Math for ML

Aniruddha Deb

Contents

1 Linear Algebra Basics **2**

2 Differential Forms **2**

3 Probability Basics **2**

3.1 Multivariate Basics 2

3.2 Multivariate Gaussian Distribution 3

3.3 Bayesian Probability 3

4 Information Theory **3**

5 Linear Regression **4**

6 References **4**

1 Linear Algebra Basics

Definition 1

The **Spectrum** of a matrix is the set of it's eigenvalues.

Definition 2

A symmetric matrix A is **Positive Semi-Definite** if for all vectors $\boldsymbol{z} \in \mathbb{R}^n$, we have $\boldsymbol{z}^T A \boldsymbol{z} \geq 0$ and **Positive Definite** if the inequality is strict.

Theorem 1

The following are equivalent (A is a symmetric matrix):

1. A is positive semidefinite
2. All the eigenvalues of A are positive
3. There exists a matrix B such that $B^T B = A$

Theorem 2

The inverse of a positive semidefinite matrix is positive semidefinite, and the eigenvalues of the inverse are inverses of the eigenvalues (eigenvectors remaining the same).

Proof. if A is PSD, then by spectral decomposition, $A = P^{-1}DP$. Therefore, $A^{-1} = P^{-1}D^{-1}P$, and D^{-1} being a diagonal matrix will simply have the inverses of the eigenvalues along it's diagonal. Hence, the eigenvalues of inverse are inverses of eigenvalues, and A^-1 is PSD as it's eigenvalues are positive. $\square$

Definition 3

The **Singular Value Decomposition** of a $m \times n$ matrix M is given by

$$M = U\Sigma V^T \tag{1}$$

where U is a $m \times m$ orthogonal matrix, Σ is a $m \times n$ diagonal matrix, and V is a $n \times n$ orthogonal matrix

A good illustration of SVD is given at Wikipedia, with some nice gifs

2 Differential Forms

Most of these are taken from the Matrix Cookbook.

$$\frac{\partial \mathbf{x}^T \mathbf{a}}{\partial \mathbf{x}} = \frac{\partial \mathbf{a}^T \mathbf{x}}{\partial \mathbf{x}} = \mathbf{a} \tag{2}$$

$$\frac{\partial \mathbf{a}^T \mathbf{X} \mathbf{b}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T \tag{3}$$

3 Probability Basics

3.1 Multivariate Basics

For a random vector $\mathbf{x}$,

$$\boldsymbol{\mu} = E(\boldsymbol{x}) = [\int x_i f_i(x_i) dx]_i \tag{4}$$

$$\boldsymbol{\Sigma} = E((\boldsymbol{x} - \boldsymbol{\mu})(\boldsymbol{x} - \boldsymbol{\mu})^T) \tag{5}$$

$$= E(\boldsymbol{x}\boldsymbol{x}^T) - \boldsymbol{\mu}\boldsymbol{\mu}^T \qquad = \tag{6}$$

where $\boldsymbol{\mu}$ is the mean and $\boldsymbol{\Sigma}$ is the covariance matrix

Theorem 3

The covariance matrix is always positive semi-definite

Proof. For a matrix A to be positive semi-definite, for every vector $\boldsymbol{z} \in \mathbb{R}^n$, $\boldsymbol{z}^T \boldsymbol{A} \boldsymbol{z} \geq 0$
Substituting this into the covariance matrix, for every $\boldsymbol{z} \in \mathbb{R}^n$, we have

$$\begin{aligned}
\boldsymbol{z}^T \boldsymbol{\Sigma} \boldsymbol{z} &= \boldsymbol{z}^T E(\boldsymbol{x}\boldsymbol{x}^T)\boldsymbol{z} - (\boldsymbol{z}^T \boldsymbol{\mu})^2 \\
&= E(\boldsymbol{z}^T \boldsymbol{x}\boldsymbol{x}^T \boldsymbol{z}) - (\boldsymbol{z}^T \boldsymbol{\mu})^2 \\
&= E((\boldsymbol{z}^T \boldsymbol{x})^2) - (\boldsymbol{z}^T \boldsymbol{\mu})^2 \\
&= \mathrm{Var}(\boldsymbol{z}^T \boldsymbol{x}) \\
&\geq 0
\end{aligned}$$

□

Definition 4

The **Mahalanobis Distance** of a point $\boldsymbol{x}$ from a probability density Q on $\boldsymbol{R}^n$ is

$$d_m(\mathbf{x}, Q) = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^- 1(\mathbf{x} - \boldsymbol{\mu})} \tag{7}$$

The mahalanobis distance can be thought of as a generalization of the Z-score $\frac{x-\mu}{\sigma}$: it gives a measure of how many standard deviations away an observation is from the mean of the distribution for a multivariate distribution.

3.2 Multivariate Gaussian Distribution

$$f(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1}(\mathbf{x} - \boldsymbol{\mu})\right) \tag{8}$$

Theorem 4

The Multivariate Gaussian is the distribution with maximum entropy subject to having a specified mean and covariance

3.3 Bayesian Probability

4 Information Theory

Definition 5

The **Information** of an event is defined as the negative logarithm of the probability of that event

$$I(x) = -log_b(P(x)) \tag{9}$$

The information of an event conveys how 'surprising' that event is. $I(x)$ describes the information of a single event, but $I(X)$ is a random variable.

Note that b defines the units of information: if $b = 2$, the units are called *bits* or *shannons*, if $b = 10$ they're called *hartleys* and if $b = e$ they're called *nats*.

Definition 6

The **Entropy** of a random variable X is the expected information of X

$$H(X) = E(I(X)) = E(-\log(X)) \tag{10}$$

These definitions apply simply to discrete variables, but can also be extended in a measure-theoretic sense (see Wikipedia) for continuous random variables. In a computational sense, if we need to compute entropy for a continuous RV, it's done by binning the RV to make a discrete RV and then performing computations on the discrete RV thus obtained.

Definition 7

The **Conditional Entropy** of one random variable with respect to another is defined as

$$H(Y|X) = H(H(Y|X=x)) = \sum_x p(x) H(Y|X=x) \tag{11}$$
$$= -\sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)} \right)$$

Definition 8

The **Kullback-Leibler divergence** is a measure of how much information is needed to discriminate between two distributions (ie whether an observation comes from distribution A or distribution B)

$$D_{kl}(A \parallel B) = \sum_x P(x) \log \left(\frac{P(x)}{Q(x)} \right) \tag{12}$$

5 Linear Regression

6 References

Bishop, Murphy, Matrix Cookbook