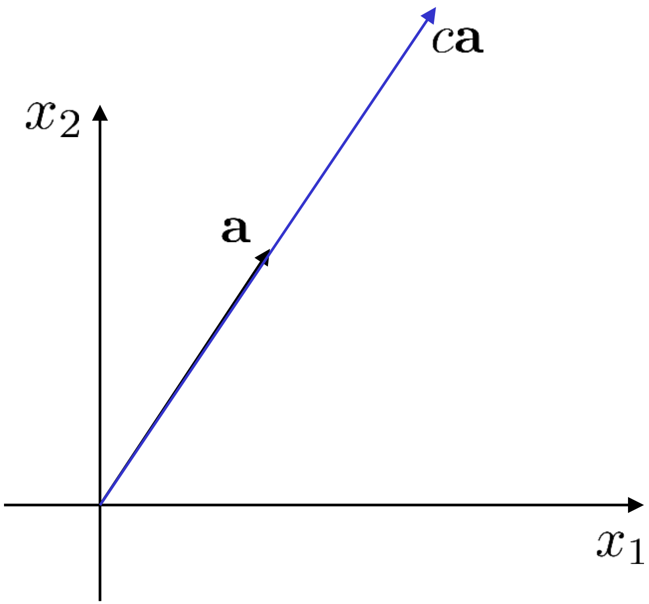
$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

$$\mathbf{a}+\mathbf{b} = \begin{pmatrix} a_1 + b_1 \\ a_2 + b_2 \\ a_3 + b_3 \end{pmatrix}$$

It's still a vector in the same space as a and b.

Vector

Scaling:



$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix}$$

$$c \in R$$

$$c\mathbf{a} = \begin{pmatrix} c \times a_1 \\ c \times a_2 \\ c \times a_3 \end{pmatrix}$$

It's still a vector in the same space as a.

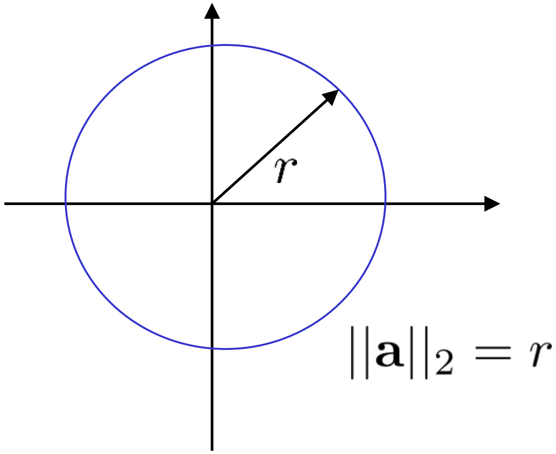
Norm

$$\mathbf{a} = (a_1, a_2, ..., a_n),\ a_i \in R$$

L2 Norm:

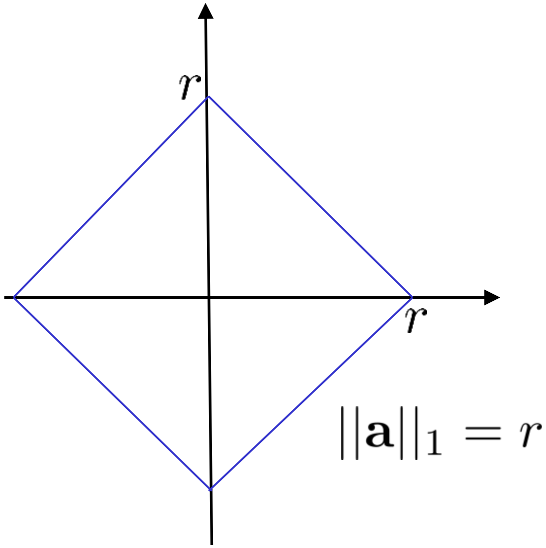
$$||\mathbf{a}||_2 = \sqrt{\sum_{i=1}^{n} a_i^2}$$

$$||\mathbf{a}||^2 = \sum_{i=1}^{n} a_i^2$$

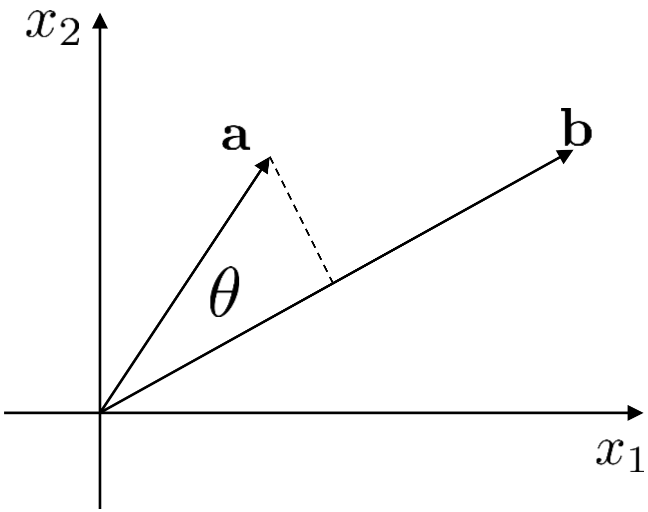


L1 Norm:

$$||\mathbf{a}||_1 = \sum_{i=1}^{n} |a_i|$$



Vector: Projection (inner product)
(one of the most important concepts in machine learning)



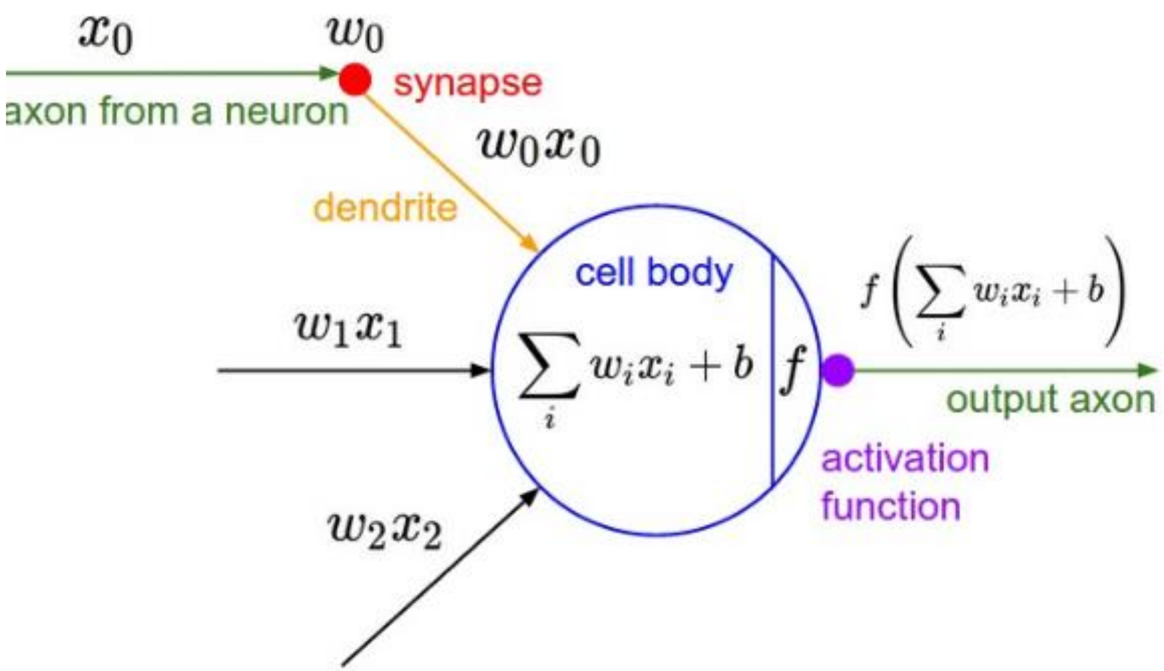
$$\mathbf{a} = \begin{pmatrix} a_1 \\ a_2 \\ a_3 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

$$< \mathbf{a}, \mathbf{b} > \ \equiv \mathbf{a} \cdot \mathbf{b} \ \equiv \mathbf{a}^T \mathbf{b} \ \equiv a_1 b_1 + a_2 b_2 + a_3 b_3$$

It's a scalar!

$$cos(\theta) = \frac{<\mathbf{a},\mathbf{b}>}{||\mathbf{a}||_2 \times ||\mathbf{b}||_2}$$

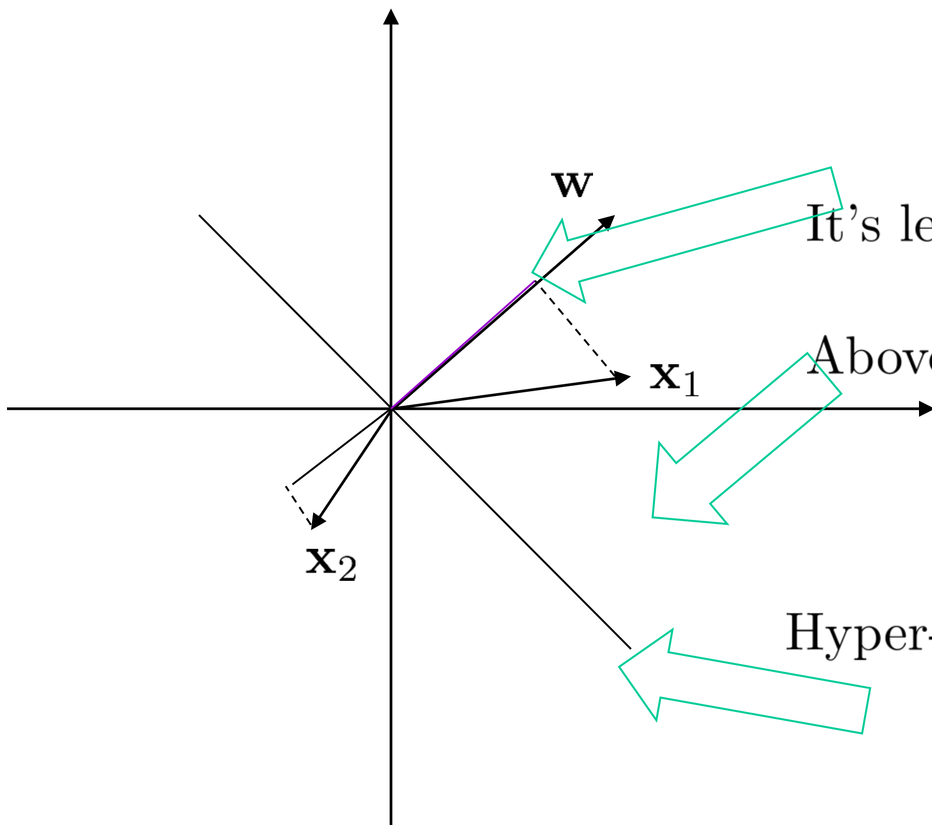
Perceptron



x_0
w_0
synapse
axon from a neuron
$w_0 x_0$
dendrite
cell body
$f\left(\sum_i w_i x_i + b\right)$
$w_1 x_1$
$\sum_i w_i x_i + b$
f
output axon
activation
function
$w_2 x_2$

Orthogonal

$||\mathbf{w}||_2 = 1$: unit vector



It's length: $| < \mathbf{w}, \mathbf{x}_1 > |$

Above or below: $sign(< \mathbf{w}, \mathbf{x}_1 >)$

Hyper-plane: $\{\mathbf{x} :< \mathbf{w}, \mathbf{x} >= 0\}$

Matrix multiplication

Vector:

$$\mathbf{a} = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \qquad \mathbf{b} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}$$

$$\mathbf{ab} = \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} = a_1 b_1 + a_2 b_2 + a_3 b_3$$

$$\mathbf{ab} \neq \mathbf{ba}$$

$$\mathbf{ba} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix} \begin{pmatrix} a_1 & a_2 & a_3 \end{pmatrix} = \begin{pmatrix} b_1 a_1 & b_1 a_2 & b_1 a_3 \\ b_2 a_1 & b_2 a_2 & b_2 a_3 \\ b_3 a_1 & b_3 a_2 & b_3 a_3 \end{pmatrix}$$

Matrix multiplication

Matrix:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \qquad B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$$

$$AB = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix} \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}$$

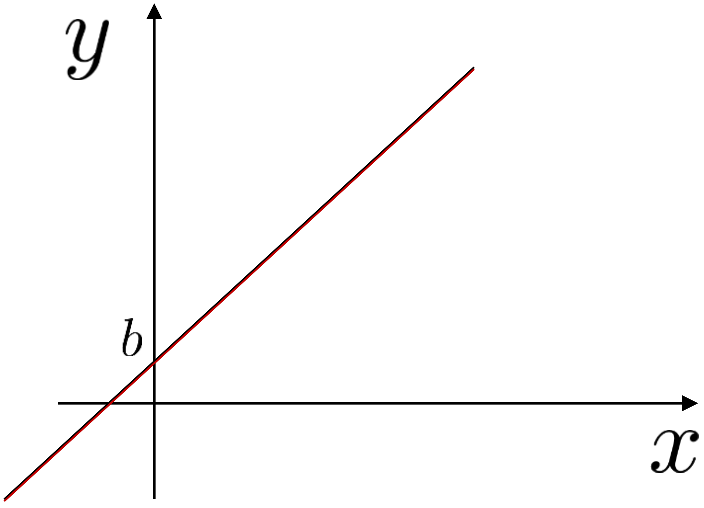
$$= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}$$

Calculus

Scalar:

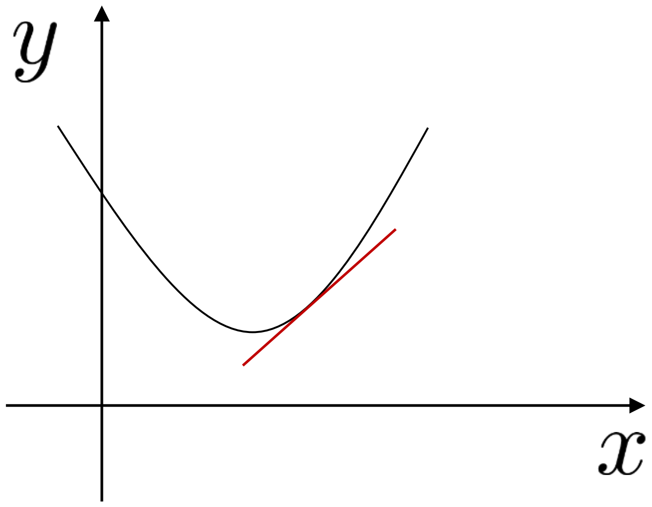
$$y = ax + b$$

$$\frac{dy}{dx} = a$$



$$y = ax^2 + bx + c$$

$$\frac{dy}{dx} = 2ax + b$$



Calculus

Vector:

Vector-by-scalar

$$Y(x) = \begin{pmatrix} y_1(x) & y_2(x) & y_3(x) \end{pmatrix}$$

$$\frac{dY(x)}{dx} = \begin{pmatrix} \frac{dy_1(x)}{dx} & \frac{dy_2(x)}{dx} & \frac{dy_3(x)}{dx} \end{pmatrix}$$

Vector-by-vector

$$Y(X) = \begin{pmatrix} y_1(X) & ,..., & y_m(X) \end{pmatrix} \qquad X = \begin{pmatrix} x_1 & ,..., & x_n \end{pmatrix}$$

$$\frac{dY(X)}{dX} = \begin{pmatrix} \frac{dy_1(X)}{\partial x_1} & ,..., & \frac{dy_m(X)}{\partial x_1} \\ . & . & . \\ \frac{dy_1(X)}{\partial x_n} & ,..., & \frac{dy_m(X)}{\partial x_n} \end{pmatrix}$$

Calculus

Matrix:

Matrix-by-scalar

$$Y(x) = \begin{pmatrix} y_{11}(x) & ,\ldots, & y_{1m}(x) \\ . & . & . \\ y_{n1}(x) & ,\ldots, & y_{nm}(x) \end{pmatrix}$$

$$\frac{dY(x)}{dx} = \begin{pmatrix} \frac{dy_{11}(x)}{dx} & ,\ldots., & \frac{dy_{1m}(x)}{dx} \\ \frac{dy_{n1}(x)}{dx} & ,\ldots., & \frac{dy_{nm}(x)}{dx} \end{pmatrix}$$

Basics abut data and linear algebra operations

$$S = \{(\mathbf{x}_i, y_i), i = 1..n\} \qquad y_i \in \{-1, +1\}$$

	age	male or female	weight (lb)	height (cm)
$y_1 = -1$ (negative)	$x_{11} = 22$	$x_{12} = M$	$x_{13} = 160$	$x_{14} = 180$
$y_2 = +1$ (positive)	$x_{21} = 51$	$x_{22} = M$	$x_{23} = 190$	$x_{24} = 175$
$y_3 = +1$ (positive)	$x_{31} = 43$	$x_{32} = F$	$x_{33} = 120$	$x_{34} = 165$

$$\Downarrow$$

$$X = \begin{pmatrix} 22 & 1 & 0 & 160 & 180 \\ 51 & 1 & 0 & 190 & 175 \\ 43 & 0 & 1 & 120 & 165 \end{pmatrix} \qquad Y = \begin{pmatrix} -1 \\ +1 \\ +1 \end{pmatrix}$$

$$W = \begin{pmatrix} 0.075 \\ 0 \\ 0 \\ -0.007 \\ -0.008 \end{pmatrix} \qquad \hat{Y} = XW = \begin{pmatrix} -0.91 \\ 1.095 \\ 1.065 \end{pmatrix}$$

Calculus

vector-by-vector

$$A = \begin{pmatrix} a_{11} & ,\dots, & a_{1m} \\ . & . & . \\ a_{n1} & ,\dots, & a_{nm} \end{pmatrix} \qquad X = \begin{pmatrix} x_1 \\ . \\ x_m \end{pmatrix}$$

$$\frac{\partial AX}{\partial X} = A^T\text{: denominator layout}$$

$$\frac{\partial X^T A^T}{\partial X} = A^T\text{: denominator layout}$$

Vector calculus

Identities: vector-by-vector $\frac{\partial \mathbf{y}}{\partial \mathbf{x}}$

Condition	Expression	Numerator layout, i.e. by $\mathbf{y}$ and $\mathbf{x}^T$	Denominator layout, i.e. by $\mathbf{y}^T$ and $\mathbf{x}$
$\mathbf{a}$ is not a function of $\mathbf{x}$	$\frac{\partial \mathbf{a}}{\partial \mathbf{x}} =$	$\mathbf{0}$	
	$\frac{\partial \mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{I}$	
$\mathbf{A}$ is not a function of $\mathbf{x}$	$\frac{\partial \mathbf{A}\mathbf{x}}{\partial \mathbf{x}} =$	$\mathbf{A}$	$\mathbf{A}^\top$
$\mathbf{A}$ is not a function of $\mathbf{x}$	$\frac{\partial \mathbf{x}^\top \mathbf{A}}{\partial \mathbf{x}} =$	$\mathbf{A}^\top$	$\mathbf{A}$
a is not a function of $\mathbf{x}$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	
$a = a(\mathbf{x})$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial a\mathbf{u}}{\partial \mathbf{x}} =$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \mathbf{u}\frac{\partial a}{\partial \mathbf{x}}$	$a\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial a}{\partial \mathbf{x}}\mathbf{u}^\top$
$\mathbf{A}$ is not a function of $\mathbf{x}$, $\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial \mathbf{A}\mathbf{u}}{\partial \mathbf{x}} =$	$\mathbf{A}\frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}}\mathbf{A}^\top$
$\mathbf{u} = \mathbf{u}(\mathbf{x})$, $\mathbf{v} = \mathbf{v}(\mathbf{x})$	$\frac{\partial (\mathbf{u}+\mathbf{v})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} + \frac{\partial \mathbf{v}}{\partial \mathbf{x}}$	
$\mathbf{u} = \mathbf{u}(\mathbf{x})$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{x}} =$	$\frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}} \frac{\partial \mathbf{u}}{\partial \mathbf{x}}$	$\frac{\partial \mathbf{u}}{\partial \mathbf{x}} \frac{\partial \mathbf{g}(\mathbf{u})}{\partial \mathbf{u}}$

https://en.wikipedia.org/wiki/Matrix_calculus

Calculus

Scalar-by-vector

A is asymmetric

$$A = \begin{pmatrix} a_{11} & ,\ldots, & a_{1m} \\ . & . & . \\ a_{m1} & ,\ldots, & a_{mm} \end{pmatrix} \qquad X = \begin{pmatrix} x_1 \\ . \\ x_m \end{pmatrix}$$

$$\frac{\partial X^T A X}{\partial X} = (A + A^T)X\text{: denominator layout}$$

A is symmetric

$$A = \begin{pmatrix} a_{11} & ,\ldots, & a_{1m} \\ . & . & . \\ a_{1m} & ,\ldots, & a_{mm} \end{pmatrix} \qquad X = \begin{pmatrix} x_1 \\ . \\ x_m \end{pmatrix}$$

$$\frac{\partial X^T A X}{\partial X} = 2AX\text{: denominator layout}$$

Matrix calculus

a is not a function of **x**	$\frac{\partial(\mathbf{a}\cdot\mathbf{x})}{\partial\mathbf{x}} = \frac{\partial(\mathbf{x}\cdot\mathbf{a})}{\partial\mathbf{x}} =$ $\frac{\partial\mathbf{a}^\top\mathbf{x}}{\partial\mathbf{x}} = \frac{\partial\mathbf{x}^\top\mathbf{a}}{\partial\mathbf{x}} =$	$\mathbf{a}^\top$	$\mathbf{a}$
A is not a function of **x** **b** is not a function of **x**	$\frac{\partial\mathbf{b}^\top\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} =$	$\mathbf{b}^\top\mathbf{A}$	$\mathbf{A}^\top\mathbf{b}$
A is not a function of **x**	$\frac{\partial\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} =$	$\mathbf{x}^\top(\mathbf{A}+\mathbf{A}^\top)$	$(\mathbf{A}+\mathbf{A}^\top)\mathbf{x}$
A is not a function of **x** **A** is symmetric	$\frac{\partial\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial\mathbf{x}} =$	$2\mathbf{x}^\top\mathbf{A}$	$2\mathbf{A}\mathbf{x}$
A is not a function of **x**	$\frac{\partial^2\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial\mathbf{x}^2} =$	$\mathbf{A}+\mathbf{A}^\top$	
A is not a function of **x** **A** is symmetric	$\frac{\partial^2\mathbf{x}^\top\mathbf{A}\mathbf{x}}{\partial\mathbf{x}^2} =$	$2\mathbf{A}$	

https://en.wikipedia.org/wiki/Matrix_calculus

Three key things you are learning in this class:

Representation: With better and better understanding of the underlining statistics about the data and methods.

Evaluation: The ideal strategy is always to aim at your target directly (take non-stop flight as opposed to having multiple stops).

Optimization: Based on the chosen representation and evaluation, you pick a strategy (mathematical/statistical) to achieve your goal.

Understanding the difference between training and testing

Regardless the situation of supervised, unsupervised (or even semi-supervised, weakly-supervised, reinforcement, …), we often define a loss (or error) function:

$$S_{training} = \{\mathbf{x}_i, i = 1..n\}$$

$$loss_{training} = \sum_{i=1}^{n} weight_i \cdot l(\mathbf{x}_i)$$

$weight_i$ and $l(\mathbf{x}_i)$ are weight and loss for each sample i

$$S_{testing} = \{\mathbf{x}_i, i = 1..u\}$$

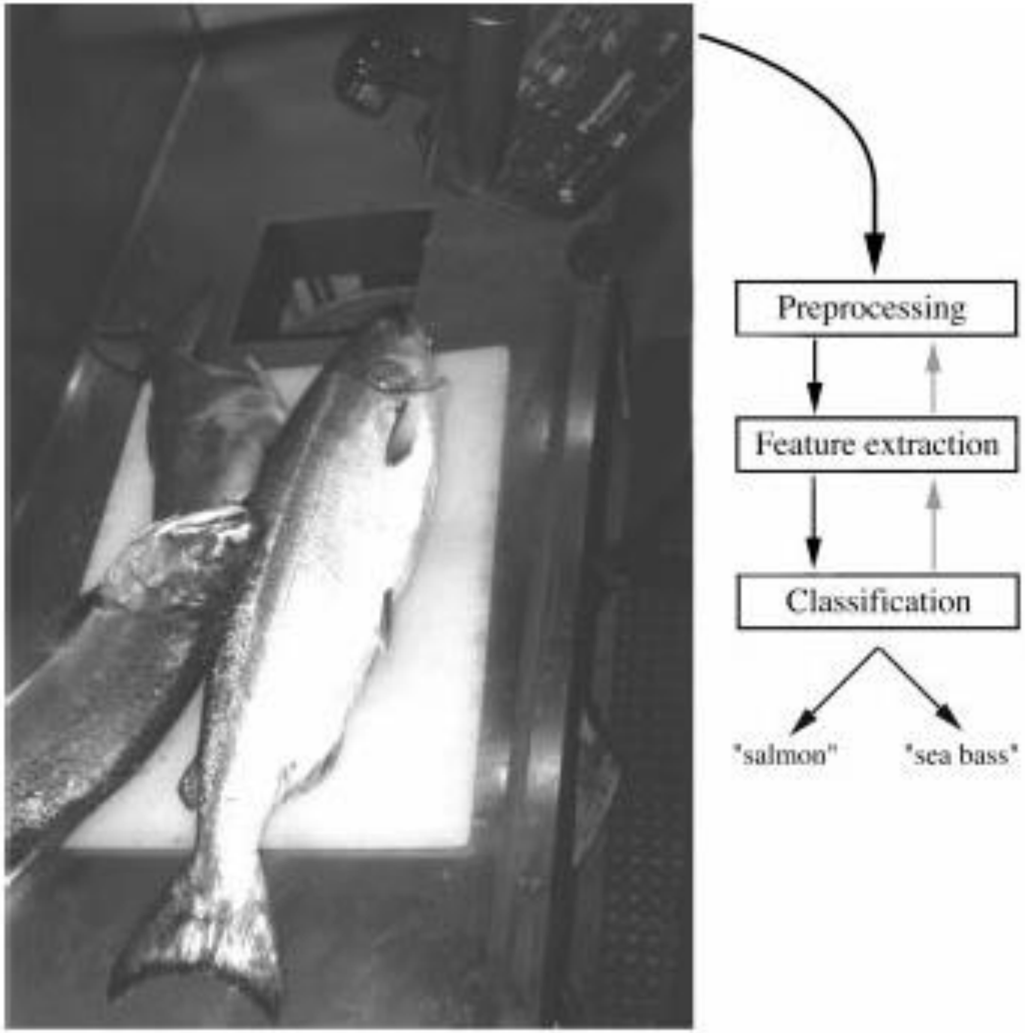
$$loss_{testing} = \sum_{i=1}^{u} weight_i \cdot l(\mathbf{x}_i)$$

$weight_i$ and $l(\mathbf{x}_i)$ are weight and loss for each sample i

$$loss_{testing} \neq loss_{training}$$

$$loss_{testing} \rightarrow loss_{training}$$

An example


Preprocessing
Feature extraction
Classification
"salmon"
"sea bass"

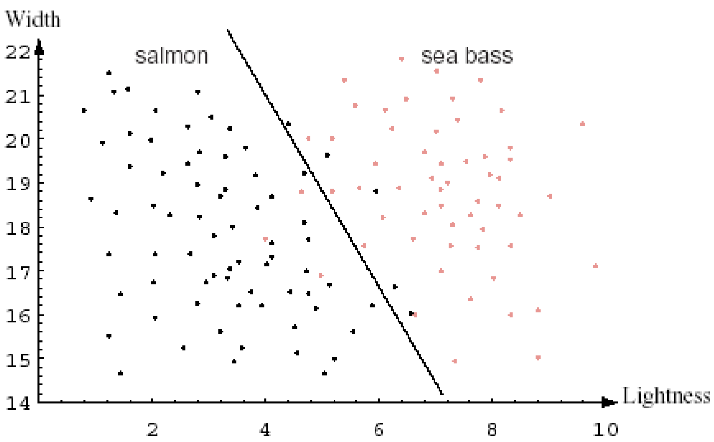
Summary of the problem

Let **x** be the input vector (observation) and y be its label:

Often, we are given a set of training data

$$S_{training} = \{(\mathbf{x}_i, y_i), i = 1..n\} \qquad \mathbf{x} = (x_1, ..., x_m), x_i \in \mathcal{R}, \quad \mathbf{x} \in \mathcal{R}^m$$

We use the training set to train a classifier f(**x**).



Given a set of testing data, we make the prediction of each input and evaluate the algorithm.

$$S_{testing} = \{(\mathbf{x}_i, y_i), i = 1..q\}.$$

For each $\mathbf{x}_i$ we want to predict its y_i.

y_i is given to evaluate the quality of a classifier and is not given in reality.