COMP 3223: Foundations of Machine Learning

Srinandan Dasmah

Matrices

You should all know the following

- Matrix notation
- Matr
- Scal
- Matr
- Matrix inverse
- System of linear equations in matrix form
- Matrix determinant
- Eigenvalues and eigenvectors

- In linear regression the model is $\mathbf{y} = \mathbf{A}\mathbf{w}$:

$$\begin{pmatrix} \vdots \\ \vdots \\ w_0\phi_0(x_N) + w_1\phi_1(x_N) + \cdots + w_p\phi \end{pmatrix} \quad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}$$

- Find weights $\{w_i\}$ that make residuals $\|\mathbf{r}$

- In linear regression the model is $\widehat{\mathbf{y}} = \mathbf{A}\mathbf{w}$:

$$w_0 \underbrace{\begin{pmatrix}\phi_0(x_1)\\ \vdots\end{pmatrix}}_{\mathrm{col}_0(\mathbf{A})} \quad \underbrace{\begin{pmatrix}\phi_1(x_1)\\ \vdots\end{pmatrix}}_{\mathrm{col}_1(\mathbf{A})} \quad \begin{pmatrix}\phi_p(x_1)\\ \vdots\end{pmatrix} = \begin{pmatrix}\widehat{y}_1\\ \widehat{y}_2\\ \vdots\\ \widehat{y}_N\end{pmatrix}$$

- Find **linear combination of columns** of
- Residual **r** not in space spanned by columns of each column:
- $\sum_n r_n\phi_i(x_n) = 0$ ... is where gradient of squared loss vanishes.

Design matrix has information on patterns in data

- In linear regression the model is $\mathbf{y} = \mathbf{A}\mathbf{w}$:

$$1 \quad \phi_1(x_N) \quad \cdots$$

- Idea of this lecture: decompose matrix using transfor
  appropriate descriptive bases

Reminder: Solving Linear Equations – Geometrical Picture

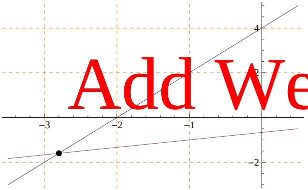
- Solve set of equations:

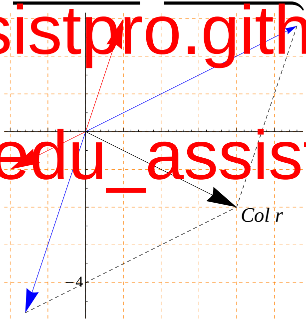
- Geo

$$\overbrace{\begin{array}{rcl} ax+by &=& r\\ cx+dy &=& s\end{array}}^{\text{rows of matrix}} \leftrightarrow x\overbrace{\begin{pmatrix} a\\ c\end{pmatrix}}^{\text{c}}$$

- Solve set of equations:

$$\begin{pmatrix} -2 & 1 \\ -1 & 3 \end{pmatrix} \begin{pmatrix} x \\ y \end{pmatrix} = \begin{pmatrix} 4 \\ -2 \end{pmatrix}$$

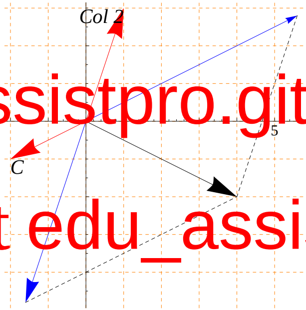




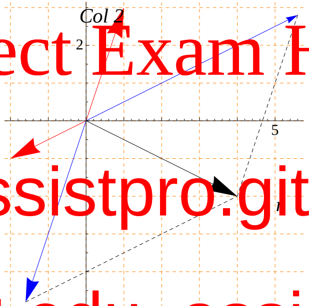
- Solution of $y - 2x = 4, 3y - x = -2$, is $(x, y) = (-2.8, -1.6)$.

Fundamental operations on vectors – multiply by scalars and perform addition

Multi
(elements of a field) and add
vectors together



Column space and Range of a matrix



Thus
only if
colum

- The column space of a matrix A (denoted
  all linear combinations of the columns of
- This is also the range of the linear map: range(A)=AV = $\{\mathbf{w} \in W : \mathbf{w} = A\mathbf{v} \text{ for some } \mathbf{v} \in V\}$

Examples illustrating linear dependence and nullspace

- Let $B = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{pmatrix}$. For vector $\mathbf{v}$ in direction $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$, $B\mathbf{v} = 0$. $\mathbf{v}$ in *nullspace* or *kernel* of B.

- For

$$\begin{pmatrix} 2 \\ 0 \\ 1 \end{pmatrix} + \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} - 1\begin{pmatrix} 3 \\ 1 \\ 1 \end{pmatrix} = \begin{pmatrix} \quad \end{pmatrix}$$

Show $\ker(A^\top) = c \begin{pmatrix} -1 \\ 1 \\ 2 \end{pmatrix}$.

Kernel or Null space of a matrix

- In the previous example $\mathbf{A} = \begin{pmatrix} 2 & 1 & 3 \\ 0 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix}$, there are 3 variables $\mathbf{v}$ in $A\mathbf{v} = \mathbf{y}$ but only two independent equations.
- If $\mathbf{Av}$ ... ns or there ...
- The k ... $V : A\mathbf{v} = 0\}$.
- Let $\mathbf{A}$ be a $3 \times p$ matrix.

$$\mathbf{A} = \begin{pmatrix} — & \mathbf{u} & — \\ — & \mathbf{v} & — \\ — & \mathbf{w} & — \end{pmatrix}$$

where $\mathbf{u}$, $\mathbf{v}$ and $\mathbf{w}$ are p-dim row vectors. Then, $\mathbf{x} \in$ker$(\mathbf{A}) \Leftrightarrow \mathbf{Ax} = 0$. This means $\mathbf{x} \perp \{\mathbf{u}, \mathbf{v}, \mathbf{w}\}$.

Rank of a matrix = number of independent equations

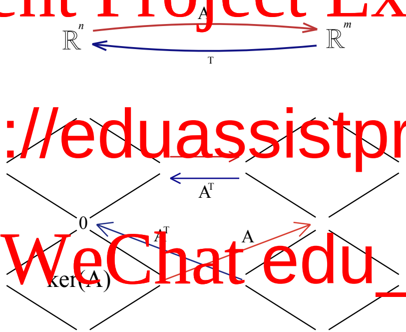
- The **rank** (column rank) of A is the dimension of the column space of A.
- A vect

nullity rank

- We can do the same for the transpose: col(
- 4 fundamental subspaces: col(A), ker($A^\top$

Four fundamental subspaces of a matrix

http://en.wikipedia.org/wiki/Fundamental_theorem_of_linear_algebra



Linear regression with fixed functions of data

- Linear combinations of fixed functions $\phi_j(\mathbf{x}_n)$:

$$1 \qquad \phi_1(x_N) \qquad \begin{pmatrix} y_1 \\ y_2 \\ \vdots \end{pmatrix}$$

- Sets of functions constitute vector spaces
- Approximate outputs/targets **y** by element o
  matrix.

Functions constitute vector spaces

- $\mathbb{R}[x]$, the space of polynomials $\sum_m a_m x^m$, where $a_m \in \mathbb{R}$ forms a vector space

$$(a_0 + a_1 x + a_2 x^2) + (b_0 + b_1 x) = (\underline{a_0} + \underline{b_0}) + (\underline{a_1} + \underline{b_1})\,x + \underbrace{a_2}_{c_2} x^2$$

Mon

$$(a_0, a_1, a_2) + (b_0, b_1, 0) = (a_0 + b$$

- Similarly the set $\mathbb{R}[x_1, \ldots, x_k]$ of polynomials i r space.
- Set of functions of the form $\sum_{|n|<N} a_n e^{in\theta}$ (Fourier series).
- Extension – replace sums (where the summation index is from a discrete set) by integrals (where the index being summed over is now continuous)

Even matrices form a vector space

- Matrices form a vector space: multiply $n \times m$ matrices A with entries $a_{ij} \in \mathbb{R}$, $i = 1, \ldots, n$, $j = 1, \ldots, m$ by scalars and add any two such matrices together:

$$3 \begin{matrix} -2 & 1 \end{matrix} \quad 2 \begin{matrix} 2 & 2 \end{matrix} = \begin{matrix} -10 & -1 \end{matrix} .$$

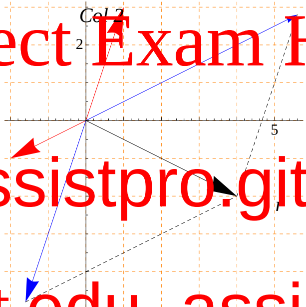
- imag

x_{11}	x_{12}		
x_{21}	x_{22}		
x_{31}	x_{32}		
$\vdots$	$\vdots$	$\ddots$	$\vdots$
x_{L1}	x_{L2}	$\cdots$	x_{LL}

$\mapsto$

$$\begin{pmatrix} {}_{1L} \\ x_{21} \\ \vdots \\ x_{LL} \end{pmatrix}$$

Reminder: Linear combination and dependence



Linea

$\mathbf{v} =$

- The vectors in the figure are linear combinations of $= \binom{0}{1}$. They are in the **span** of $\{\mathbf{e}_1, \mathbf{e}_2\}$.
- $\mathbf{v} = a_1\mathbf{e}_1 + a_2\mathbf{e}_2 = \binom{a_1}{a_2}$ can be zero iff $a_1 = 0 = a_2$.

Reminder: Linear independence & Basis

- A set of vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n$ are called **linearly independent** if none of them can be represented as a linear combination of the others.

- Equi ndent:

$$\text{If } \sum_{i=1}^{n} \alpha_i \mathbf{v}_i = 0, \text{ then}$$

.

- A **basis** for V is a set $B \subset V$ which is both spanning and independent. A finite dimensional vector space has a finite basis, and its dimension dim V is the number of elements in B.

Dot Products, Orthogonality and Norms

- We can associate, with two vectors $\mathbf{v}$ and $\mathbf{w}$ an element of $\mathbb{R}$ called their **scalar** (or **dot**) **product**:

- Two v ie. $\mathbf{v}_1 \cdot \mathbf{v}_2 = 0$. If k vectors $\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_k$ are mutually orthogonal, ie. $\mathbf{v}_i \cdot \mathbf{v}_j = 0$ for $i \neq j$, they are called an **orthogonal se**

- Euclidean norm for $\mathbf{v} \in V$, $\dim V = N$,

$$\|\mathbf{v}\| := \sqrt{\mathbf{v}^\top \mathbf{v}} = \sqrt{v_1^2 + v_2^2 + \cdots + v_N^2}$$

- If all vectors are of unit length $\|\mathbf{v}_i\| = 1$, the set is called **orthonormal**.

- **Project** Φ ctor $\mathbf{v}$
- ve if α is

$\mathbf{\Phi}_n \quad \mathbf{v}$

Note, pr

since **dot** produ .

Example: expand vector in orthogonal basis

- Let $\mathbf{e}_1 = \binom{1}{0}$, $\mathbf{e}_2 = \binom{0}{1}$. Expand $\mathbf{v} = \binom{-5}{3}$ as a *linear combination* of the set $\{\mathbf{e}_i\}$, ie. find numbers α_1, α_2 such that

- Solution: Multiply $\mathbf{v}$ by $\mathbf{e}_j$, use orthogonality ($= -5$, $\mathbf{e}_2 \cdot \binom{-5}{3} = 3$.

$$\begin{pmatrix} -5 \\ 3 \end{pmatrix} = (-5)\begin{pmatrix} 1 \\ 0 \end{pmatrix}$$

Expanding a vector in a set of orthogonal vectors

- Let $\{\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n\}$ be a set of orthogonal $n \times 1$ column vectors and $\mathbf{v}$ an arbitrary $n \times 1$ column vector.

- Task: Expand $\mathbf{v}$ as *linear combination* of set $\mathbf{v}_i \;_i$, ie. find numbers α_1, α
- Solu thogonality) to get

$$\begin{aligned} \mathbf{v}_j \cdot \mathbf{v} &= \alpha_1 \mathbf{v}_j \cdot \mathbf{v}_1 + \cdots + \\ &= \alpha_j \mathbf{v}_j \cdot \mathbf{v}_j \end{aligned}$$

Hence

$$\alpha_j = \frac{\mathbf{v}_j \cdot \mathbf{v}}{\mathbf{v}_j \cdot \mathbf{v}_j} = \frac{{}_j}{\|\mathbf{v}_j\|^2}$$

- Seek to characterise design matrix in terms of some orthonormal bases

Design matrix is not square

- The domain and range of matrix have different dimensions
- Need descriptive basis for each action of matrix on vector space pieced together from its action on orthonormal basis
- For in
  vect

$$\mathbf{Aw} = w_0 \mathbf{A} \begin{pmatrix} \vdots \\ 0 \end{pmatrix} + w_1 \mathbf{A} \begin{pmatrix} \vdots \end{pmatrix}$$

- $\mathbf{Aw} = w_0 \mathrm{col}_0(\mathbf{A}) + \cdots + w_p \mathrm{col}_p(\mathbf{A})$
- Introduce **singular value decomposition** (SVD) to find approximate subspaces
- Generalise notion of eigenvalue/eigenvector pair

Approximating subspaces

- Point $\boldsymbol{\Phi}_n = (\phi_{n1}, \phi_{n2}, \ldots, \phi_{np})$
- Project $\boldsymbol{\Phi}$ along $\mathbf{v}$
- $\|\boldsymbol{\Phi}_n\|$ [illegible]
- ... su [illegible] $\mathbf{v}$
- $(\mathrm{dist}_{n,\mathbf{v}})^2 = -(\mathrm{proj}_{n,\mathbf{v}})^2 + \|\boldsymbol{\Phi}_n\|^2$
- $\boldsymbol{\Phi}_n$ is $\mathbf{v}$-independent
- Minimising distance to $\mathbf{v}$ equivalent to maxim [illegible] projection along $\mathbf{v}$

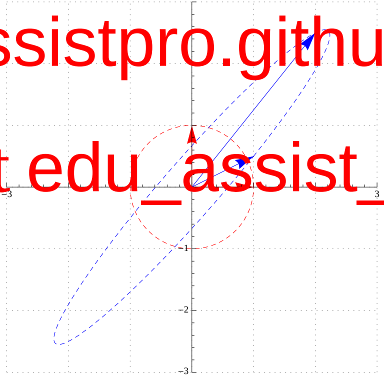
Two interpretations of best fit subspaces

1. mini
2. maxi
   cont
   dimension)

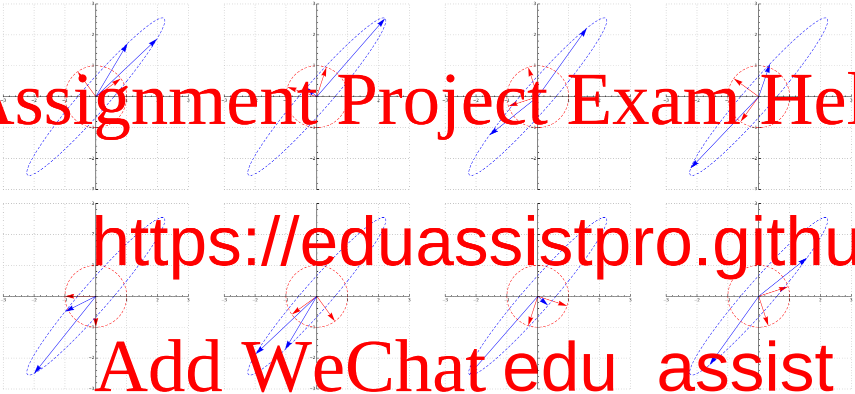
Singular Value Decomposition (SVD) of a Matrix

- The action of an arbitrary matrix on a vector space can be pieced together from its action on an orthonormal basis in that vector space.
- SVD measures how a circle is mapped into an ellipse;
- If matrix A is n-by-m, SVD of A characterises how an m-dimensional hype

Action of $\begin{pmatrix} 1.0 & 2.0 \\ 0.5 & 2.5 \end{pmatrix}$ on unit vectors $\binom{1}{0}$ and $\binom{0}{1}$. The lengths of the semi-major axes of the hyper-ellipse are properties of the map.



In pictures: mapping a unit circle into an ellipse



- Even when the vectors in the domain and range of the map change, their locus displays the geometrical character of the transformation enacted by the matrix.
- While the displayed pairs of vectors in the domain (red) are orthogonal by construction, the pairs they map to (blue) are usually not.

Example of SVD

- The action of an arbitrary matrix on a vector space can be pieced together from its action on an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ in that vector space. Here 2-dimensional. So, $\mathbf{w} = (\mathbf{v}_1^\top \mathbf{w})\mathbf{v}_1 + (\mathbf{v}_2^\top \mathbf{w})\mathbf{v}_2$.
- The ra lars for $j = 1$
- Acti

$$
\begin{aligned}
\mathbf{A}\mathbf{w} &= (\mathbf{v}_1^\top \mathbf{w})\mathbf{A}\mathbf{v} \\
&= (\mathbf{v}_1^\top \mathbf{w})\sigma_1 \mathbf{u} \\
&= (\sigma_1 \mathbf{u}_1 \mathbf{v} \\
\Rightarrow \mathbf{A} &= \mathbf{v}_1^\top \sigma_1 \mathbf{u}_1 +
\end{aligned}
$$

- Express that as $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top$, with $\mathbf{U}$ containing the columns of $\mathbf{u}_i$, $\mathbf{V}$ the columns of $\mathbf{v}_i$, and Σ a diagonal matrix with σ_i along the diagonal.

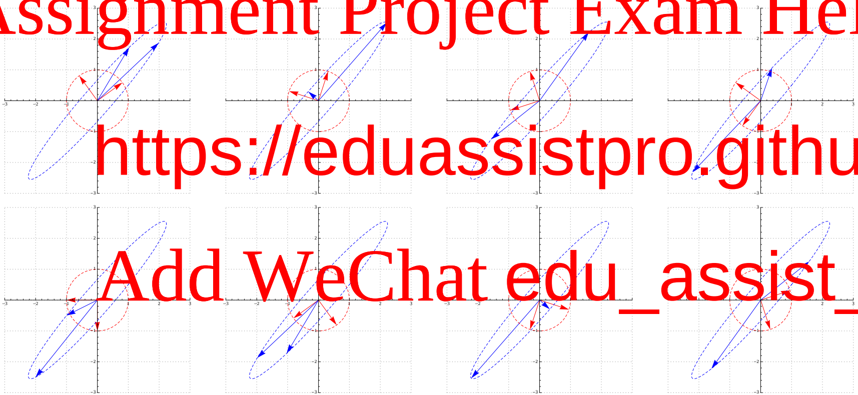
The full SVD describes both the domain and range of a matrix by orthonormal bases

- For an … nd a $m \times m$ matr … on-zero entri …

$$\mathbf{A} = \mathbf{U}\Sigma$$

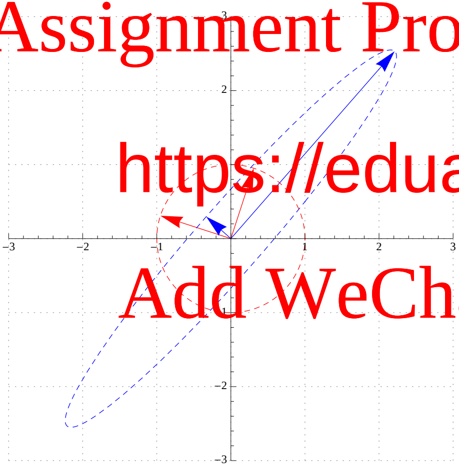
- The columns of $\mathbf{V}$ and $\mathbf{U}$ are the right and left si… diagonal entries of Σ are the singular values of

Geometry of SVD: choice of basis vectors lying on circle and map



♡ Choose the pre-image of the orthogonal pair in the range of the map.

Singular vectors describe spheres and ellipsoids by semi-major axes



- There is one choice of vector pairs
  s
  l pair
  ose.
- ngular
  s
- obtained by finding vectors $\mathbf{v}, \mathbf{u}$ that yield maximum lengths lengths of $\mathbf{A}\mathbf{v}$ and $\mathbf{A}^\top\mathbf{u}$

Linear regression using SVD: find $\mathbf{w}$ for smallest $\|\mathbf{Aw} - \mathbf{y}\|_2$

- A vect $\boldsymbol{\phi} = \alpha^* \mathbf{u}$.
  Proo

$$\alpha \in \mathbb{R} \qquad \overline{\mathbf{u}^\top \mathbf{u}}$$

- Recall linear regression: slope $w_1 = (\mathbf{y}^\top \mathbf{x})$

- Projection along $\mathbf{u}$.
- Use SVD to find singular vectors $\mathbf{u}_i$ and find projections $\mathbf{y} \cdot \mathbf{u}_i$.

Linear regression by SVD: express weights and targets in terms of singular vectors

- $\mathbf{A} = \mathbf{U}\Sigma\mathbf{V}^\top = \sum_{k=1}^{r} \mathbf{u}_k \sigma_k \mathbf{v}_k^\top$ or, equivalently, $\mathbf{A}\mathbf{v}_i = \sigma_i \mathbf{u}_i$.
- Weight space spanned by $\mathbf{v}_k \in \mathbb{R}^{P+1}$, output space spanned by $\mathbf{u}_k \in \mathbb{R}^N$:

- Near $_k^\top \mathbf{y})$.
- Mod

$$\mathbf{A}\mathbf{w} = \mathbf{A}(\sum_k \alpha_k \mathbf{v}_k) = \sum_k \alpha$$

- Best fit vector to $\mathbf{y}$ along each $\mathbf{u}_k$ is $\beta_k \mathbf{u}_k$ $\mathbf{A}$ along direction $\mathbf{u}_k$ is $\alpha_k \sigma_k \mathbf{u}_k$.
- Equating coefficients along $\mathbf{u}_i$,

$$\alpha_i \sigma_i = \beta_i = \mathbf{u}_i^\top \mathbf{y} \implies \alpha_i = \frac{\mathbf{u}_i^\top \mathbf{y}}{\sigma_i}.$$

Linear regression by SVD: small singular values are unwelcome

- The best fit weight vector is

- Sing
  could track "noise" in targets of training set, not "signa
  generalise.

- Requires **regularisation**: add positive constant
  minimising

$$\mathbf{w}^* = \operatorname*{argmin}_{\mathbf{w}} \|\mathbf{y} - \mathbf{A}\mathbf{w}\|^2 + \lambda\|\mathbf{w}\|^2 \implies \mathbf{w}^* = \sum_i \frac{\sigma_i}{\sigma_i^2 + \lambda}(\mathbf{u}_i^\top \mathbf{y})\mathbf{v}_i.$$