ML & LinAlg Math Cheat Sheet

December 9, 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 $\boldsymbol{x}^\top = [x_1 \dots x_N]$. A matrix is indicated by a bolded upper-case variable, such that an $N \times M$ matrix is

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

For some random variable x, let $\mathbb{E}[x]$ denote its expected value.

2 Derivative

2.a Vector Gradient

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

3 Determinant Operator

3.a Determinant Properties

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

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

3.b Derivatives of Determinants

3.b.i $\nabla_{\boldsymbol{A}} \ln |\boldsymbol{A}| = \boldsymbol{A}^{-1}$

Proof TBD.

4 Trace Operator

Defined for $N \times N$ square matrix $\boldsymbol{A}$ as

$$\operatorname{tr}(\boldsymbol{A}) \overset{\text{def}}{=} \sum_i^N a_{ii} \tag{2}$$

4.a Trace Properties

4.a.i $\operatorname{tr}(c\boldsymbol{A} + d\boldsymbol{B}) = c \operatorname{tr}(\boldsymbol{A}) + d \operatorname{tr}(\boldsymbol{B})$

For scalars c and d, square matrices $\boldsymbol{A}$ and $\boldsymbol{B}$.

4.a.ii $\operatorname{tr}(\boldsymbol{AB}) = \operatorname{tr}(\boldsymbol{BA}) = \operatorname{tr}(\boldsymbol{A}^\top \boldsymbol{B}) = \operatorname{tr}(\boldsymbol{AB}^\top) = \sum_{i,j} a_{ij} b_{ij}$

And clearly, also $\operatorname{tr}(\boldsymbol{B}^\top \boldsymbol{A}) = \operatorname{tr}(\boldsymbol{BA}^\top) = \sum_{i,j} a_{ij} b_{ij} = \operatorname{tr}(\boldsymbol{AB})$.

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

This can be seen from

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} = \left[\begin{array}{ccc} \sum_{i=1}^{N} x_i a_{i,1} & \cdots & \sum_{i=1}^{N} x_i a_{i,N} \end{array}\right] \boldsymbol{x} = \sum_{j=1}^{N} x_j \sum_{i=1}^{N} x_i a_{i,j} = \sum_{j=1}^{N}\sum_{i=1}^{N} a_{i,j} (\boldsymbol{x}\boldsymbol{x}^\top)_{i,j} = \operatorname{tr}(\boldsymbol{A}\boldsymbol{x}\boldsymbol{x}^\top).$$

The last equality follows from section 4.a.ii.

4.b Trace Derivatives

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

For square matrix $\boldsymbol{A}$. Note that $\boldsymbol{x}^\top(\boldsymbol{A} + \boldsymbol{A}^\top) = 2\boldsymbol{x}^\top \boldsymbol{A}$ for symmetric $\boldsymbol{A}$.

See appendix A.a.i for proof.

4.b.ii $\nabla_{\boldsymbol{A}} \operatorname{tr}(\boldsymbol{x}\boldsymbol{x}^\top \boldsymbol{A}) = \boldsymbol{x}\boldsymbol{x}^\top$

Proof left as an exercise for the reader.

4.c Trace Relation to Determinant

TBD

5 Expected Values

For $\boldsymbol{x} \in \mathbb{R}^d$, with expected value $\mathbb{E}[\boldsymbol{x}] = \boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma} = \mathbb{E}[(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])(\boldsymbol{x} - \mathbb{E}[\boldsymbol{x}])^\top]$,

$$\mathbb{E}[x_i^2] = \Sigma_{i,i} + \mu_i^2 \tag{3}$$

$$\mathbb{E}[\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x}] = \operatorname{tr}(\boldsymbol{A}\boldsymbol{\Sigma}) + \mu^\top \boldsymbol{A}\mu \tag{4}$$

$$\mathbb{E}_x\left[(y - \boldsymbol{x}\top\boldsymbol{w})^2\right] = (y - \boldsymbol{w}^\top\boldsymbol{\mu})^2 + \boldsymbol{w}^\top\boldsymbol{\Sigma}\boldsymbol{w}. \tag{5}$$

A Proofs

A.a Trace

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

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 $j \in \{1, \ldots, N\}$:

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

where $[\boldsymbol{a}^\top]_j$ is the jth column of $A^\top$.

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

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