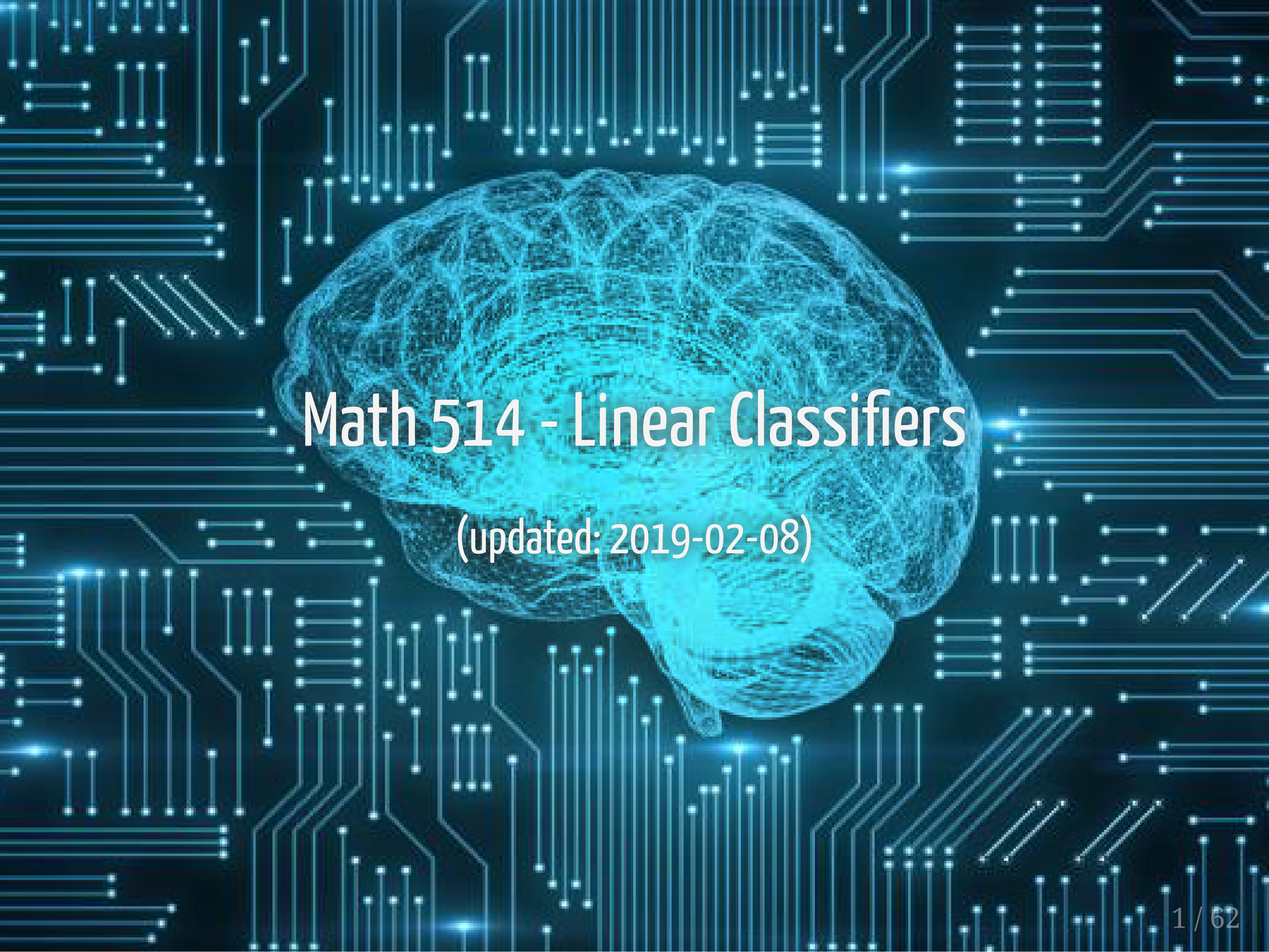


Math 514 - Linear Classifiers

(updated: 2019-02-08)

Class 4 - Gradient Free Learning

With this class begin the introduction of 1 new network/class

Perceptron network

- Perceptron learning algorithm(s)
  - Vanilla Perceptron
  - Averaged Perceptron
  - Voted Perceptron
- Perceptron convergence proof

Fisher Linear Discriminant

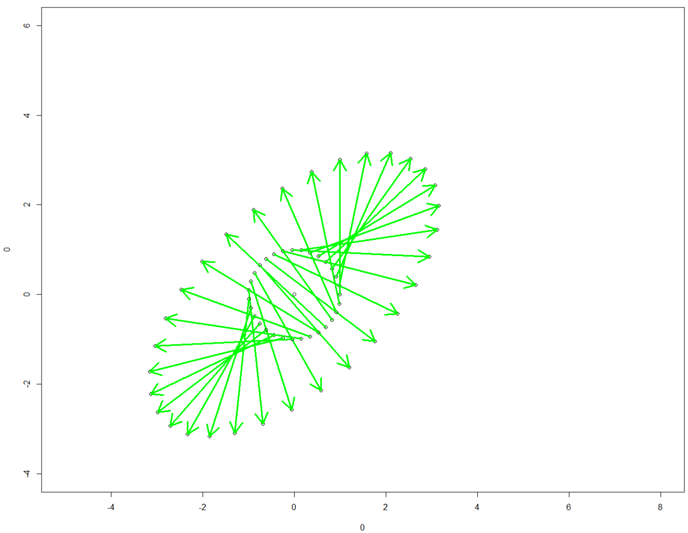
- Classical statistical technique heavily used by computer vision industry
- Projection method, like PCA, but supervised
- Good baseline for comparison
- If mapped to higher dimensions, can be very effective

Voted Perceptron

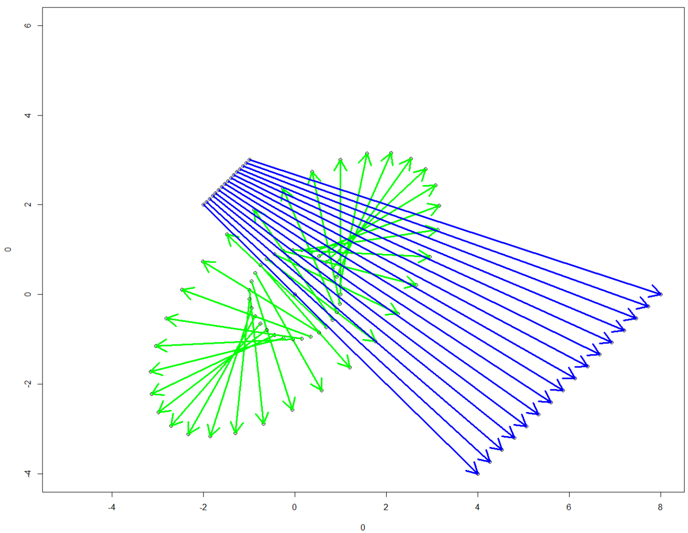
Large Margin Classification using the Perceptron Algorithm by Freund and Schapire (1999)

> The performance of our algorithm is close to, but not as good as, the performance of maximal-margin classifiers on the same problem, while saving significantly on computation time and programming effort.
>
> In this paper, we introduce a new and simpler algorithm for linear classification which takes advantage of data that are linearly separable with large margins. We named the new algorithm the voted-perceptron algorithm. The algorithm is based on the well known perceptron algorithm of Rosenblatt (1958, 1962)

Matrix as Map



Circle $\rightarrow$ Ellipse



Line $\rightarrow$ Line

Geometric View of Linear Algebra

A matrix $M \in \mathbb{R}^{m \times n}$ maps any line in $\mathbb{R}^n$ into a line in $\mathbb{R}^m$. In particular, this is true for lines through the origin (vectors).

Your two first questions might be:

1. For a given matrix M, are any lines $\alpha\mathbf{x}$ mapped to $\mathbf{0}$?
2. For a given matrix M, are any lines $\alpha\mathbf{x}$ mapped to themselves?

These two questions essentially divide basic Linear Algebra into two large pieces

1. Null spaces, invertibility
2. Eigenvalues, Eigenvectors
   - SVD, PCA

Vectors

For this course, typically think of vectors and matrices as just sets of numbers:

Notation

Vectors - Lower case Roman letters, bolded

$$\mathbf{x}_i = 2^i, \quad i = 1, \cdots, n \qquad \mathbf{x} = \begin{bmatrix} 2 \\ \vdots \\ 2^n \end{bmatrix}$$

$$\mathbf{x}^T = [2, \cdots, 2^n]$$

$$a\,\mathbf{x} + b\,\mathbf{y} = \begin{bmatrix} a\,\mathbf{x}_1 + b\,\mathbf{y}_1 \\ \vdots \\ a\,\mathbf{x}_n + b\,\mathbf{y}_n \end{bmatrix}$$

Matrices

Notation

Matrices - Upper case Roman letters, not bolded

For $A \in \mathbb{R}^{m \times n}$, i indexes rows and j indexes columns.

$$(A)_{i,j} = a_{ij} = i - j, \quad i = 1, \cdots, m, \quad j = 1, \cdots, n$$

$$\begin{bmatrix} 0 & -1 & -2 & \cdots & (1-n) \\ 1 & 0 & & & \\ \vdots & & & & \\ m-1 & & \cdots & & m-n \end{bmatrix}$$

Abstract Linear Algebra

- Understanding matrix multiplication - why isn't it elementwise like addition?

  For linear operators L and M, want matrix multiplication to represent the composition $L \circ M$

$$
\begin{aligned}
(L \circ M)(x) &\to A_{LM}x \\
L(x) &\to A_L x \\
M(x) &\to A_M x \\
A_{LM} &= A_L A_M
\end{aligned}
$$

- Understanding why similar matrices share eigenvalues

$$B = P^{-1}AP$$

  We will see that A and B are different representations of the same linear operator. Since scalars are not affected by coordinate changes $\sigma(A) = \sigma(B)$. The eigenvectors are also the same, just represented in different coordinate systems.

Generic NN Node

Dot Product

Given vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, the dot or inner product is defined as

$$u \cdot v = \sum_{i=1}^{n} u_i \, v_i$$

Note 1: $\|u\|^2 = \sum_{i=1}^{n} u^2 = u \cdot u \geq 0$

Note 2: vector notation $u \cdot v = u^T v$

Algebraic Properties

Commutative

$$u \cdot v = v \cdot u$$

Scalar Multiplication

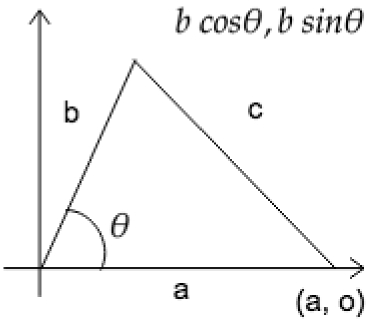
For $\alpha \epsilon \mathbb{R}$

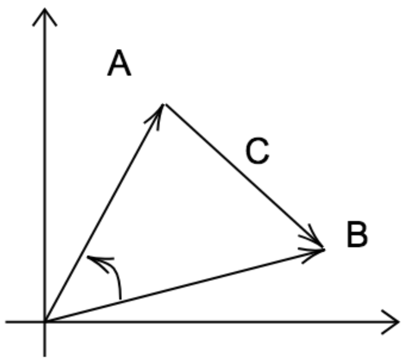
$$(\alpha u) \cdot v = u \cdot (\alpha v) = \alpha (u \cdot v)$$

Distributive Property

$$u \cdot (v + w) = u \cdot v + u \cdot w$$

Plane Geometry of Dot Product





Law of Cosines

$$
\begin{aligned}
c^2 &= (\cos\theta - a)^2 + (b\sin - 0)^2 \\
&= b^2\cos^2\theta - 2ab\cos\theta \\
&= a^2 + b^2\sin^2\theta \\
&= a^2 + b^2 - 2ab\cos\theta
\end{aligned}
$$

Law of Cosines, vector form

$$
\begin{aligned}
&C = B - A \\
&C \cdot C = (B - A) \cdot (B - A) \\
&\ = B \cdot B + A \cdot A - 2A \cdot B \\
&\|C\|^2 = \|B\|^2 + \|A\|^2 - 2A \cdot B
\end{aligned}
$$

Comparison with Law of Cosines gives -

$$
\begin{gathered}
2A \cdot B \leftrightarrow 2ab\cos\theta \\
A \cdot B = \|A\|\,\|B\|\cos\theta
\end{gathered}
$$

Geometry of Dot Product

Pythagoras Law: $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n, \mathbf{u} \perp \mathbf{v} \iff \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 = \|\mathbf{u} + \mathbf{v}\|^2$

$$
\begin{aligned}
\|\mathbf{u}+\mathbf{v}\|^2 &= (\mathbf{u}+\mathbf{v})^T(\mathbf{u}+\mathbf{v}) \\
&= \mathbf{u}^T\mathbf{u} + \mathbf{v}^T\mathbf{v} + 2\,\mathbf{u}^T\mathbf{v} \\
&= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 + 2\,\mathbf{u}\cdot\mathbf{v} \\
\Rightarrow \|\mathbf{u}+\mathbf{v}\|^2 &= \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2 \quad \text{if} \quad \mathbf{u}\cdot\mathbf{v} = 0
\end{aligned}
$$

It follows that

$$\mathbf{u} \perp \mathbf{v} \rightarrow \mathbf{u}\cdot\mathbf{v}$$

and

$$\mathbf{u}\cdot\mathbf{v} \rightarrow \mathbf{u} \perp \mathbf{v}$$

Note: If $\mathbf{u} \perp v \Leftrightarrow \mathbf{u} \cdot v = 0$ then $v \perp \mathbf{u}$ so $\|\mathbf{u} - v\|^2 = \|\mathbf{u}\|^2 + \|v\|^2$

Linear Subspace

A subspace of $\mathbb{R}^n$ is a collection of vectors $\mathbf{S}$ such that

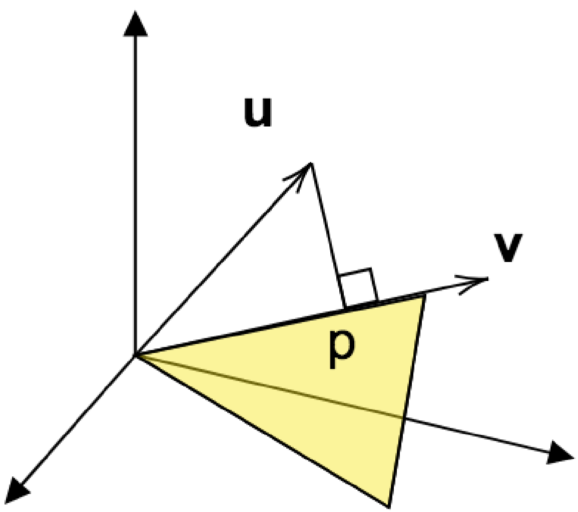
- The zero vector $\mathbf{0} \in S$
- If $\mathbf{u}$ and $\mathbf{v}$ are in S then $\mathbf{u} + \mathbf{v} \in S$
- If a is a scalar and $\mathbf{u} \in S$ then $a\,\mathbf{u} \in S$

Properties of Subspaces

1. Given a set of vectors $v_1 \cdots v_n \in \mathbb{R}^n$ then $\mathrm{span}(v_1 \cdots v_n)$ is a subspace
2. The set of vectors $\mathbf{x}$ satisfying $A\mathbf{x} = \mathbf{0}$ is a subspace (**Null space of A**)
3. Every non-zero subspace S of $\mathbb{R}^n$ has a finite basis. The number of vectors in a basis is the **dimension** of the subspace.
4. Convince yourself that the following are subspaces of the space of all $n \times n$ real matrices
   - The set S^n of all symmetric $n \times n$ matrices
   - The set T^n of all skew-symmetric $n \times n$ matrices
5. The set U^n of invertible $n \times n$ matrices is not a subspace

Projection Onto 1^d Subspace

Given vectors $\mathbf{u}$ and $\mathbf{v}$ in $\mathbb{R}^n$, define the projection of $\mathbf{u}$ on $\mathbf{v}$as the point on $\mathbf{v}$ which is closest to $\mathbf{u}$



Want α such that $\mathbf{p} = \alpha\ \mathbf{v}$ minimizes $\|\mathbf{u} - \alpha\mathbf{v}\|^2$

$$\mathbf{u} - \mathbf{p} = \mathbf{u} - \alpha\mathbf{v}$$
$$\mathbf{v} \cdot (\mathbf{u} - \alpha\mathbf{v}) = \mathbf{u} \cdot \mathbf{v} - \alpha\mathbf{v} \cdot \mathbf{v} = 0$$

Solving for α

$$\alpha = \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}}$$

Gives the projection vector $\mathbf{p}$

$$\begin{aligned} \mathbf{p} &= \frac{\mathbf{u} \cdot \mathbf{v}}{\mathbf{v} \cdot \mathbf{v}}\ \mathbf{v} \\ &= \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{v}\|}\ \hat{\mathbf{v}} \\ &= (\mathbf{u} \cdot \hat{\mathbf{v}})\ \hat{\mathbf{v}} \end{aligned}$$

Cauchy-Schwartz Inequality

For vectors $\mathbf{u},\ \mathbf{v}$ in $\mathbb{R}^n$

$|\mathbf{u}\cdot\mathbf{v}| \leq \|\mathbf{u}\|\|\mathbf{v}\|$

Proof

- Inequality holds if $\mathbf{v}=0$. Assume $\mathbf{v}\neq 0$ and project $\mathbf{u}$ onto $\mathbf{v}$

$$\begin{aligned}
\|\mathbf{u}\|^2 &= \|\mathbf{u}_{||}\|^2 + \|\mathbf{u}_{\perp}\|^2 \\
&\geq \|\mathbf{u}_{||}\|^2 \\
&= \|(\frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}})\mathbf{v}\|^2 \\
&= (\frac{\mathbf{u}\cdot\mathbf{v}}{\mathbf{v}\cdot\mathbf{v}})^2\ \|\mathbf{v}\|^2 \\
&= \frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{v}\|^2}
\end{aligned}$$

Result follows

$$\frac{(\mathbf{u}\cdot\mathbf{v})^2}{\|\mathbf{u}\|^2\|\mathbf{v}\|^2} \leq 1$$

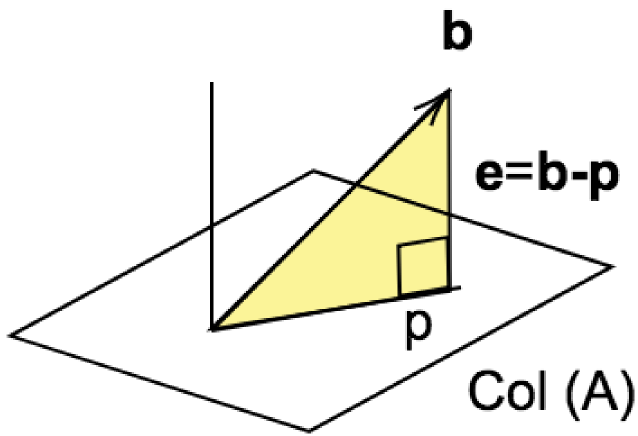
$$-1 \leq \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|} \leq 1$$

Define

$$\cos\theta_{\mathbf{uv}} = \frac{\mathbf{u}\cdot\mathbf{v}}{\|\mathbf{u}\|\|\mathbf{v}\|}$$

Projection Onto Subspace

Let A be an $m \times n$ matrix (m>n) with independent column vectors. For a vector $\mathbf{b}$, want to find orthogonal projection onto the column space of A.



- Want $\mathbf{p} = Ax$ for some x
- $\mathbf{p}$ is in col(A) so $\mathbf{e}$ = $\mathbf{b}$ - $\mathbf{p}$ is $\perp$ to all columns of A
- $\mathbf{e} \perp$ col(A) so $A^T\mathbf{e} = 0$

$$\begin{aligned} A^T\mathbf{e} &= A^T(\mathbf{b} - \mathbf{p}) \\ &= A^T(\mathbf{b} - \mathbf{Ax}) \\ &= A^T\,\mathbf{b} - A^T A\mathbf{x} \\ &= 0 \end{aligned}$$

Projection Onto Subspace (cont'd)

Solve to get:

$$\mathbf{x} = (A^T A)^{-1} A^T \mathbf{b}$$

$$\mathbf{p} = A\,\mathbf{x} = A(A^T A)^{-1} A^T \mathbf{b}$$

The matrix

$P \equiv A(A^T A)^{-1} A^T$ is the projection matrix into col(A)

$$P\,P = P = P^T$$

$$\begin{aligned} P^T &= (A(A^T A)^{-1} A^T)^T \\ &= A(A^T A)^{-1} A^T \end{aligned}$$

Note: If a matrix M is symmetric and invertible, then M^{-1} is symmetric P^{-1} is symmetric

$$\begin{aligned} PP &= [A(A^T A)^{-1} A^T][A(A^T A)^{-1} A^T] \\ &= A(A^T A)^{-1} A^T \\ &= P \end{aligned}$$

Discriminant Function/Decision Boundary

Consider linear discriminant function

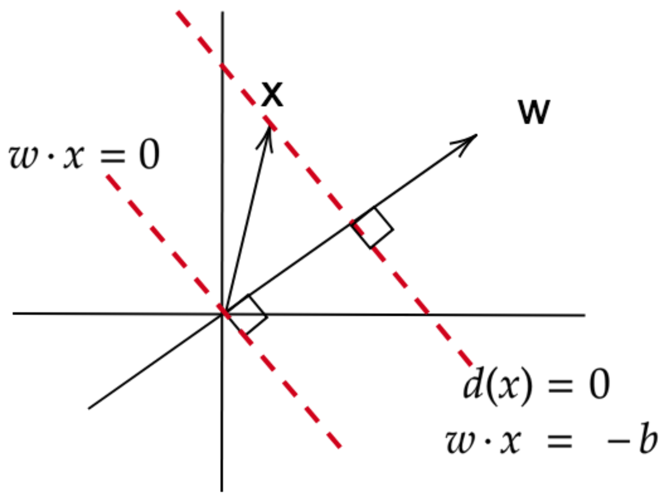
$$d(\mathbf{x}) = b + \mathbf{w} \cdot \mathbf{x}$$

for vectors $\mathbf{w}, \mathbf{x} \in \mathbb{R}^n$

if $d(\mathbf{x}) = 0$ then

$$\mathbf{w} \cdot \mathbf{x} = -b$$

$d(\mathbf{x}) = 0$ defines a $(n-1)$ dimensional hyperplane which splits $\mathbb{R}^n$



Linear Decision Boundaries

For $\mathbf{x} \in \mathbb{R}^d$, define discriminant $g(\mathbf{x}) = w_0 + \mathbf{w}^T\mathbf{x}$. The decision boundary is defined by the set of $\mathbf{x}$ satisfying $g(\mathbf{x}) = 0$.

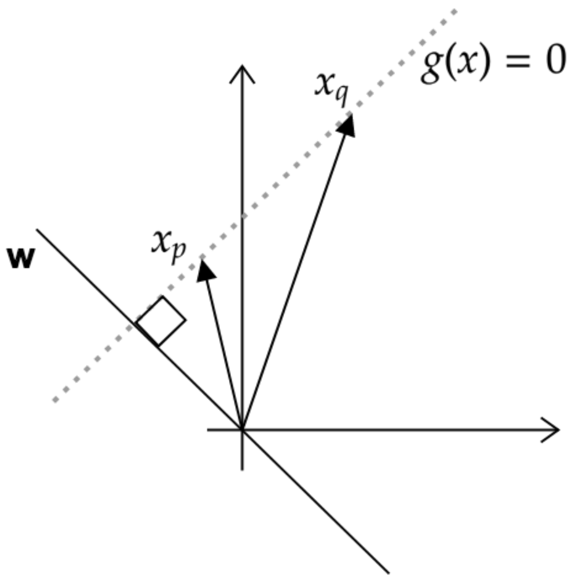
Consider points $\mathbf{x}_p$, $\mathbf{x}_q$ on the decision boundary

$$g(\mathbf{x}_p) = g(\mathbf{x}_q) = 0$$
$$w_0 + \mathbf{w}^T\mathbf{x}_p = w_0 + \mathbf{w}^T\mathbf{x}_q$$
$$\mathbf{w}^T(\mathbf{x}_p - \mathbf{x}_q) = 0$$

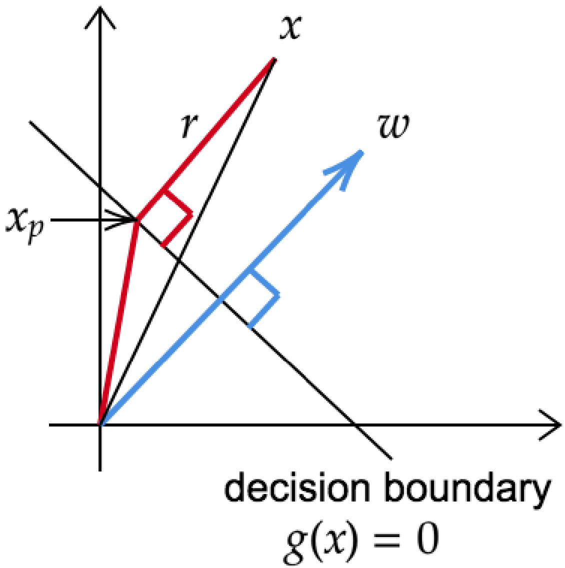
By construction $\mathbf{x}_p - \mathbf{x}_q$ lies in the direction of the decision boundary

$$\mathbf{w}^T(\mathbf{x}_p - \mathbf{x}_q) = 0$$

Shows decision boundary direction is $\perp$ to $\mathbf{w}$



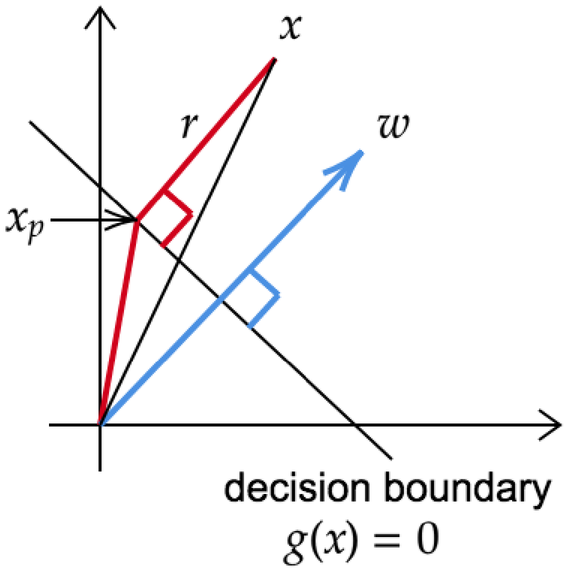
Linear Decision Boundaries



Computing orthogonal distance to boundary:

$$
\begin{aligned}
\mathbf{x} &= \mathbf{x}_p + r\frac{w}{\|w\|} \\
g(\mathbf{x}) &= w_0 + w^T(\mathbf{x}_p + r\frac{w}{\|w\|}) \\
&= \underbrace{w_0 + w^T\mathbf{x}_p}_{g(\mathbf{x}_p)=0} + r\frac{\mathbf{w}^T\mathbf{w}}{\|w\|} \\
g(\mathbf{x}) &= r\|w\| \\
r &= \frac{g(\mathbf{x})}{\|w\|}
\end{aligned}
$$

Linear Decision Boundaries



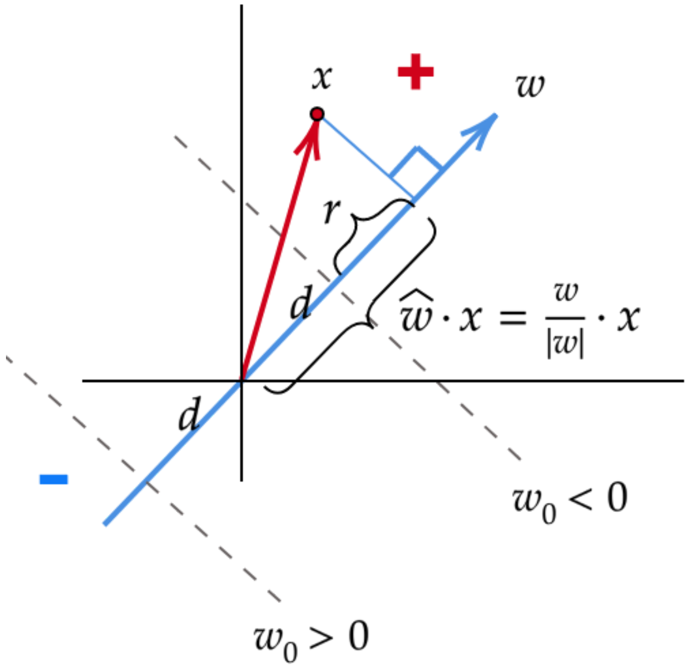
- $\frac{g(\mathbf{x})}{\|w\|}$ is the distance from $\mathbf{x}$ to the decision boundary
- if $\|w\| = 1$ then $g(\mathbf{x})$ is the distance
- Distance to origin from decision boundary:

$$r_0 = \frac{g(0)}{\|w\|} = \frac{w_0}{\|w\|}$$

- Distance is 'signed'

Linear Decision Boundaries

A second time:



$$
\begin{aligned}
g(\mathbf{x}) &= w_0 + w \cdot \mathbf{x} \\
d &= \frac{w_0}{\|w\|} \\
r &= \hat{w} \cdot \mathbf{x} + d \\
&= \hat{w} \cdot \mathbf{x} + \frac{w_0}{\|w\|} \\
&= \frac{1}{\|w\|}(w_0 + w \cdot \mathbf{x}) \\
&= \frac{g(x)}{\|w\|}
\end{aligned}
$$

Matrix Multiplication

Scalar Times Matrix

For matrix A $\in \mathbb{R}^{m\times n}$ and scalar b, the product $b\,A$ is elementwise multiplication

$$(bA)_{i,j} = bA_{i,j}$$

$$bA = \begin{bmatrix} b\,a_{1,1} & b\,a_{1,2} & \cdots \\ b\,a_{2,1} & \ddots & \\ \vdots & & \end{bmatrix}$$

```r
# scalar multiplication uses
# the '*' operator
A=matrix(1:6,nrow=3)
b=2
print(b*A)
```

```
##      [,1] [,2]
## [1,]    2    8
## [2,]    4   10
## [3,]    6   12
```

```r
(b*A)[3,2]==b*A[3,2]
```

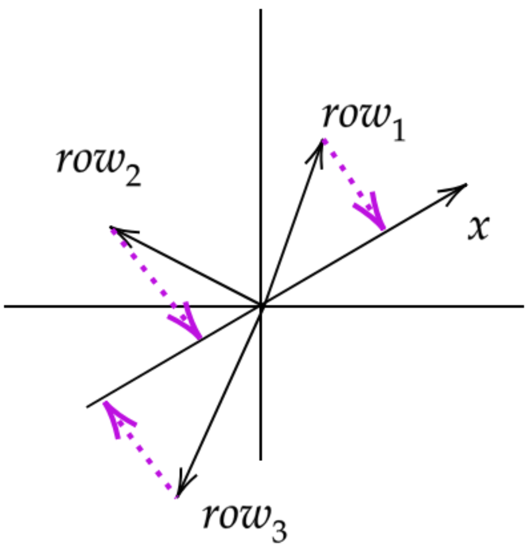
```
## [1] TRUE
```

Matrix Multiplication

Matrix Times Vector - Row projections onto vector

For matrix A $\in \mathbb{R}^{m \times n}$ and vector $\mathbf{b} \in \mathbb{R}^n$, the product $A\mathbf{b} \in \mathbb{R}^m$ is defined as

$$(A\mathbf{b})_i = row_i(A) \cdot \mathbf{b}$$




```
# matrix by matrix multiplicatio
# uses the '%*%' operator
A=matrix(1:6,nrow=3)
b=c(1,2)
print(A%*%b)
```

```
##      [,1]
## [1,]    9
## [2,]   12
## [3,]   15
```

```
dot=function(x,y){sum(x*y)}
(A%*%b)[2]==dot(A[2,],b)
```

```
## [1] TRUE
```

Matrix Multiplication

Matrix Times Vector - Weighted sum of columns

For matrix A $\in \mathbb{R}^{m \times n}$ and vector $\mathbf{b} \in \mathbb{R}^n$, the product $A\mathbf{b} \in \mathbb{R}^m$ is defined as

$$A\,\mathbf{b} = \begin{bmatrix} b_1\,a_{1,1} + b_2\,a_{1,2} + \ldots + b_n\,a_{1,n} \\ b_1\,a_{2,1} + b_2\,a_{2,2} + \ldots + b_n\,a_{2,n} \\ \vdots \\ b_1\,a_{m,1} + b_2\,a_{m,2} + \ldots + b_n\,a_{m,n} \end{bmatrix}$$

$$= \sum_i b_i\; col_i(A)$$

Clearly, $A\mathbf{b}$ is in the column space of A.

```
# matrix by matrix multiplicatic
# uses the '%*%' operator
A=matrix(1:6,nrow=3)
b=c(1,2)
print(A%*%b)
```

```
##       [,1]
## [1,]    9
## [2,]   12
## [3,]   15
```

```
(A%*%b)[2]==(b[1]*A[,1]+b[2]*A[,
```

```
## [1] TRUE
```

Matrix Multiplication

Matrix Times Matrix - Hadamard

For matrix A $\in \mathbb{R}^{m\times n}$ and matrix B $\in \mathbb{R}^{m\times n}$, the product $A \odot B \in \mathbb{R}^{m\times n}$ is defined as

$$(AB)_{i,j} = a_{i,j}b_{i,j}$$

```
# matrix by matrix multiplicatio
# uses the '%*%' operator
A=matrix(1:6,nrow=3)
B=matrix(1:6,nrow=3)
print(A*B)
```

```
##      [,1] [,2]
## [1,]    1   16
## [2,]    4   25
## [3,]    9   36
```

```
(A*B)[3,2]==A[3,2]*B[3,2]
```

```
## [1] TRUE
```

Matrix Multiplication

Matrix Times Matrix - Traditional View

For matrix A $\in \mathbb{R}^{m\times n}$ and matrix B $\in \mathbb{R}^{n\times p}$, the product AB $\in \mathbb{R}^{m\times p}$ is defined as

$$(AB)_{i,j} = row_i(A) \cdot col_j(B) \\ = \sum_k a_{i,k} b_{k,j}$$

Note: if $p = 1$ then B is a vector.

```
# matrix by matrix multiplicatio
# uses the '%*%' operator
A=matrix(1:6,nrow=3)
B=matrix(1:8,nrow=2)
print(A%*%B)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    9   19   29   39
## [2,]   12   26   40   54
## [3,]   15   33   51   69
```

```
dot=function(x,y){sum(x*y)}
(A%*%B)[2,3]==dot(A[2,],B[,3])
```

```
## [1] TRUE
```

Matrix Multiplication

Matrix Times Matrix - Second View

For matrix A $\in \mathbb{R}^{m\times n}$ and matrix B $\in \mathbb{R}^{n\times p}$, the product AB $\in \mathbb{R}^{m\times p}$ is defined as

$$AB = [A\,col_1(B), \dots, A\,col_p(B)]$$

```
# matrix by matrix multiplicatic
# uses the '%*%' operator
A=matrix(1:6,nrow=3); B=matrix(1
print(A%*%B)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    9   19   29   39
## [2,]   12   26   40   54
## [3,]   15   33   51   69
```

```
AB=matrix(0,nrow=nrow(A),ncol=nc
for(i in 1:ncol(B)){
  AB[,i]=A%*%B[,i,drop=FALSE]
}
all((A%*%B)==AB)
```

```
## [1] TRUE
```

Matrix Multiplication

Matrix Times Matrix - Third View

For matrix A $\in \mathbb{R}^{m\times n}$ and matrix B $\in \mathbb{R}^{n\times p}$, the product AB $\in \mathbb{R}^{m\times p}$ is defined as

$$AB = \begin{bmatrix} row_1(A)B \\ \vdots \\ row_m(A)B \end{bmatrix}$$

```
# matrix by matrix multiplicatio
# uses the '%*%' operator
A=matrix(1:6,nrow=3); B=matrix(1
print(A%*%B)
```

```
##      [,1] [,2] [,3] [,4]
## [1,]    9   19   29   39
## [2,]   12   26   40   54
## [3,]   15   33   51   69
```

```
AB=matrix(0,nrow=nrow(A),ncol=nc
for(i in 1:nrow(A)){
  AB[i,]=A[i,,drop=FALSE]%*%B
}
all((A%*%B)==AB)
```

```
## [1] TRUE
```

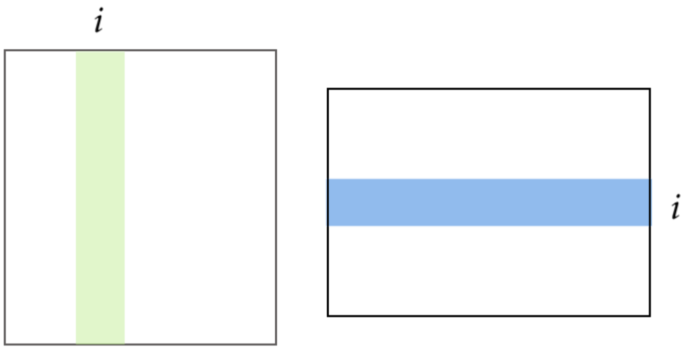
Matrix Multiplication

> "There is a fourth way to multiply matrices. Not many people realize how important this is. I feel like a magician explaining a trick. Magicians won't do it but mathematicians try."
>
> --*Strang* (pg 72)

Matrix-Matrix Multiplication - Take 4

For matrix A $\in \mathbb{R}^{m \times n}$ and matrix B $\in \mathbb{R}^{n \times p}$, the product AB $\in \mathbb{R}^{m \times p}$ is

$$AB = \sum_i col_i(A) \otimes row_i(B)$$



Matrix Multiplication

The rule for matrix multiplication is defined so that the composition of linear operators is expressed as the product of matrices.

$$U \leftrightarrow M_U$$

$$V \leftrightarrow M_V$$

$$(U \circ V)\mathbf{x} \leftrightarrow M_U(M_V\mathbf{x})$$

In coordinates $AB = \sum_j a_{ij}b_{jk}$

Properties

Associative

- $A(BC) = (AB)C$

Not commutative

- $AB \neq BA$

Distributive

- $A(B + C) = AB + AC$

Matrix Multiplication

Rank one matrix

Let $\mathbf{u}$ be a vector in $\mathbb{R}^d$. Define the matrix U as

$$U = \mathbf{u}\mathbf{u}^T$$

Consider $U\mathbf{v}$.

$$\begin{aligned} U\mathbf{v} &= (\mathbf{u}\mathbf{u}^T)\mathbf{v} \\ &= \mathbf{u}(\mathbf{u}^T\mathbf{v}) \\ &= (\mathbf{u} \cdot \mathbf{v})\mathbf{u} \end{aligned}$$

So $U\mathbf{v}$ is in the direction of $\mathbf{u}$

Special Matrices

Delta Matrix

The δ matrix

$$\delta_{ij} = \begin{cases} 1 \text{ if} & i = j \\ 0 & \text{o.w.} \end{cases}$$
$$i = 1, \cdots, m \quad j = 1, \cdots, n$$

Matrix Transpose

$(A^T)_{ij} = (A)_{j,i}$

if $A = A^T$, then A is called symmetric

Transpose of Product

$(AB)^T = B^T A^T$ (Will show later)

$A^T A$ and AA^T are symmetric

$$(A^T A)^T = (A^T (A^T)^T) = A^T A$$

$$(AA^T)^T = ((A^T)^T A^T) = AA^T$$

Orthogonal Matrices

If Q is a square matrix with orthonormal columns, then Q is called an orthogonal matrix.

$$col_i(Q) \cdot col_j(Q) = \delta_{i,j}$$

It follows that

$$Q^TQ = I$$
$$QQ^T = I$$

So, Q^T is a left and right inverse.

Properties

$$\|Q\mathbf{x}\| = \sqrt{\|Q\mathbf{x}\|^2} = \sqrt{\mathbf{x}^TQ^TQx} = \sqrt{\mathbf{x}^T\mathbf{x}} = \|\mathbf{x}\|$$

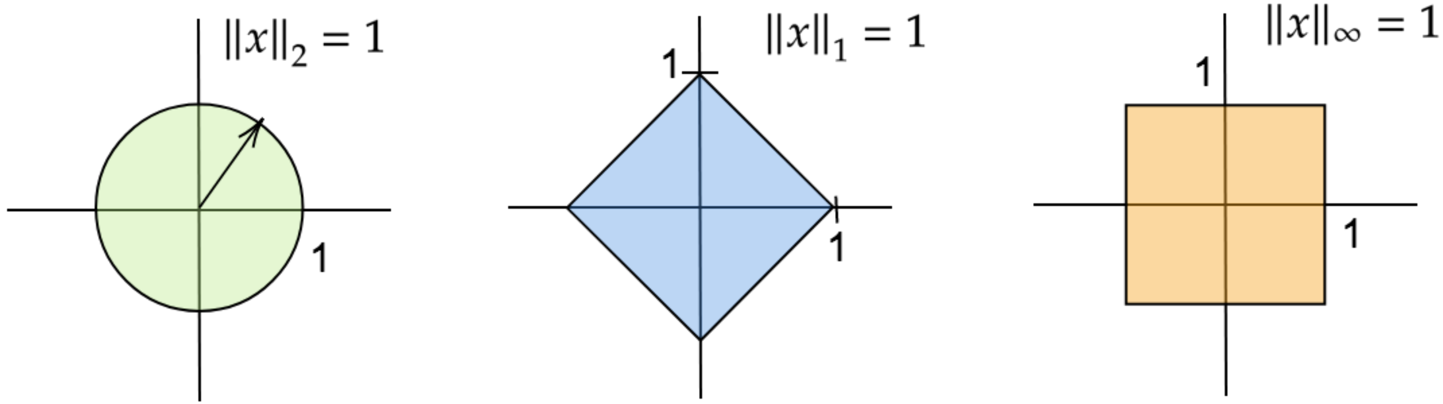
$$(Q\mathbf{x})^T(Q\mathbf{y}) = \mathbf{x}^TQ^TQ\mathbf{y} = \mathbf{x}^T\mathbf{y}$$

Vector Norms

Vector Norm (length)

$$
\begin{aligned}
\|\mathbf{x}\|_2 &= (\sum \mathbf{x}_i^2)^{\frac{1}{2}} \\
&= (\mathbf{x}^T\mathbf{x})^{\frac{1}{2}} \\
&= (\mathbf{x} \cdot \mathbf{x})^{\frac{1}{2}} \\
\|\mathbf{x}\|_1 &= \sum |\mathbf{x}_i| \\
\|\mathbf{x}\|_\infty &= \max_i |\mathbf{x}_i|
\end{aligned}
$$

Unit Vectors



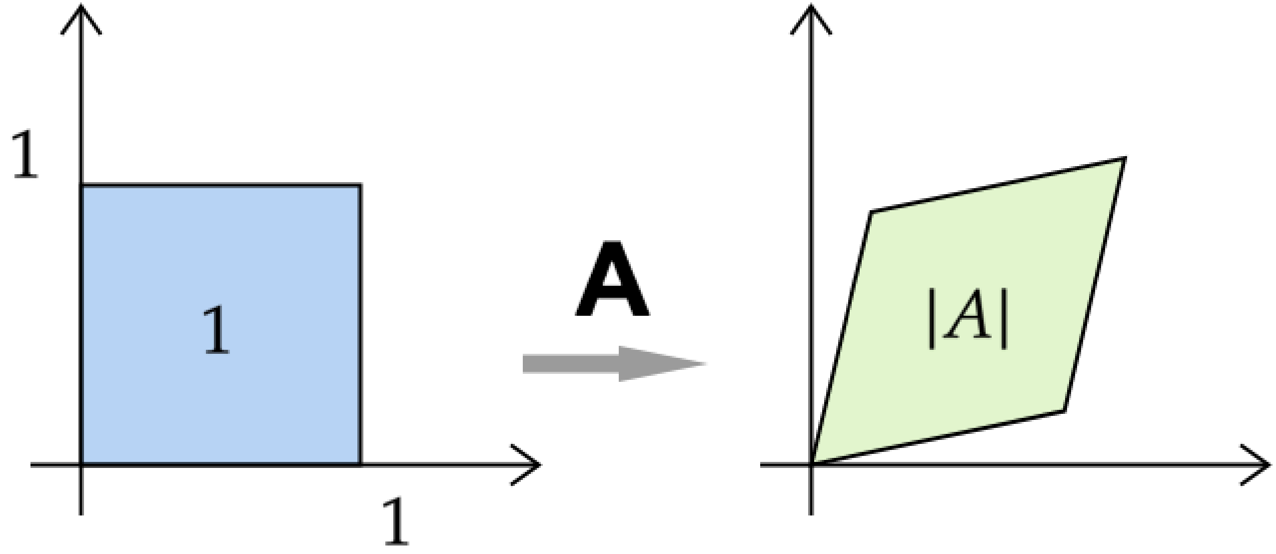
Linear Algebra

Trace

Sum along main diagonal $tr(A) = \sum_i a_{ii}\ i = 1, \cdots, n$

Determinant

$det(A) = |A|$



Linear Algebra

Determinant

Gives the area/volume magnification of a linear transformation $\mathbf{x} \rightarrow A\mathbf{x}$

- Determinant can be positive, negative or zero.
- Will show later that matrix multiplication is linear so that lines $\rightarrow$ lines.

$$A(t\mathbf{x} + (1-t)\mathbf{y}) = tA\mathbf{x} + (1-t)A\mathbf{y}$$

- $det(AB) = det(A)det(B)$

Matrix Inverse

Matrix Inverse (Square Matrices)

If A is invertible then

- $AA^{-1} = A^{-1}A = I$
- $det(A) = |A| \neq 0$
- $A\mathbf{x} = 0 \Rightarrow \mathbf{x} = A^{-1} \cdot 0 = 0$
- $(AB)^{-1} = B^{-1}A^{-1}$
- $(B^{-1}A^{-1})(AB) = B^{-1}IB = B^{-1}B = I$

Determinants

$$det(I) = 1$$

$$det(A^{-1}A) = det(A^{-1})det(A) = 1$$

$$det(A^{-1}) = \frac{1}{det(A)}$$

Linear Algebra

if A is not invertible then there is a an $\mathbf{x} \neq 0$ such that $A\mathbf{x} = 0$

- Imagine forming a volume with $\mathbf{x}$ as one side.
- $A\mathbf{x} = 0$ implies one side of volume is mapped to length zero.
- Area magnification is zero, so $det(A) = 0$.

$$A^{-1} \text{ exists iff } det(A) \neq 0$$

Linear Algebra

Vector Norms - behave like lengths

1. $\|\mathbf{x}\| \geq \forall \mathbf{x} \in \mathbb{R}^n$ and $\|\mathbf{x}\| = 0$ iff $\mathbf{x} = 0$
2. $\|a\mathbf{x}\| = |a|\|\mathbf{x}\|$
3. $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$

Matrix Norms - measure stretching

1. $\|A\| \geq 0 \ \forall A \in \mathbb{R}^{m\times\, n}$ and $\|A\| = 0$ iff $A = 0$
2. $\|aA\| = |a|\|A\|$
3. $\|A + B\| \leq \|A\| + \|B\|$
4. $\|AB\| \leq \|A\|\|B\|$, $A \in \mathbb{R}^{p\ x\ n}$

Linear Algebra

Properties of Matrix Norms

$$
\begin{aligned}
\|A+B\| &\leq \|A\| + \|B\| \\
\|(A+B)x\| &\leq \|Ax\| + \|Bx\| \\
&\leq \|A\|\|x\| + \|B\|\|x\| \\
\max_{x \neq 0} \frac{\|(A+B)x\|}{\|x\|} &\leq \|A\| + \|B\| \\
\Rightarrow \|A+B\| &\leq \|A\| + \|B\|
\end{aligned}
$$

$$
\begin{aligned}
\|AB\| &\leq \|A\|\|B\| \\
\|ABx\| &\leq \|A\|\|Bx\| \\
&\leq \|A\|\|B\|\|x\| \\
\max_{x \neq 0} \frac{\|ABx\|}{\|x\|} &\leq \|A\|\|B\|
\end{aligned}
$$

Linear Algebra

Any vector norm induces a matrix norm

$$\|A\|_\alpha = \sup_{x \neq 0} \frac{\|Ax\|_\alpha}{\|x\|_\alpha} = \max_{\|x\|_\alpha = 1} \|Ax\|_\alpha$$

1. $\|A\|_1 = \max_j \sum_{i=1}^m |a_{ij}|$ maximum absolute column sum
2. $\|A\|_\infty = \max_i \sum_{j=1}^n |a_{ij}|$ maximum absolute row sum
3. $\|A\|_2 \geq \sigma_{\max}(A)$

Linear Algebra

Froebenius Norm

The Froebenius norm is not induced by a vector norm. The Froebenius norm is the L_2 norm if A is treated as an element of $\mathbb{R}^{n \times n}$

$$\|A\|_F = (Trace(A^T A))^{\frac{1}{2}}$$

$$\begin{aligned}
\|A\|_F^2 &= Trace(A^T A) \\
&= \sum_i (A^T A)_{ii}) \\
&= \sum_i \sum_j \left[(A^T)_{ij} \; A_{ji} \right] \\
&= \sum_i \sum_j (A_{ij})^2
\end{aligned}$$

Linear Algebra

If U and V are orthogonal then $\|UA\|_F = \|AV\|_F = \|A\|_F$

$$\begin{aligned}\|UA\|_F^2 &= Trace[(UA)^T(UA)] \\ &= Trace[A^T U^T UA] \\ &= Trace[A^T A] \\ &= \|A\|_F^2\end{aligned}$$

$$\begin{aligned}\|Ax\| &\leq \|A\|_F \|x\| \\ A &= U\Sigma V^T \quad \text{full SVD} \\ \|A\|_F &= \|U\Sigma V^T\|_F \\ &= \|\Sigma V^T\|_F \\ &= \|\Sigma\|_F \\ &= (\sum_i \sigma_i^2)^{\frac{1}{2}}\end{aligned}$$

It follows that $\|A\|_F \geq \sigma_{max}(A)$

$$\|Ax\| \leq \sigma_{max}\|x\| \leq \|A\|_F\|x\|$$

Linear Algebra

Change of Basis

Given the canonical basis $\hat{e}_i = (0, \cdots, 1, \cdots 0)$

$$\mathbf{x} = \sum_i \mathbf{x}_i \hat{e}_i = E\mathbf{x}$$

For new basis $\hat{f}_i$,

$$\mathbf{x} = \sum_i \mathbf{x}_i \hat{f}_i = F\mathbf{x}^{'}$$

Where F is the matrix with columns $\hat{f}_i$ expressed in the $\hat{e}_i$ basis

Given linear transformation A, want to find $A^{'}$

$$A\mathbf{x} = \mathbf{y}$$
$$AF\mathbf{x}^{'} = F\mathbf{y}^{'}$$
$$F^{-1}AF\mathbf{x}^{'} = \mathbf{y}^{'}$$

$$A^{'} = F^{-1}AF$$

So $A^{'}$ is **similar** to A. Will show later that similar matrices have the same eigenvalues and eigenvectors.

Linear Algebra

Change of Basis example

$$A = \frac{1}{4}\begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix}$$

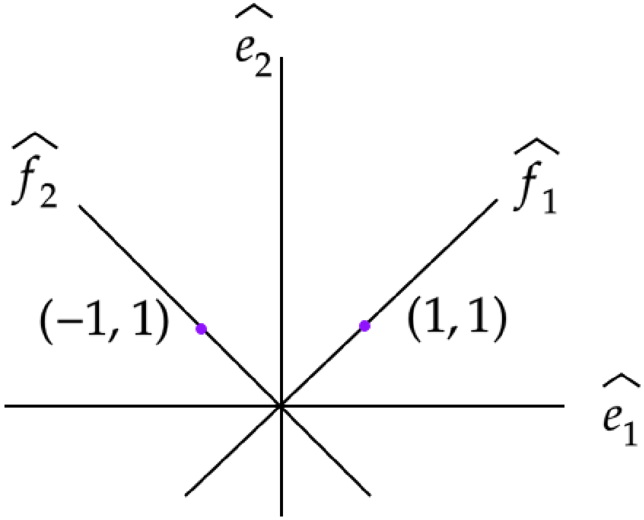
$$A\begin{bmatrix} 1 \\ 1 \end{bmatrix} = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$$

$$A\begin{bmatrix} 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{2} \\ \frac{1}{2} \end{bmatrix}$$

$$\left.\begin{array}{r} \hat{f}_1 = \frac{1}{\sqrt{2}}(1,1) \\ \hat{f}_2 = \frac{1}{\sqrt{2}}(-1,1) \end{array}\right\}_F = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

$$F^{-1} = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}$$

Linear Algebra



Let $\mathbf{x} = (1, 1)$

$$\mathbf{x}' = F^{-1}\mathbf{x} =$$

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} 1 \\ 1 \end{pmatrix} = \frac{1}{\sqrt{2}}\begin{pmatrix} 2 \\ 0 \end{pmatrix} = \begin{pmatrix} \sqrt{2} \\ 0 \end{pmatrix}$$

Linear Algebra

Change of Basis Example

$$B = F^{-1}AF = (\frac{1}{\sqrt{2}})^2 \frac{1}{4} \begin{pmatrix} 1 & 1 \\ -1 & 1 \end{pmatrix} \begin{pmatrix} 5 & 3 \\ 3 & 5 \end{pmatrix} \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix} = \begin{pmatrix} 2 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}$$

in new coordinates clearly see that:

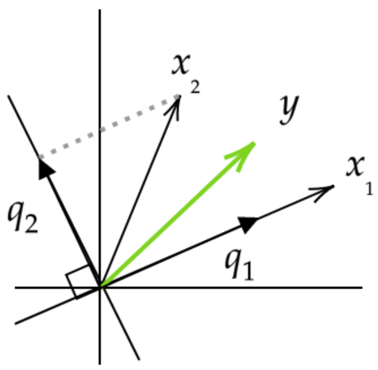
- linear operator stretches by 2 in one direction
- shrinks by $\frac{1}{2}$ in other direction

Independence, Span, Rank

- A set of vectors in $\mathbb{R}^n$ are independent if $\sum_{i=1}^{m} \alpha_i \mathbf{x}_i = 0$ implies $\alpha_i = 0 \quad \forall \quad i$
- If any $\mathbf{x}_i$ is a multiple of $\mathbf{x}_j$ for $j \neq i$ then the set is **dependent**
- If $m > n$ then the vectors form a dependent set. There are at most n independent vector in $\mathbb{R}^n$
- The Gram-Schmidt procedure produces an orthonormal basis for the subspace spanned by an independent set of vectors

Independence, Span, Rank

- Use Gram-Schmidt to form orthonormal basis from set of independent vectors



$$q_1 = \frac{x_1}{\|x\|}$$

$$q_i = \frac{(x_i - P\,x_i)}{\|x_i - Px_i\|}$$

$$P\,x_i = \sum_{j=1}^{i-1} (x_i \cdot q_j)\; q_j$$

- Once the $\mathbf{q}_i$ have been constructed

from the independent $\mathbf{x}_i$, then any vector $\mathbf{y}$ can be expressed as

$$\mathbf{y} = \sum_i (\mathbf{y} \cdot \mathbf{q_i})\mathbf{q_i} \quad \text{coordinate expansion}$$

Independence, Span, Rank

Let the columns A_i of matrix A be independent

- $A\mathbf{x} = \sum_i A_i x_i$ is a linear combination of independent vectors so there is no $\mathbf{x} \neq 0$ such that $A\mathbf{x} = 0$
- if A is $n \times n$ with independent columns then A is invertible
- If A be $\in \mathbb{R}^{m \times n}$ If $m > n$ then Ax is over-determined

Rank

Let A be $\in \mathbb{R}^{m \times n}$. Want to show

row rank(A)=column rank(A)

- Let r be the column rank of A. Let C be the matrix whose columns form a basis of $col(A)$ so the columns of A are linear combinations of the columns of C
- This implies there is a matrix $R \in \mathbb{R}^{r \times n}$ such that $A = CR$ (Place the linear combination coefficients in columns of R)
- $A = CR$ so the rows of A are linear combinations of the rows of R. R is $r \times n$ so $\mathrm{rowrank}(A) \leq r \rightarrow \mathrm{rowrank}(A) \leq \mathrm{colrank}(A)$
- Repeat using A^T to complete proof

Matrix Calculus

Quote from Wiki page

> Two competing notational conventions split the field of matrix calculus into two separate groups. The two groups can be distinguished by whether they write the derivative of a scalar with respect to a vector as a column vector or a row vector.

Numerator Layout

$$\frac{\partial f}{\partial \mathbf{x}} = [\frac{\partial f}{\partial \mathbf{x_1}} \dots \frac{\partial f}{\partial \mathbf{x_d}}]$$

Denominator Layout

$$\frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f}{\partial \mathbf{x_1}} \\ \vdots \\ \frac{\partial f}{\partial \mathbf{x_d}} \end{bmatrix}$$

Matrix Calculus Wiki Page

Index Notation

$$
\begin{array}{rl}
\text{Index Notation} & \text{Vector Notation} \\
\mathbf{x}_i & \rightarrow \mathbf{x} \quad \text{or} \quad \mathbf{x}^T \\
\sum_i \mathbf{w}_i \mathbf{x}_i & \rightarrow \mathbf{w} \cdot \mathbf{x} \\
\sum_i \mathbf{x}_i^2 & \rightarrow \mathbf{x}^T \mathbf{x} = \mathbf{x} \cdot \mathbf{x} = \|\mathbf{x}\|^2 \\
\sum_i a_{ij} \mathbf{x}_i & \rightarrow A^T \mathbf{x} \quad \text{or} \quad \mathbf{x}^T A \\
a_{ij} b_{ij} & \rightarrow A \odot B \qquad \text{(Hadamard Product)} \\
\sum_{ij} a_{ij} \mathbf{x}_i \mathbf{x}_j & \rightarrow \mathbf{x}^T A \mathbf{x}
\end{array}
$$

Derivatives of Matrices and Vectors

By convention, if $f : \mathbb{R}^n \to \mathbb{R}^m$, then the Jacobian matrix is written:

$$J(f(\mathbf{x})) = \frac{\partial f}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}$$

So the i^{th} column is $\frac{\partial f}{\partial x_i}$ and the i^{th} row is $\nabla_x f_i$

Note 1: If $x \in \mathbb{R}$, J is a column vector.

Note 2: If $f : \mathbb{R}^n \to \mathbb{R}$, J is a row vector

Derivatives of Matrices and Vectors (cont'd)

In the following, will use

- $(AB)^T = B^T A^T$
- $(AB)^{-1} = B^{-1} A^{-1}$ for invertible A, B

Derivative of Discriminant

$$
\begin{aligned}
y &= b + \mathbf{w} \cdot \mathbf{x} \\
&= b + w_1 x_1 + \cdots + w_n x_n \\
&= b + \mathbf{w}^T \mathbf{x}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial y}{\partial x_k} &= w_k \\
\frac{\partial y}{\partial \mathbf{x}} &= \mathbf{w}^T \qquad \text{row vector} \\
\frac{\partial y}{\partial \mathbf{w}} &= \mathbf{x} \\
\frac{\partial y}{\partial b} &= 1
\end{aligned}
$$

Derivatives of Matrices and Vectors (cont'd)

In the following, will use

- $(AB)^T = B^T A^T$
- $(AB)^{-1} = B^{-1} A^{-1}$ for invertible A, B

Derivative of Matrix Equation

$$\mathbf{y} = A\mathbf{x}$$
$$y_i = \sum_j A_{ij} x_j$$

$$\frac{\partial y_i}{\partial x_j} = A_{i,j}$$
$$\frac{\partial \mathbf{y}}{\partial \mathbf{x}} = A \qquad \text{(following convention for J)}$$
$$\frac{\partial y_i}{\partial A_{p,q}} = \sum_j \frac{\partial A_{ij}}{\partial A_{pq}} x_j$$
$$= \sum_j \delta_{i,p} \delta_{j,q} x_j$$

Derivatives of Matrices and Vectors (cont'd)

3a. $y = \mathbf{x}^T A \mathbf{x}, \qquad y \in \mathbb{R}$

$$y = \sum_{ij} a_{i,j} \mathbf{x}_i \mathbf{x}_j$$

$$\frac{\partial y}{\partial \mathbf{x}_d} = \sum_{i,j} a_{ij} (\mathbf{x}_i \frac{\partial \mathbf{x}_j}{\partial \mathbf{x}_d} + \frac{\partial \mathbf{x}_i}{\partial \mathbf{x}_d} x_j)$$

$$= \sum_{i,j} a_{ij} (\mathbf{x}_i \delta_{j,d} + \delta_{i,d} x_j)$$

$$= \sum_i a_{i,d} x_i + \sum_j a_{d,j} x_j$$

$$\begin{aligned} \frac{\partial y}{\partial \mathbf{x}} &= x^T A + x^T A^T \\ &= x^T (A + A^T) \end{aligned}$$

3b.

$$y = \|A\mathbf{x}\|^2 = x^T A^T A x$$

- Use (3a) and fact that $A^T A$ is symmetric to get

$$\frac{\partial y}{\partial \mathbf{x}} = 2\mathbf{x}^T (A^T A)$$

Derivatives of Matrices and Vectors (cont'd)

4.

$y = \|\mathbf{x}\|^2 = x^T x = \sum_i x_i^2$

$$\begin{aligned} \frac{\partial y}{\partial x_k} &= 2\sum_i x_i \frac{\partial x_i}{\partial x_k} \\ &= 2\sum_i x_i \delta_{ik} \\ &= 2\mathbf{x}_k \end{aligned}$$

$$\frac{\partial y}{\partial \mathbf{x}} = 2\mathbf{x}$$

$$(A)_{ij} = a_{ij}$$

$$(\frac{\partial A}{\partial x})_{ij} = \frac{\partial a_{ij}}{\partial x}$$

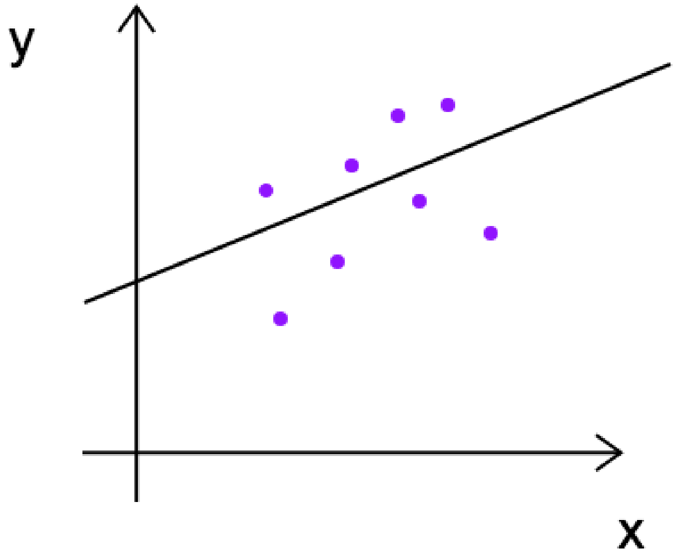
$$AA^{-1} = I$$

$$\frac{\partial A}{\partial \mathbf{x}} A^{-1} + A\frac{\partial A^{-1}}{\partial \mathbf{x}} = 0$$

$$\frac{\partial A^{-1}}{\partial x} = -A^{-1}\frac{\partial A}{\partial x}A^{-1}$$

Least Squares

Let A be a matrix with each row a set of n measurements $\mathbf{x}, \mathbf{y}$ called samples. The idea is to find $\mathbf{x} \in \mathbb{R}^n$ which minimizes the errors $A\mathbf{x} - \mathbf{y}$ where $\mathbf{y}$ is an m-dimensional vector of responses.



- Cost $\mathcal{C} = \frac{1}{2}\|A\mathbf{x} - \mathbf{y}\|^2$
- Could use projection results. Solution is project of y onto subspace spanned by columns of A

Least Squares (cont'd)

$$\begin{aligned}\frac{\partial \mathcal{C}}{\partial \mathbf{x}} &= \frac{1}{2}\frac{\partial}{\partial \mathbf{x}}(A\mathbf{x}-\mathbf{y})^T(A\mathbf{x}-\mathbf{y}) \\ &= \frac{1}{2}\frac{\partial}{\partial \mathbf{x}}\left[\mathbf{x}^T A^T A\mathbf{x} - \mathbf{x}^T A^T\mathbf{y} - \mathbf{y}^T A\mathbf{x} + \mathbf{y}^T\mathbf{y}\right]\end{aligned}$$

Numerator Convention

$$\frac{2\mathcal{C}}{\partial \mathbf{x}} = \frac{1}{2}\left[2\mathbf{x}^T\left(A^T A\right) - 2\mathbf{y}^T A\right] = 0$$

Taking transpose

$$(A^T A)\mathbf{x} = A^T\mathbf{y}$$

$$\mathbf{x} = (A^T A)^{-1}A^T\mathbf{y}$$

Denominator Convention

$$\frac{2\mathcal{C}}{\partial \mathbf{x}} = \frac{1}{2}\left[2\left(A^T A\right)\mathbf{x} - 2A\mathbf{y}\right] = 0$$

$$(A^T A)\mathbf{x} = A^T\mathbf{y}$$

$$\mathbf{x} = (A^T A)^{-1}A^T\mathbf{y}$$