Lecture 2.1: Vector Calculus
CSC 84020 - Machine Learning

Andrew Rosenberg

February 5, 2009

Today

- Last Time
  - Probability Review
- Today
  - Vector Calculus

Background

Let's talk.

- Linear Algebra
  - Vectors
  - Matrices
  - Basis Spaces
  - Eigenvectors/values?
  - Inversion and transposition
- Calculus
  - Derivation
  - Integration
- Vector Calculus
  - Gradients
  - Derivation w.r.t. a vector

Linear Algebra Basics

- What is a vector?
- What is a matrix?

- Transposition
- Adding matrices and vectors
- Multiplying matrices.

Definitions

A vector is a one dimensional array.
We denote vectors as either **x**, **x**.
If we don't specify otherwise assume **x** is a *column vector*.

$$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ \dots \\ x_{n-1} \end{pmatrix}$$

Definitions

A matrix is a higher dimensional array.
We typically denote matrices as capital letters e.g., A.
If A is an n-by-m matrix, it has the following structure

$$A = \begin{pmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,m-1} \\ a_{1,0} & a_{1,1} & & a_{1,m-1} \\ \vdots & & \ddots & \vdots \\ a_{n-1,0} & a_{n-1,1} & \cdots & a_{n-1,m-1} \end{pmatrix}$$

Matrix transposition

Transposing a matrix or vector swaps rows and columns.

A column-vector becomes a row-vector

$$\mathbf{x} = \begin{pmatrix} x_0 \\ x_1 \\ \dots \\ x_{n-1} \end{pmatrix}$$

$$\mathbf{x}^T = \begin{pmatrix} x_0 & x_1 & \dots & x_{n-1} \end{pmatrix}$$

Matrix transposition

Transposing a matrix or vector swaps rows and columns.

A column-vector becomes a row-vector

$$A = \begin{pmatrix} a_{0,0} & a_{0,1} & \cdots & a_{0,m-1} \\ a_{1,0} & a_{1,1} & & a_{1,m-1} \\ \vdots & & \ddots & \vdots \\ a_{n-1,0} & a_{n-1,1} & \cdots & a_{n-1,m-1} \end{pmatrix}$$

$$A^T = \begin{pmatrix} a_{0,0} & a_{1,0} & \cdots & a_{n-1,0} \\ a_{0,1} & a_{1,1} & & a_{1,m-1} \\ \vdots & & \ddots & \vdots \\ a_{0,m-1} & a_{1,m-1} & \cdots & a_{n-1,m-1} \end{pmatrix}$$

If A is n-by-m, then A^T is m-by-n.

Adding Matrices

Matrices can only be added if they have the same dimension.

$$A+B = \begin{pmatrix} a_{0,0}+b_{0,0} & a_{0,1}+b_{0,1} & \dots & a_{0,m-1}+b_{0,m-1} \\ a_{1,0}+b_{1,0} & a_{1,1}+b_{1,1} & & a_{1,m-1}+b_{1,m-1} \\ \vdots & & \ddots & \vdots \\ a_{n-1,0}+b_{n-1,0} & a_{n-1,1}+b_{n-1,1} & \dots & a_{n-1,m-1}+b_{n-1,m-1} \end{pmatrix}$$

Multiplying matrices

To multiply two matrices, the *inner dimensions* must match.

- An n-by-m can be multiplied by an n'-by-m' matrix iff $m = n'$.

$$AB = C$$

$$c_{ij} = \sum_{k=0}^{m} a_{ik} * b_{kj}$$

That is, multiply the i-th row by the j-th column.

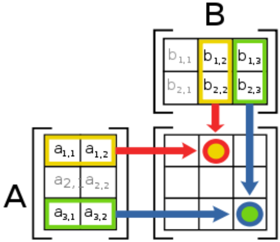


Image from wikipedia.

Useful matrix operations

- Inversion
- Norm
- Eigenvector decomposition

Matrix Inversion

The inverse of an n-by-m matrix A is denoted A^{-1}, and has the following property.

$$AA^{-1} = I$$

Where I is the **identity matrix**, an n-by-n matrix where $I_{ij} = 1$ iff $i = j$ and 0 otherwise.
If A is a **square** matrix (iff $n = m$) then,

$$A^{-1}A = I$$

Matrix Inversion

The inverse of an n-by-m matrix A is denoted A^{-1}, and has the following property.

$$AA^{-1} = I$$

Where I is the **identity matrix**, an n-by-n matrix where $I_{ij} = 1$ iff $i = j$ and 0 otherwise.
If A is a **square** matrix (iff $n = m$) then,

$$A^{-1}A = I$$

What is the inverse of a vector? $\mathbf{x}^{-1} = ?$

Some useful Matrix Inversion Properties

$$(A^{-1})^{-1} = A$$

$$(kA)^{-1} = k^{-1}A^{-1}$$

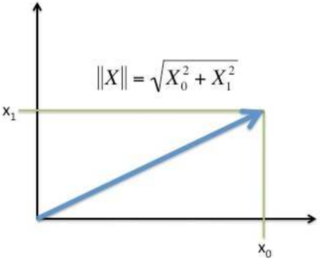
$$(A^T)^{-1} = (A^{-1})^T$$

$$(AB)^{-1} = B^{-1}A^{-1}$$

The norm of a vector

The **norm** of a vector $\mathbf{x}$ is written $||\mathbf{x}||$.
The norm represents the euclidean length of a vector.

$$
\begin{aligned}
||\mathbf{x}|| &= \sqrt{\sum_{i=0}^{n-1} x_i^2} \\
&= \sqrt{x_0^2 + x_1^2 + \ldots + x_{n-1}^2}
\end{aligned}
$$



Positive Definite/Semi-Definite

A **positive definite** matrix, M has the property that

$$x^T M x > 0$$

A **positive semi-definite** matrix, M has the property that

$$x^T M x \geq 0$$

Why might we care about these matrices?

Eigenvectors

For a square matrix A, the eigenvector is defined as

$$A\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

Where $\mathbf{u}_i$ is an **eigenvector** and λ_i is its corresponding **eigenvalue**.

In general, eigenvalues are complex numbers, but if A is symmetric, they are real.

Eigenvectors describe how a matrix transforms a vector, and can be used to define a basis space, namely the eigenspace.

Who cares? The eigenvectors of a covariance matrix have some very interesting properties.

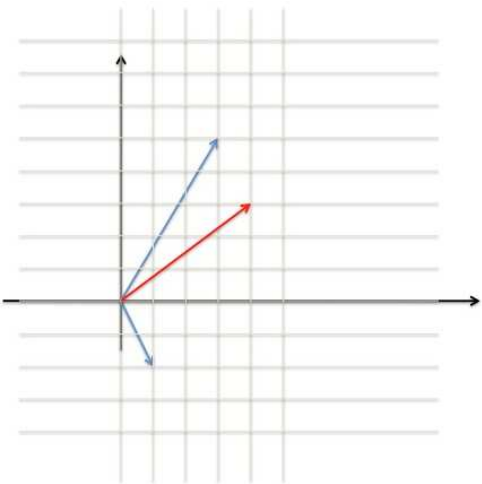
Basis Spaces

Basis spaces allow vectors to be represented in different spaces. Our normal 2-dimensional basis space is generated by the vectors $[0, 1]$, $[1, 0]$.

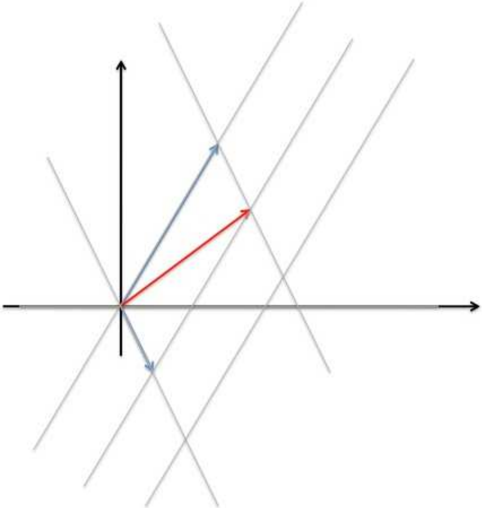
- Any 2-d vector can be expressed as the sum of linear factors of these two **basis vectors**.

However, any two non-colinear vectors can **generate** a 2-d basis space. In this basis space, the generating vectors are perpendicular.

Basis Spaces



Basis Spaces



Basis Spaces

Why do we care?

Dimensionality reduction.

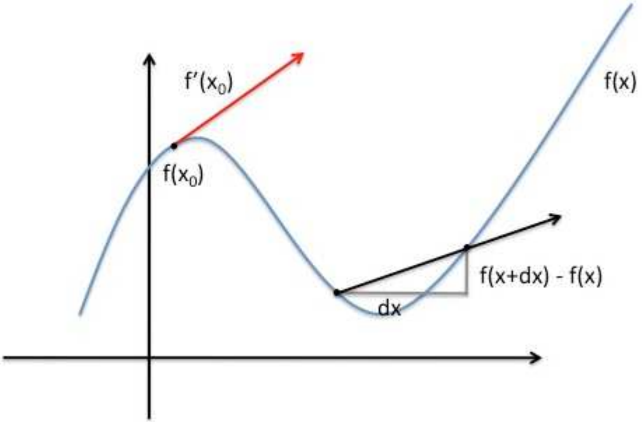
Calculus Basics

- What is a derivative?
- What is an integral?

Derivatives

A **derivative**, $\frac{d}{dx}f(x)$ can be thought of as defining the **slope** of a function $f(x)$. This is sometimes also written as $f'(x)$.

Derivative Example



Integrals

Integrals are an inverse operation of the derivative (plus a constant).

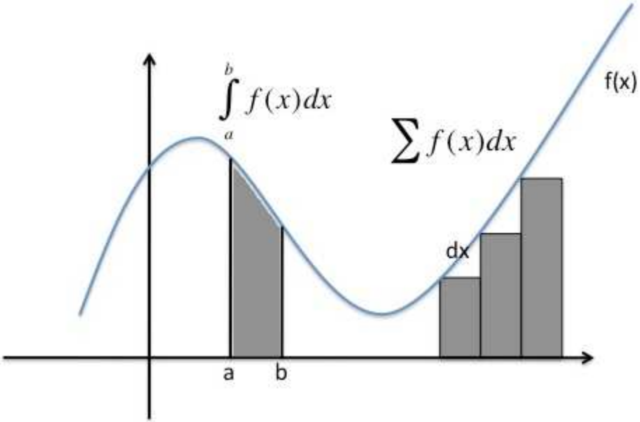
$$\int f(x)dx = F(x) + c$$

$$F'(x) = f(x)$$

An integral can be thought of as a calculation of the area under the curve defined by $f(x)$.

A **definite** integral evaluates the area over a finite region. An **indefinite** integral is calculated over the range of $(-\infty, \infty)$.

Integration Example



Useful calculus operations

- Product, quotient, summation rules for derivatives.
- Useful integration and derivative identities.
- Chain rule
- Integration by parts
- Variable substitution (don't forget the Jacobian!)

Calculus Identities

Summation rule

$$g(x) = f_0(x) + f_0(x)$$

$$g'(x) = f_0'(x) + f_1'(x)$$

Product Rule

$$g(x) = f_0(x)f_1(x)$$

$$g'(x) = f_0(x)f_1'(x) + f_0'(x)f_1(x)$$

Quotient Rule

$$g(x) = \frac{f_0(x)}{f_1(x)}$$

$$g'(x) = \frac{f_0(x)f_1'(x) - f_0'(x)f_1(x)}{f_1^2(x)}$$

Calculus Identities

Constant multipliers

$$g(x) = cf(x)$$
$$g'(x) = cf'(x)$$

Exponent Rule

$$g(x) = f(x)^k$$
$$g'(x) = kf(x)^{k-1}$$

Chain Rule

$$g(x) = f_0(f_1(x))$$
$$g'(x) = f_0'(f_1(x))f_1'(x)$$

Calculus Identities

Exponent Rule

$$g(x) = e^x$$

$$g'(x) = e^x$$

$$g(x) = k^x$$

$$g'(x) = \ln(k)k^x$$

Logarithm Rule

$$g(x) = \ln(x)$$

$$g'(x) = \frac{1}{x}$$

$$g(x) = \log_b(x)$$

$$g'(x) = \frac{1}{x \ln b}$$

Calculus Operations

Integration by Parts

$$\int f(x)\frac{dg(x)}{dx}dx = f(x)g(x) - \int g(x)\frac{df(x)}{dx}dx$$

Variable Substitution

$$\int_a^b f(g(x))g'(x)dx = \int_{g(a)}^{g(b)} f(x)dx$$

Vector Calculus

- Derivation with respect to to a vector or matrix.
- Gradient of a vector.
- Change of variables with a vector.

Derivation with respect to a vector

Given a vector $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})^T$, and a function $f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ how can we find $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$?

Derivation with respect to a vector

Given a vector $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})^T$, and a function $f(\mathbf{x}) : \mathbb{R}^n \to \mathbb{R}$ how can we find $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$?

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_0} \\ \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_{n-1}} \end{pmatrix}$$

This is also called the **gradient** of the function, and is often written $\nabla f(x)$ or ∇f.

Derivation with respect to a vector

Given a vector $\mathbf{x} = (x_0, x_1, \ldots, x_{n-1})^T$, and a function $f(\mathbf{x}) : \mathbb{R}^n \rightarrow \mathbb{R}$ how can we find $\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}}$?

$$\frac{\partial f(\mathbf{x})}{\partial \mathbf{x}} = \begin{pmatrix} \frac{\partial f(\mathbf{x})}{\partial x_0} \\ \frac{\partial f(\mathbf{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f(\mathbf{x})}{\partial x_{n-1}} \end{pmatrix}$$

This is also called the **gradient** of the function, and is often written $\nabla f(x)$ or ∇f.

Why might this be useful?

Useful Vector Calculus identities

Given a vector $\mathbf{x}$ with $|\mathbf{x}| = n$ and a scalar variable y.

$$\frac{\partial \mathbf{x}}{\partial y} = \begin{pmatrix} \frac{\partial \mathbf{x}_0}{\partial y} \\ \frac{\partial \mathbf{x}_1}{\partial y} \\ \vdots \\ \frac{\partial \mathbf{x}_{n-1}}{\partial y} \end{pmatrix}$$

Useful Vector Calculus identities

Given a vector $\mathbf{x}$ with $|\mathbf{x}| = n$ and a vector $\mathbf{y}$ with $|\mathbf{y}| = m$.

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial \mathbf{x}_0}{\partial \mathbf{y}_0} & \frac{\partial \mathbf{x}_0}{\partial \mathbf{y}_1} & \cdots & \frac{\partial \mathbf{x}_0}{\partial \mathbf{y}_{m-1}} \\ \frac{\partial \mathbf{x}_1}{\partial \mathbf{y}_0} & \frac{\partial \mathbf{x}_1}{\partial \mathbf{y}_1} & \cdots & \frac{\partial \mathbf{x}_1}{\partial \mathbf{y}_{m-1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{y}_0} & \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{y}_1} & \cdots & \frac{\partial \mathbf{x}_{n-1}}{\partial \mathbf{y}_{m-1}} \end{pmatrix}$$

Vector Calculus Identities

Similar to – Scalar Multiplication Rule

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^T\mathbf{a}) = \frac{\partial}{\partial \mathbf{x}}(\mathbf{a}^T\mathbf{x}) = \mathbf{a}$$

Similar to – Product Rule

$$\frac{\partial}{\partial x}(AB) = \frac{\partial A}{\partial x}B + A\frac{\partial B}{\partial x}$$

Derivative of an Matrix inverse.

$$\frac{\partial}{\partial x}(A^{-1}) = -A^{-1}\frac{\partial A}{\partial x}A^{-1}$$

Change of Variable in an Integral

$$\int f(\mathbf{x})d\mathbf{x} = \int f(\mathbf{u})\left|\frac{\partial \mathbf{x}}{\partial \mathbf{u}}\right| d\mathbf{u}$$

Calculating the Expectation of a Gaussian

Now we have enough tools to calculate the expectation of a variable given a Gaussian Distribution.

Recall:

$$
\begin{aligned}
\mathbb{E}[x|\mu,\sigma^2] &= \int p(x|\mu,\sigma^2)x dx \\
&= \int N(x|\mu,\sigma^2)x dx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} x dx
\end{aligned}
$$

Calculating the Expectation of a Gaussian

$$\mathbb{E}[x|\mu,\sigma^2] = \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} xdx$$

$$u = x - \mu$$
$$du = dx$$

$$\begin{aligned}
\mathbb{E}[x|\mu,\sigma^2] &= \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\} xdx \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} (u+\mu)du \\
&= \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} udu + \mu \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} du
\end{aligned}$$

Calculating the Expectation of a Gaussian

$$\mathbb{E}[x|\mu,\sigma^2] = \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} udu + \mu \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} du$$

$$\int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} du = 1$$

$$\mathbb{E}[x|\mu,\sigma^2] = \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} udu + \mu$$

Aside: A function is **Odd** iff $f(x) = -f(-x)$.

Odd functions have the property $\int_{-\infty}^{\infty} f(x)dx = 0$.

A function is **Even** iff $f(x) = f(-x)$.

The product of an odd function and an even function is an odd function.

Calculating the Expectation of a Gaussian

$$\mathbb{E}[x|\mu, \sigma^2] = \int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} udu + \mu$$

$$exp\left\{-\frac{1}{2\sigma^2}u^2\right\} \text{ is } \textbf{even}$$

$$u \text{ is } \textbf{odd}$$

$$exp\left\{-\frac{1}{2\sigma^2}u^2\right\} u \text{ is } \textbf{odd}$$

$$\int \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}u^2\right\} udu = 0$$

$$\mathbb{E}[x|\mu, \sigma^2] = \mu$$

Why does Machine Learning need these tools?

Calculus

- We need to find maximum likelihoods or minimum risks. This optimization is accomplished with derivatives.
- Integration allows us to marginalize continuous probability density functions.

Linear Algebra

- We will be working in high-dimension spaces.
- Vectors and Matrices allow us to refer to high dimensional points – groups of features – as vectors.
- Matrices allow us to describe the *feature space*.

Why does machine learning need these tools

Vector Calculus

- We need to do all of the calculus operations in high-dimensional feature spaces.
- We will want to optimize multiple values simultaneously – Gradient Descent.
- We will need to take a marginal over a high dimensional distributions – Gaussians.

Broader Context

What we have so far:

- Entities in the world are represented as feature vectors and maybe a label.
- We want to construct statistical models of the feature vectors.
- Finding the **most** likely model is an optimization problem.
- Since the feature vectors may have more than one dimension, linear algebra can help us work with them.

Bye

- Next
  - Linear Regression