Linear Algebra

Lecture Notes

Rostyslav Hryniv

Ukrainian Catholic University
Business Analytics and Computer Science Programmes

4th term
Spring 2019

Lecture 8. Projections

Outline

1. Distance to a subspace
   - Shortest distance
2. Projections
   - Orthogonal decomposition
3. Least squares
   - Least squares solutions to $A\mathbf{x} = \mathbf{b}$
   - Linear, polynomial and multiple regression

Summary of the last lecture:

- In inner product vector spaces
  - the norm $\|\mathbf{u}\| = \langle \mathbf{u}, \mathbf{u} \rangle^{1/2}$ of $\mathbf{u}$ and
  - the distance $\|\mathbf{u} - \mathbf{v}\|$ between $\mathbf{u}$ and $\mathbf{v}$

  can be introduced;
- the inner product $\langle \cdot, \cdot \rangle$ and the related norm $\| \cdot \|$ satisfy
  - the Cauchy–Bunyakovski–Schwarz inequality
  - the triangle inequality
- every $m \times n$ matrix A generates two pairs of orthogonal subspaces (column space=range and nullspace of A and its transposed A^T)
  - $\mathbb{R}^m = \mathcal{C}(A) \oplus \mathcal{N}(A^T)$
  - $\mathbb{R}^n = \mathcal{C}(A^T) \oplus \mathcal{N}(A)$
- orthogonal vectors $\mathbf{u}$ and $\mathbf{v}$ satisfy the Pythagorean theorem:

$$\boxed{\|\mathbf{u} \pm \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2}$$

- a point Q in a subspace W is closest to P $\iff$ $\overrightarrow{PQ} \perp W$

Pythagorean theorem and shortest distance

Useful identity:

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

Parallelogram rule:

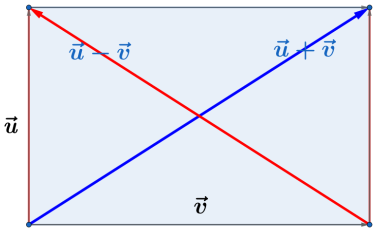
$$\|\mathbf{u} + \mathbf{v}\|^2 + \|\mathbf{u} - \mathbf{v}\|^2 = 2\|\mathbf{u}\|^2 + 2\|\mathbf{v}\|^2$$

Pythagorean theorem

u and **v** are orthogonal

$$\Updownarrow$$

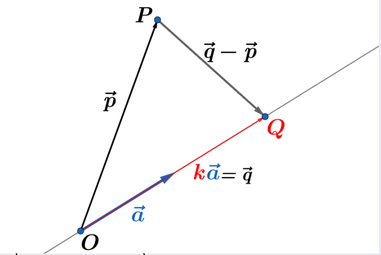
$$\boxed{\|\mathbf{u} \pm \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2}$$



Shortest distance to a line

ℓ is a line in $\mathbb{R}^n$ through O in direction of $\mathbf{a}$; P is a point in $\mathbb{R}^n$ outside ℓ

Problem: Find a point Q on ℓ that is closest to P.



Solution: Set $\mathbf{p} := \overrightarrow{OP}$, $\mathbf{q} := \overrightarrow{OQ} = k\mathbf{a}$. The optimal $\hat{k}$ minimizes

$$|PQ|^2 = \|k\mathbf{a}-\mathbf{p}\|^2 = k^2\|\mathbf{a}\|^2 - 2k\langle \mathbf{p},\mathbf{a}\rangle + \|\mathbf{p}\|^2 \quad\Longrightarrow\quad \hat{k} = \langle \mathbf{p},\mathbf{a}\rangle / \|\mathbf{a}\|^2$$

Observe that $\overrightarrow{PQ} = \mathbf{q}-\mathbf{p}$ and $\mathbf{a}$ are then orthogonal:

$$\langle \hat{k}\mathbf{a}-\mathbf{p},\mathbf{a}\rangle = \hat{k}\langle \mathbf{a},\mathbf{a}\rangle - \langle \mathbf{p},\mathbf{a}\rangle = 0$$

Shortest distance and orthogonality

Orthogonal PQ is the shortest one!

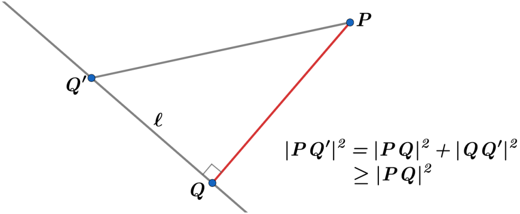
ℓ is a line in $\mathbb{R}^n$ in direction of **a**; P is a point in $\mathbb{R}^n$ outside ℓ

Claim: If Q on ℓ is s.t. $\mathbf{u} := \overrightarrow{PQ} \perp \mathbf{a}$, then $|PQ|$ is the smallest one

Reason: for any other point P' on ℓ, we have

$$|PQ'|^2 = \|\overrightarrow{PQ'}\|^2 = \|\overrightarrow{PQ} + \overrightarrow{QQ'}\|^2 = \|\overrightarrow{PQ}\|^2 + \|\overrightarrow{QQ'}\|^2 \geq \|\overrightarrow{PQ}\|^2$$

and the inequality is strict unless $Q = Q'$



Conclusion:

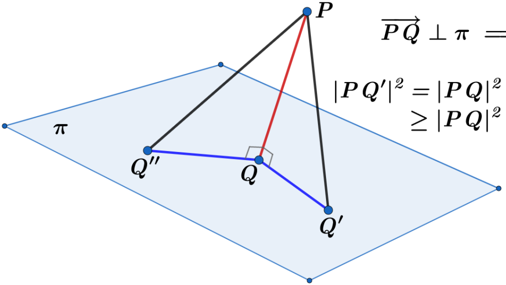
$$\boxed{Q \in \ell \text{ minimizes } |PQ| \iff \overrightarrow{PQ} \perp \ell}$$

Shortest distance to a plane

Remark

The same arguments work if instead of a line ℓ we take a plane π:

$$\boxed{Q \in \pi \text{ minimizes } |PQ| \iff \overrightarrow{PQ} \perp \pi}$$



$$\overrightarrow{PQ} \perp \pi \implies$$

$$|PQ'|^2 = |PQ|^2 + |QQ'^2| \geq |PQ|^2$$

Remark

In fact, instead of line ℓ or plane π we can take any subspace W in $\mathbb{R}^n$

Orthogonal decomposition

Theorem (Orthogonal decomposition)

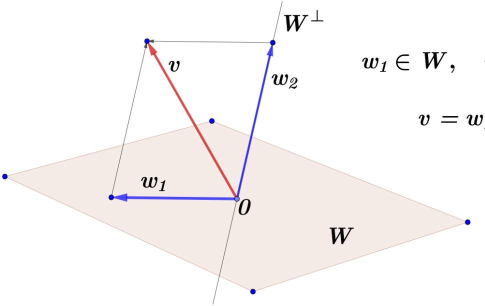
Assume W is a subspace of a vector space V with scalar product. Then $\forall\ \mathbf{v} \in V$: $\exists$ unique $\mathbf{w}_1 \in W$ and $\mathbf{w}_2 \in W^{\perp}$ s.t. $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$

Proof.

Existence: follows from the fact that $V = W \oplus W^{\perp}$

Uniqueness: if $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2 = \mathbf{u}_1 + \mathbf{u}_2$, then $\mathbf{w}_1 - \mathbf{u}_1 = \mathbf{u}_2 - \mathbf{w}_2$

$$\implies \mathbf{w}_1 - \mathbf{u}_1 = \mathbf{u}_2 - \mathbf{w}_2 = \mathbf{0}$$ □



$w_1 \in W, \quad w_2 \in W^{\perp}$

$v = w_1 + w_2$

Projectors

Definition (Projections and projectors)

Let $V = W \oplus W^\perp$ and $\mathbf{v} = \mathbf{w}_1 + \mathbf{w}_2$ with $\mathbf{w}_1 \in W$ and $\mathbf{w}_2 \in W^\perp$. Then

$\mathbf{w}_1$ is the orthogonal projection of $\mathbf{v}$ on W

$\mathbf{w}_2$ is the orthogonal projection of $\mathbf{v}$ on $W^\perp$

$P_j : \mathbf{v} \mapsto \mathbf{w}_j$ is the orthogonal projector onto W $(j = 1)$ or $W^\perp$ $(j = 2)$

Properties of P_j

- $P_j(a\mathbf{u} + b\mathbf{v}) = aP_j(\mathbf{u}) + bP_j(\mathbf{v})$ (linearity)
- $P_j^2 = P_j$ (idempotent)
- $P_1 P_2 = P_2 P_1 = 0$ (orthogonality)
- $P_1 + P_2 = I$ (completeness)

Extremal properties of projections:

$w_1 = P_1\mathbf{v}$ is the vector in W that is closest to $\mathbf{v}$;

$w_2 = P_2\mathbf{v}$ is the vector in $W^\perp$ that is closest to $\mathbf{v}$

Projection onto a line

Problem: Find the projection of $\mathbf{b} \in \mathbb{R}^m$ onto the line in direction $\mathbf{a} \in \mathbb{R}^m$

- The projection of $\mathbf{b}$ is a vector $\hat{x}\mathbf{a}$ orthogonal to error $\mathbf{e} := \mathbf{b} - \hat{x}\mathbf{a}$:

$$\mathbf{a}^T\mathbf{e} = \mathbf{a}^T(\mathbf{b} - \hat{x}\mathbf{a}) = 0 \quad \Longleftrightarrow \quad \hat{x} = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}$$

- the projection $\mathbf{p} := \hat{x}\mathbf{a} = \frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}}\mathbf{a}$ is a vector on the line closest to $\mathbf{b}$:

$$\begin{aligned}\|\mathbf{b} - t\mathbf{a}\|^2 &= \|\mathbf{b} - \mathbf{p} + \mathbf{p} - t\mathbf{a}\|^2 = \|\mathbf{b} - \mathbf{p}\|^2 + \|\mathbf{p} - t\mathbf{a}\|^2 \\ &= \|\mathbf{b} - \mathbf{p}\|^2 + (\hat{x} - t)^2\|\mathbf{a}\|^2 \geq \|\mathbf{b} - \mathbf{p}\|^2\end{aligned}$$

- Projection matrix P:

$$\mathbf{p} = \mathbf{a}\frac{\mathbf{a}^T\mathbf{b}}{\mathbf{a}^T\mathbf{a}} = P\mathbf{b} \implies \boxed{P = \frac{\mathbf{a}\mathbf{a}^T}{\mathbf{a}^T\mathbf{a}}}$$

Example: $\mathbf{b} = (1,1,1)^T$, $\mathbf{a} = (2,1,2)^T \implies$

$$\hat{x} = \frac{5}{9}, \quad \mathbf{e} = \begin{pmatrix} -1/9 \\ 4/9 \\ -1/9 \end{pmatrix} \perp \mathbf{a}, \quad P = \frac{1}{9}\mathbf{a}\mathbf{a}^T = \frac{1}{9}\begin{pmatrix} 4 & 2 & 4 \\ 2 & 1 & 2 \\ 4 & 2 & 4 \end{pmatrix}$$

Standard error, covariance, and correlation via LA

- In statistics, the sample mean of a data vector $\mathbf{x} = (x_1, \ldots, x_n)$ is

$$\bar{\mathbf{x}} := \frac{1}{n} \sum x_k = \frac{\langle \mathbf{x}, \mathbf{1} \rangle}{\langle \mathbf{1}, \mathbf{1} \rangle}$$

- $\bar{\mathbf{x}}\mathbf{1}$ is the orth. projection of $\mathbf{x}$ onto the constant space $V = \mathrm{ls}\{\mathbf{1}\}$
- the length $\|\mathbf{x} - \bar{\mathbf{x}}\mathbf{1}\|$ of $P_{V^\perp}\mathbf{x} = \mathbf{x} - \bar{\mathbf{x}}\mathbf{1}$ is the standard error $\mathrm{sd}(\mathbf{x})$
- clearly, $\mathbf{x} = \bar{\mathbf{x}}\mathbf{1} + P_{V^\perp}\mathbf{x}$ is the decomposition of $\mathbf{x}$ w.r.t. $V \oplus V^\perp$
- on $V^\perp$, $\langle \mathbf{x}, \mathbf{y} \rangle = \sum (x_k - \bar{\mathbf{x}})(y_k - \bar{\mathbf{y}})$ is the covariance
- the correlation,

$$\mathrm{cor}(\mathbf{x}, \mathbf{y}) := \frac{\sum (x_k - \bar{\mathbf{x}})(y_k - \bar{\mathbf{y}})}{\mathrm{sd}(\mathbf{x})\, \mathrm{sd}(\mathbf{y})}$$

  is the cosine between the vectors $P_{V^\perp}\mathbf{x}$ and $P_{V^\perp}\mathbf{y}$
- $|\mathrm{cor}(\mathbf{x}, \mathbf{y})| = 1 \iff$ the vectors $P_{V^\perp}\mathbf{x}$ and $P_{V^\perp}\mathbf{y}$ are collinear

Projection onto a subspace

Take $\mathbf{b} \in \mathbb{R}^m$ and assume $\mathbf{a}_1, \ldots, \mathbf{a}_n$ in $\mathbb{R}^m$ are linearly independent

Problem

Find a projection $\mathbf{p} = \hat{x}_1\mathbf{a}_1 + \cdots \hat{x}_n\mathbf{a}_n$ *of* $\mathbf{b}$ *onto* $W = \text{ls}\{\mathbf{a}_1, \ldots, \mathbf{a}_n\}$

- $\mathbf{p} = A\hat{\mathbf{x}}$, where A has columns $\mathbf{a}_1, \ldots, \mathbf{a}_n$
- the error $\mathbf{e} = \mathbf{b} - A\hat{\mathbf{x}}$ must be orthogonal to the $\mathcal{C}(A) = W$
- orthogonality condition: $\mathbf{a}_j^T\mathbf{e} = 0 \iff A^T(\mathbf{b} - A\hat{\mathbf{x}}) = \mathbf{0}$
- thus: $\boxed{A^TA\hat{\mathbf{x}} = A^T\mathbf{b}}$ gives $\hat{\mathbf{x}} \in \mathbb{R}^n$ s.t. $A\hat{\mathbf{x}}$ is closest to $\mathbf{b}$
- A^TA is nonsingular: if $A^TA\mathbf{y} = 0$, then
  $$\langle A^TA\mathbf{y}, \mathbf{y}\rangle = \|A\mathbf{y}\|^2 = 0 \implies A\mathbf{y} = \mathbf{0} \implies \mathbf{y} = \mathbf{0}$$
- $\mathbf{a}_j$-coordinates: $\hat{\mathbf{x}} = (A^TA)^{-1}A^T\mathbf{b}$
- projection: $\mathbf{p} = A(A^TA)^{-1}A^T\mathbf{b}$
- projection matrix: $P = A(A^TA)^{-1}A^T$

Approximate solution to a linear system $A\mathbf{x} = \mathbf{b}$

- If $\mathbf{b} \notin \mathcal{C}(A)$, then the system $A\mathbf{x} = \mathbf{b}$ is not soluble
- this is a typical situation for overdetermined systems ("tall" matrices A, i.e., more equations than variables)
- can ask for $\mathbf{x} = \mathbf{x}_0$ minimizing the error $\mathbf{e} := \mathbf{b} - A\mathbf{x}$
- as we know, the shortest $\mathbf{e}_0 := \mathbf{b} - A\mathbf{x}_0$ is orthogonal to $\mathcal{C}(A)$
- $\mathbf{b} = \mathbf{e}_0 + A\mathbf{x}_0$ is an orthogonal decomposition of $\mathbf{b}$ wrt
$$\mathbb{R}^m = \mathcal{N}(A^\top) \oplus \mathcal{C}(A):$$
    - $A\mathbf{x}_0$ is the component of $\mathbf{b}$ in $\mathcal{C}(A)$
    - $\mathbf{e}_0$ is orthogonal to $\mathcal{C}(A) \implies$ belongs to the left null-space $\mathcal{N}(A^\top)$
- therefore, $A^\top \mathbf{e}_0 = \mathbf{0}$ so that $A^\top A\mathbf{x}_0 = A^\top \mathbf{b}$
- assume columns of A are linearly independent; then $A^\top A$ is invertible (already know this!)
- we thus get the same results but in different interpretation:
    - best solution: $\mathbf{x}_0 = (A^\top A)^{-1} A^\top \mathbf{b}$;
    - projection of $\mathbf{b}$: $P\mathbf{b} = A\mathbf{x}_0 = A(A^\top A)^{-1} A^\top \mathbf{b}$
    - projection matrix: $P = A(A^\top A)^{-1} A^\top$

Least squares solution to $A\mathbf{x} = \mathbf{b}$

- As was shown, solving $\boxed{A^T A\hat{\mathbf{x}} = A^T\mathbf{b}}$ gives the projection $\mathbf{p} = A\hat{\mathbf{x}}$ of $\mathbf{b}$ onto the column space of A
- if $\mathbf{b} \notin \mathcal{C}(A)$, this is the least squares solution of $A\mathbf{x} \approx \mathbf{b}$ with the smallest error $\|\mathbf{b} - A\hat{\mathbf{x}}\|^2$
- this solution can be obtained by minimizing $f(\mathbf{x}) := \|A\mathbf{x} - \mathbf{b}\|^2$
- indeed,

$$\begin{aligned}\|A\mathbf{x} - \mathbf{b}\|^2 &= \langle A\mathbf{x}, A\mathbf{x}\rangle - 2\langle A\mathbf{x}, \mathbf{b}\rangle + \|\mathbf{b}\|^2 \\ &= \langle A^T A\mathbf{x}, \mathbf{x}\rangle - 2\langle A^T\mathbf{b}, \mathbf{x}\rangle + \|\mathbf{b}\|^2\end{aligned}$$

so that (justify!)

$$\operatorname{grad}\|A\mathbf{x} - \mathbf{b}\|^2 = 2A^T A\mathbf{x} - 2A^T\mathbf{b}$$

and

$$\operatorname{grad}\|A\mathbf{x} - \mathbf{b}\|^2 = \mathbf{0} \iff A^T A\mathbf{x} = A^T\mathbf{b}$$

Linear regression

Example: fitting a straight line $b = C + Dt$ to m points

- $(0,6)$, $(1,0)$, $(2,0)$ do not lie on one line
- best fit: minimize the sum of squared errors $\sum(b_k - C - Dt_k)^2$
- minimize $\|A\mathbf{x} - \mathbf{b}\|^2$ for $A = \begin{pmatrix} 1 & 0 \\ 1 & 1 \\ 1 & 2 \end{pmatrix}$, $\mathbf{x} = \begin{pmatrix} C \\ D \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 6 \\ 0 \\ 0 \end{pmatrix}$
- that is the setting we get in linear regression models

Polynomial regression

Fitting a quadratic polynomial $b = C + Dt + Et^2$ to m points

- $(0,6)$, $(1,0)$, $(2,0)$, $(3,1)$ do not belong to one parabola
- best fit: minimize the squared error $\sum(b_k - C - Dt_k - Et_k^2)^2$
- minimize $\|A\mathbf{x} - \mathbf{b}\|^2$ for $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix}$, $\mathbf{x} = \begin{pmatrix} C \\ D \\ E \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 6 \\ 0 \\ 0 \\ 1 \end{pmatrix}$
- that is the setting of the polynomial regression

Multiple regression

Example: fitting a plane $b = C + Ds + Et$ to m points

- Data points (s, t, b): $(0, 0, 6)$, $(1, 1, 0)$, $(2, 4, 0)$, $(3, 9, 1)$;
- these points do not belong to one plane
- best fit: minimize the squared error
$$\sum (b_k - C - Ds - Et)^2$$
- minimize $\|A\mathbf{x} - \mathbf{b}\|^2$ for $A = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{pmatrix}$, $\mathbf{x} = \begin{pmatrix} C \\ D \\ E \end{pmatrix}$, $\mathbf{b} = \begin{pmatrix} 6 \\ 0 \\ 0 \\ 1 \end{pmatrix}$
- that is the setting of the multiple regression