Introduction to Machine Learning

What you can use it for
- pattern recognition (faces, digits, speech),
- bioinformatics (gene finding, introns)
- internet (spam filtering, search engines)
- prediction (stock market)

What you get
- skills in programming, numerical analysis, optimization
- practical experience with data
- easy do-it-yourself algorithms


Overview

Week 1
Linear Algebra, Hilbert Spaces, Numerical Mathematics

Week 2
Learning Theory, Statistics, Risk Functional, Common Distributions, Perceptron

Week 3
Regression, Squared Loss, Noise Models and Loss, Regularization

Week 4
Kernels, Kernel Perceptron, Kernel Regression

Week 5
Large Margin and Optimization, SV Classification, Regression, Novelty Detection

Week 6
Applications and Mini Projects: Text Categorization and Bad Digits

Practical Issues

Scoring
This is a 3 credit point unit. Exercises and programming each count $\frac{1}{3}$, the final exam counts $\frac{1}{2}$.

Problem Sheets
Due Monday at 10am in the mailbox. Late submissions cost 20% a day.
You are expected to work together in groups of 3 and submit one solution sheet per group. If you copy from other groups you will not get points for these solutions.

Tutorials
Ben O’Loghlin (ben@synerg.anu.edu.au) will hold the tutorials (Thursday 2-5pm) which include solutions of the exercise sheets and some programming practice with the SVLab toolbox.

Final Exam
Probably Monday, June 18 (slides, personal notes, calculator and tables are OK).

A Crash-Course in Math

Topics
- Vector spaces, Hilbert and Banach Spaces, Metrics and Norms
- Matrices, Eigenvalues, Orthogonal Transformations, Singular Values
- Operators, Operator Norms, Function Spaces revisited

Rationale
- We need this toolbox to describe the functions we will be dealing with and to set up the optimization/learning problems.
**Metric**

**Definition 1 (Metric)**
Denote by $\mathcal{X}$ a space. Then $d : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+^1$ is a metric on $\mathcal{X}$ if for all $x, y, z \in \mathcal{X}$
1. $d(x, y) = 0$ is equivalent to $x = y$
2. $d(x, y) = d(y, x)$
3. $d(x, z) \leq d(x, y) + d(y, z)$ (Triangle Inequality)

**Example 1 (Trivial Metric)**
For arbitrary $\mathcal{X}$ define $d(x, y) = 1$ if $x \neq y$ and $d(x, y) = 0$ if $x = y$.

**Example 2 (Manhattan Distance)**
For $\mathcal{X} = \mathbb{R}^n$ define $d(x, y) := \sum_{i=1}^{n} |x_i - y_i|$.

---

**Vector Spaces**

**Definition 2 (Vector Space)**
A space $\mathcal{X}$ on which for all $x, y \in \mathcal{X}$ and for all $\alpha \in \mathbb{R}$ the following operations are defined:
1. $x + y \in \mathcal{X}$ (Addition)
2. $\alpha x \in \mathcal{X}$ (Multiplication)

**Definition 3 (Cauchy Series)**
Given a space $\mathcal{X}$, a series $x_i \in \mathcal{X}$ with $i \in \mathbb{N}$ is a Cauchy series if for any $\varepsilon$ there exists an $n_0$ such that for all $m, n \geq n_0$ we have $d(x_m, x_n) \leq \varepsilon$.

**Definition 4 (Completeness)**
A space $\mathcal{X}$ is complete if the limits of every Cauchy series are elements of $\mathcal{X}$.
We call $\bar{\mathcal{X}}$ the completion of $\mathcal{X}$, i.e. the union of $\mathcal{X}$ and all its limits of Cauchy series.

---

**Vector Spaces: Examples**

**Rational Numbers**
Addition and multiplication are obviously OK. However, the space is not complete. For instance, we can find a Cauchy series of $x_i \in \mathbb{Q}$ converging to $\sqrt{2}$.

**Real Numbers**
Addition and multiplication are obviously OK. The same holds for limits (recall algebra lectures).

$\mathbb{R}^n$
Prototypical example of a vector space. addition, multiplication, and limits are obviously OK, e.g., take $\mathcal{X} = \mathbb{R}^3$ and $x = (2, 33.4, 4.2, 2.999, 6)$.

**Polynomials**
Functions such as $f(x) := a + bx + cx^2 + dx^3$ obviously form a vector space. For polynomials of finite order $n$ we can even find a mapping between $\mathcal{X}$ and $\mathbb{R}^n$.

---

**Series**
series $(a_i)$ of numbers with $a_i \in \mathbb{R}$ and $i \in \mathbb{N}$ are clearly vector spaces.

**Fourier Expansions**
expansions via the discrete Fourier transform form a vector space where
$$f(x) = \sum_{j=1}^{n} s_j \sin(jx) + c_j \cos(jx)$$

**Functions**
many classes of functions, e.g., $f : [0, 1] \to \mathbb{R}$.

**Counterexamples**
- $f : [0, 1] \to [0, 1]$ does not form a vector space!
- $\mathbb{Z}$ is not a vector space, unless we only allow multiplications by integers.
- The alphabet $\{a_1, \ldots, a_k\}$ is not a vector space (still it can be an interesting mathematical object, e.g. when determining similarity of documents).
Banach Spaces

Definition 5 (Norm)
Given a vector space $X$, a mapping $\| \cdot \| : X \to \mathbb{R}_+^*$ is called a norm if for all $x, y \in X$ and all $a \in \mathbb{R}$ it satisfies

1. $\|x\| = 0$ if and only if $x = 0$
2. $\|ax\| = |a| \|x\|$ (scaling)
3. $\|x + y\| \leq \|x\| + \|y\|$ (triangle inequality)

A mapping $\| \cdot \|$ not satisfying (1) is called a pseudo norm.

A metric via $d(x, y) := \|x - y\|$.  

Definition 6 (Banach Space)
A complete vector space $X$ together with a norm $\| \cdot \|$.  

Banach Spaces: Examples

$\ell_p^n$ Spaces
Take the $\mathbb{R}^n$ endowed with the norm $\|x\| := \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$ where $p > 0$. Note that in $\mathbb{R}^n$ all norms are equivalent, i.e. there exist $c, C \in \mathbb{R}^+$ such that

$$c\|x\|_c \leq \|x\|_b \leq C\|x\|_c$$

for all $x \in X$ and likewise $1 \leq \|x\|_b \leq \|x\|_c \leq \frac{1}{c} \|x\|_b$

$\ell_p^n$ Spaces
These are subspaces of $\mathbb{R}^n$ with $\|x\| := \left(\sum_{i=1}^{n} |x_i|^p\right)^{\frac{1}{p}}$

Not for all series $x_i$ the sum converges, e.g. $x_i = \frac{1}{i}$ is in $\ell_2$ but not in $\ell_1$.

Function Spaces $L_p^n(X)$
We replace sums by integrals over $X$ and obtain $\|f\| := \left(\int_X |f(x)|^p dx\right)^{\frac{1}{p}}$. Again, not for all $f$ this integral is defined, i.e. they are not elements of the corresponding $L_p^n(X)$.

Hilbert Spaces

Definition 7 (Dot Product)
Given a vector space $X$, a mapping $\langle \cdot, \cdot \rangle$ with $X \times X \to \mathbb{R}$ which for all $a \in \mathbb{R}$ and $x, y, z \in X$ satisfies

1. $\langle x, y \rangle = \langle y, x \rangle$ (symmetry)
2. $\langle ax, y \rangle = a \langle x, y \rangle$ (linearity)
3. $\langle x, y + z \rangle = \langle x, y \rangle + \langle x, z \rangle$ (additivity)

Definition 8 (Hilbert Space)
A complete vector space $X$, endowed with a dot product $\langle \cdot, \cdot \rangle$.

The dot product automatically generates a norm (and a metric) by $\|x\| := \sqrt{\langle x, x \rangle}$. Thus Hilbert spaces are special case of a Banach space.

These are the spaces we will work with in this lecture.

Hilbert Spaces: Examples

Euclidean Spaces
Use standard dot product for $x, y \in \mathbb{R}^n$ given by $\langle x, y \rangle := \sum_{i=1}^{n} x_i y_i$

Function Spaces ($L_2^n(X)$) Functions on $X$ with $f : X \to \mathbb{C}$ for all $f \in F$. Here we can define the dot product for $f, g \in F$ by $\langle f, g \rangle := \int_X \overline{f(x)} g(x) dx$. Note that we take the complex conjugate of $f$. Also note that all we did was to replace the sum by an integral.

$\ell_2$ (Infinite) series of real numbers, $\ell_2 \subset \mathbb{R}^n$. We define a dot product for $x, y \in \ell_2$ by $\langle x, y \rangle = \sum_{i=1}^{\infty} x_i y_i$

Polarization Inequality
We can recover the dot product from the norm $\|x\|$ by computing $\|x + y\|^2 - \|x\|^2 - \|y\|^2 = 2\langle x, y \rangle$. 


Matrices

In the following we assume that a matrix $M \in \mathbb{R}^{n \times n}$ corresponds to a linear map from $\mathbb{R}^n$ to $\mathbb{R}^n$ and is given by its entries $M_{ij} \in \mathbb{R}$.

Symmetry

A symmetric matrix $M \in \mathbb{R}^{n \times n}$ satisfies $M_{ij} = M_{ji}$.

Antisymmetry

An antisymmetric matrix $M \in \mathbb{R}^{n \times n}$ satisfies $M_{ij} = -M_{ji}$.

Rank

Denote by $I$ the image of $\mathbb{R}^n$ under $M \in \mathbb{R}^{n \times n}$. Since $M$ is a linear map, we can find a $I$ as a linear combination of vectors. rank$(M)$ is the smallest number of such vectors that span $I$.

Orthogonality

A matrix $M \in \mathbb{R}^{n \times n}$ with $M^T M = I$ is called an orthogonal matrix (if $M \in \mathbb{C}^{n \times n}$ it is called unitary). This means $M^{-1} = M^T$.

Matrix Invariants

Trace:

$$\text{tr} M = \sum_{i=1}^{n} M_{ii}$$

One can show $\text{tr}(AB) = \text{tr}(BA)$ and thus for symmetric matrices

$$\text{tr} M = \text{tr}(O^T AO) = \text{tr}(AOO^T) = \text{tr} \Lambda = \sum_{i=1}^{n} \lambda_i$$

Determinant:

Antisymmetric multilinear form, i.e. swapping columns or rows changes the sign, adding elements in rows and columns is linear. Useful form

$$\det M = \prod_{i=1}^{n} \lambda_i$$

Both trace and determinant are invariant under orthogonal transformations $M \rightarrow O^T M O$ where $O \in \text{SO}(n)$ for or of the matrix.

Matrices, Part II

Orthogonality, Part II

It consists of mutually orthogonal rows and columns. The corresponding matrix group is often denoted by $O(n)$ (the orthogonal group). If it is only a rotation, it is called $\text{SO}(n)$ (special orthogonal group).

Note that from $M^T M = I$ it also follows that $MM^T = I \Rightarrow (MM^T)M = M$ (and all matrices have full rank).

Example

Rotation matrices in $\mathbb{R}^2$ are given by

$$M = \begin{bmatrix} \cos \phi & -\sin \phi \\ \sin \phi & \cos \phi \end{bmatrix}$$

here $\det M = 1$.

Mirror matrices are

$$M = \begin{bmatrix} 1 & 0 \\ 0 & -1 \end{bmatrix}$$

and

$$M = \begin{bmatrix} -1 & 0 \\ 0 & 1 \end{bmatrix}$$

here $\det M = 1$.

Matrix Norms

Operator Norm

The norm of a linear operator $A$ between two Banach spaces $X$ and $Y$ is defined as

$$||A|| := \max_{x \in X \setminus \{0\}} \frac{||Ax||}{||x||}$$

This clearly satisfies all conditions of a norm:

- $||\alpha A|| = \max_{x \in X} \left| |\alpha| ||x|| \right| = |\alpha|||A||$.
- $||A + B|| = \max_{x \in X} \left| \frac{||Ax + Bx||}{||x||} \right| \leq \max_{x \in X} \frac{||Ax||}{||x||} + \max_{x \in X} \frac{||Bx||}{||x||} = ||A|| + ||B||$
- $||A|| = 0$ implies $\max_{x \in X} ||x|| = 0$ and thus $Ax = 0$ for all $x$. This means that $A = 0$.

Frobenius Norm

For a matrix $M \in \mathbb{R}^{n \times n}$ we can define a norm in analogy to the vector norm by

$$||M||^2_{Frob} = \sum_{i} \sum_{j=1}^{n} M_{ij}^2$$
Eigensystems

Definition 9 (Eigenvalues, Eigenvectors)
Denote by $M \in \mathbb{R}^{n \times n}$ matrix, then an eigenvalue $\lambda \in \mathbb{R}$ and eigenvector $\mathbf{x} \in \mathbb{R}^n$ satisfy

$$M \mathbf{x} = \lambda \mathbf{x}$$

Ananlogously for operators $A : \mathcal{X} \rightarrow \mathcal{X}$ we have $A \mathbf{x} = \lambda \mathbf{x}$.

Caveat
We cannot always find a complete eigensystem. Example: $\begin{bmatrix} 1 & 2 \\ 0 & 1 \end{bmatrix}$

Symmetric Matrices
All eigenvalues of symmetric matrices are real and symmetric matrices are fully diagonalizable, i.e. we can find $m$ eigenvectors.

Matrix Norms Revisited

Operator Norm: Using $M \in \mathbb{R}^{n \times n}$ we have

$$\|M\|^2 = \max_{\mathbf{x} \in \mathbb{R}^n, \|\mathbf{x}\|=1} \|M\mathbf{x}\|^2$$

$$= \max_{\mathbf{x} \in \mathbb{R}^n} \|M\mathbf{x}\|^2$$

$$= \max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^T M^T M \mathbf{x}$$

$$= \max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^T O A O^T O A O \mathbf{x}$$

$$= \max_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{x}\|^2 \Lambda$$

$$\|M\|^2_{\text{Frob}} = \max_{\mathbf{x} \in \mathbb{R}^n} \mathbf{x}^T M^T M \mathbf{x}$$

$$= \max_{\mathbf{x} \in \mathbb{R}^n} \|\mathbf{x}\|^2 \Lambda$$

Positive Matrices

Positive Definite Matrix:
A matrix $M \in \mathbb{R}^{n \times n}$ for which for all $\mathbf{x} \in \mathbb{R}^n$ we have

$$\mathbf{x}^T M \mathbf{x} \geq 0 \text{ if } \mathbf{x} \neq \mathbf{0}$$

This matrix has only positive eigenvalues since for all eigenvectors $\mathbf{x}$ we have $\mathbf{x}^T M \mathbf{x} = \lambda \mathbf{x}^T \mathbf{x} = \lambda \|\mathbf{x}\|^2 > 0$ and thus $\lambda > 0$.

Induced Norms and Metrics:
Every positive definite matrix induces a norm via

$$\|\mathbf{x}\|_M := \mathbf{x}^T M \mathbf{x}$$

- Linearity is obvious, so is uniqueness
- The triangle inequality can be seen by writing

$$\|\mathbf{x} + \mathbf{x}'\|_M^2 = (\mathbf{x} + \mathbf{x}')^T M^2 (\mathbf{x} + \mathbf{x}') = \|M^2 (\mathbf{x} + \mathbf{x}')\|^2$$

and using the triangle inequality for $M^2 \mathbf{x}$ and $M^2 \mathbf{x}'$. 
Singular Value Decompositions

Idea:
Can we find something similar to the eigenvalue / eigenvector decomposition for arbitrary matrices?

Decomposition:
Without loss of generality assume $m \geq n$ For $M \in \mathbb{R}^{m \times n}$ we may write $M$ as $U \Lambda O$
where $U \in \mathbb{R}^{m \times m}$, $O \in \mathbb{R}^{n \times n}$, and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_k)$.
Furthermore $O^T O = O O^T = U^T U = I$.

Useful Trick:
Nonzero eigenvalues of $M^T M$ and $M M^T$ are the same. This is so since

\[ M^T M x = \lambda x \quad \text{and hence} \quad (M M^T) M x = \lambda M x \quad \text{or equivalently} \quad (M M^T) x' = \lambda x'. \]