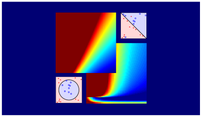
Machine Learning Foundations
(機器學習基石)



Lecture 9: Linear Regression, Extended

Hsuan-Tien Lin (林軒田)
htlin@csie.ntu.edu.tw

Department of Computer Science
& Information Engineering

National Taiwan University
(國立台灣大學資訊工程系)

The Hat Matrix

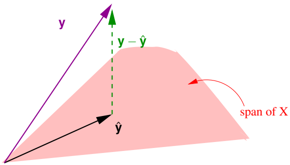
when X^TX invertible, hat matrix $H = XX^\dagger = X(X^TX)^{-1}X^T$

Claim: $H^{1126} = H$

proof (when X^TX invertible):

$$
\begin{aligned}
H^{1126} &= HHH^{1124} \\
&= X(X^TX)^{-1}X^TX(X^TX)^{-1}X^TH^{1124} \\
&= X(X^TX)^{-1}(X^TX)(X^TX)^{-1}X^TH^{1124} \\
&= X(X^TX)^{-1}X^TH^{1124} \\
&= H^{1125}
\end{aligned}
$$

... and you know the rest



geometrically, **projecting** 1126 **times** $\equiv$ projecting once

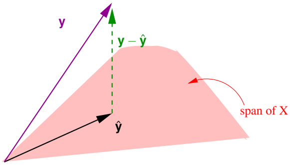
Trace of The Hat Matrix

when $\mathrm{X}^T\mathrm{X}$ invertible, hat matrix $\mathrm{H} = \mathrm{X}\mathrm{X}^\dagger = \mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T$

Claim: $\text{trace}(\mathrm{H}) = d + 1$ when $\mathrm{X}^T\mathrm{X}$ invertible

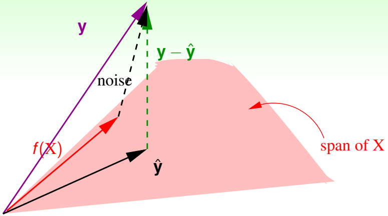
proof:

$$\begin{aligned}
\text{trace}(\mathrm{H}) &= \text{trace}(\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}\mathrm{X}^T) \\
&= \text{trace}(\mathrm{X}^T\mathrm{X}(\mathrm{X}^T\mathrm{X})^{-1}) \\
&= \text{trace}(\mathrm{I}_{d+1}) \\
&= d + 1
\end{aligned}$$



geometrically, H projects to
a $(d+1)$-dimensional subspace

An Illustrative 'Proof', Corrected



- if $\mathbf{y}$ comes from some ideal $f(\mathrm{X}) \in$ span plus **noise**
- **noise** with per-dimension 'noise level' σ^2 transformed by $\mathrm{I} - \mathrm{H}$ to be $\mathbf{y} - \hat{\mathbf{y}}$

$$E_{\text{in}}(\mathbf{w}_{\text{LIN}}) = \frac{1}{N}\|\mathbf{y} - \hat{\mathbf{y}}\|^2 = \frac{1}{N}\|(\mathrm{I} - \mathrm{H})\textbf{noise}\|^2 = \frac{1}{N}\big(N - (d+1)\big)\sigma^2$$

$$\overline{E_{\text{in}}} = \sigma^2 \cdot \left(1 - \frac{d+1}{N}\right)$$

$$\overline{E_{\text{out}}} = \sigma^2 \cdot \left(1 + \frac{d+1}{N}\right)$$ **(complicated!)**