ML & LinAlg Math Cheat Sheet

October 24, 2017

Contents

1 Notation

Vectors are column vectors denoted by lower-case bolded variables, such that

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}.$$

A row vector is denoted $\boldsymbol{x}^{\top} = [x_1 \ldots x_N]$. A matrix is indicated by a bolded upper-case variable, such that an $N \times M$ matrix is

$$\boldsymbol{A} = \{a_{ij}\} = [\boldsymbol{a}_1 \cdots \boldsymbol{a}_M] = \begin{bmatrix} \boldsymbol{a}_1^{\top} \\ \vdots \\ \boldsymbol{a}_N^{\top} \end{bmatrix} = \begin{bmatrix} a_{1,1}^{\top} & \cdots & a_{1,M} \\ \vdots & \ddots & \vdots \\ a_N^{\top} & \cdots & a_{N,M} \end{bmatrix}.$$

2 Derivative

2.a Vector Gradient

$$\nabla_{\boldsymbol{x}}\boldsymbol{y} = [\frac{\partial \boldsymbol{y}}{\partial x_1}, \ldots, \frac{\partial \boldsymbol{y}}{\partial x_N}] \tag{1}$$

3   Determinant Operator

3.a   Random Properties

For scalar c and $N \times N$ identity matrix I,

$$\det(cI) = c^N.$$

4   Trace Operator

4.a   Derivatives

4.a.i $\nabla_{\boldsymbol{x}} \operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{A}) = 2\boldsymbol{x}^\top \boldsymbol{A}$

For square $N \times N$ matrix $\boldsymbol{A}$. See appendix A.a.i for proof.

4.b   Relation to Determinant

A   Proofs

A.a   Trace

A.a.i $\nabla_{\boldsymbol{x}} \operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{A}) = 2\boldsymbol{x}^\top \boldsymbol{A}$

This proof can likely be generalized to non-square matrixes (and possibly some communicativeness, given the flexibility afforded by the trace), but the restricted case is presented here.

For square $N \times N$ matrix $\boldsymbol{A}$,

$$\nabla_{\boldsymbol{x}} \operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{A}) = \frac{d}{d\boldsymbol{x}} \operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{A}) = \frac{d}{d\boldsymbol{x}} \sum_i^N \sum_k^N x_i x_k a_{ik}.$$

Recall eq. (1), and consider for any $j \in \{1, \ldots, N\}$:

$$
\begin{aligned}
\frac{\partial}{\partial x_j} \sum_i^N \sum_k^N x_i x_k a_{ik} &= x_1 a_{j,1} + x_2 a_{j,2} + \cdots \frac{\partial}{\partial x_j} \sum_k^N x_j x_k a_{jk} + \cdots x_N a_{j,N} \\
&= \sum_i^N x_i a_{ji} - x_j a_{jj} + \sum_k^N x_k a_{jk} - x_j a_{jj} + \frac{\partial}{\partial x_j} x_j x_j a_{jj} \\
&= 2 \sum_i^N x_i a_{ji} - 2 x_j a_{jj} + 2 x_j a_{jj} \\
&= 2 \boldsymbol{x}^\top \boldsymbol{a}_j.
\end{aligned}
$$

This equally applies for any j in $1 \ldots N$, and so for the full gradient:

$$\nabla_{\boldsymbol{x}} \operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{A}) = \frac{d}{d\boldsymbol{x}} \sum_i^N \sum_k^N x_i x_k a_{ik} = [2\boldsymbol{x}^\top \boldsymbol{a}_1 \cdots 2\boldsymbol{x}^\top \boldsymbol{a}_N] = 2\boldsymbol{x}^\top \boldsymbol{A}.$$