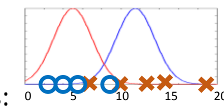
Logistic Classifier

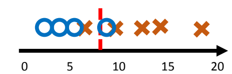
CISC 5800

Professor Daniel Leeds

Classification strategy: generative vs. discriminative



- Generative, e.g., Bayes/Naïve Bayes:
  - Identify probability distribution for each class
  - Determine class with maximum probability for data example
- Discriminative, e.g., Logistic Regression:
  - Identify boundary between classes
  - Determine which side of boundary new data example exists on



Linear algebra: data features

- Vector – list of numbers: each number describes a data **feature**
- Matrix – list of lists of numbers: features for each data point

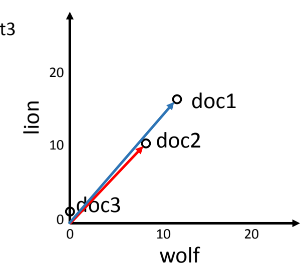
	Document 1	Document 2	Document 3
Wolf	12	8	0
Lion	16	10	2
Monkey	14	11	1
Broker	0	1	14
Analyst	1	0	10
Dividend	1	1	12
⋮	⋮	⋮	⋮

# of word occurrences

Feature space

- Each data feature defines a dimension in space

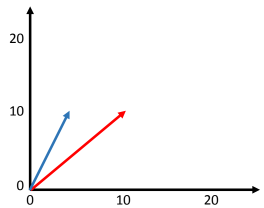
	Document1	Document2	Document3
Wolf	12	8	0
Lion	16	10	2
Monkey	14	11	1
Broker	0	1	14
Analyst	1	0	10
Dividend	1	1	12
⋮	⋮	⋮	⋮



The dot product

The dot product compares two vectors:

- $\boldsymbol{a} = \begin{bmatrix} a_1 \\ \vdots \\ a_n \end{bmatrix}$, $\boldsymbol{b} = \begin{bmatrix} b_1 \\ \vdots \\ b_n \end{bmatrix}$ $\quad\quad \boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i \; = \boldsymbol{a}^T \boldsymbol{b}$



$\begin{bmatrix} 5 \\ 10 \end{bmatrix} \cdot \begin{bmatrix} 10 \\ 10 \end{bmatrix} =$

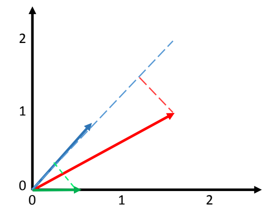
5

The dot product, continued $\boldsymbol{a} \cdot \boldsymbol{b} = \sum_{i=1}^{n} a_i b_i$

Magnitude of a vector is sum of squares of the elements

$|\boldsymbol{a}| = \sqrt{\sum_i a_i^2}$

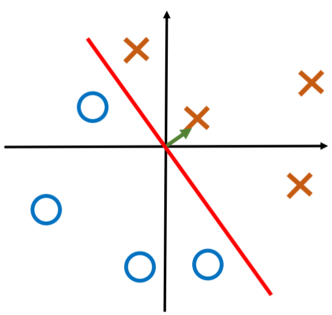
If $\boldsymbol{a}$ has unit magnitude, $\boldsymbol{a} \cdot \boldsymbol{b}$ is "projection" of $\boldsymbol{b}$ onto $\boldsymbol{a}$



$\begin{bmatrix} 0.6 \\ 0.8 \end{bmatrix} \cdot \begin{bmatrix} 1.5 \\ 1 \end{bmatrix} =$

7

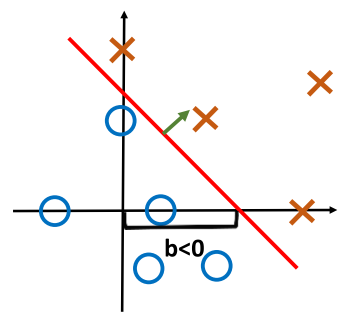
Separating boundary, defined by **w**



- Separating **hyperplane** splits **class 0** and **class 1**
- Plane is defined by line **w** perpendicular to plane
- Is data point **x** in class 0 or class 1? $\mathbf{w}^T\mathbf{x}$+b **>** 0 class **1**
  $\mathbf{w}^T\mathbf{x}$+b **<** 0 class **0**

10

Separating boundary, defined by **w** and b



Example:

$\mathbf{w} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$ b=-4

$\mathbf{x}^1 = \begin{bmatrix} 5 \\ 0 \end{bmatrix}$

$\mathbf{x}^2 = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$

11

Notational simplification

Recall: $\boldsymbol{w}^T\boldsymbol{x} = \boldsymbol{w}\cdot\boldsymbol{x} = \sum_{i=1}^{n} w_i x_i$

Define $x'_{1:n} = x_{1:n}$ and $x'_{n+1} = 1$ for all inputs $\boldsymbol{x}$ and $w'_{1:n} = w_{1:n}$ and $w'_{n+1} = b$

Now $\boldsymbol{w}'^T\boldsymbol{x}' = \boldsymbol{w}^T\boldsymbol{x} + b$

Let's assume x_{n+1}=1 always, and w_{n+1}=b always

From real-number projection to 0/1 label

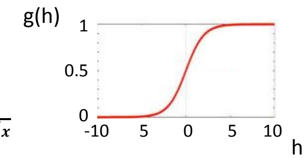
Binary classification: 0 is class A, 1 is class B

Sigmoid function stands in for p(**x**|y)

Sigmoid: $g(h) = \frac{1}{1+e^{-h}}$

$p(y=0|x;\theta) = 1 - g(\boldsymbol{w}^T\boldsymbol{x}) = \frac{e^{-\boldsymbol{w}^T\boldsymbol{x}}}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}}$

$p(y=1|x;\theta) = g(\boldsymbol{w}^T\boldsymbol{x}) = \frac{1}{1+e^{-\boldsymbol{w}^T\boldsymbol{x}}}$



$\boldsymbol{w}^T\boldsymbol{x} = \sum_j \boldsymbol{w}_j\boldsymbol{x}_j + b$

14

Learning parameters for classification

Similar to MLE for Bayes classifier

"Likelihood" for data points $y^1, \ldots, y^n$
(different from Bayesian likelihood)

- If y^i in class A, y^i =0, multiply (1-g($\boldsymbol{x}^i$;$\boldsymbol{w}$))
- If y^i in class B, y^i=1, multiply (g($\boldsymbol{x}^i$;$\boldsymbol{w}$))

$$\underset{w}{\operatorname{argmax}} L(y|x;w) = \prod_i \left(1 - g(\boldsymbol{x}^i;\boldsymbol{w})\right)^{(1-y^i)} g(\boldsymbol{x}^i;\boldsymbol{w})^{y^i}$$

15